2017

On the Relationship Between Sampling Rate and Hidden Markov Models Accuracy in Non-intrusive Load Monitoring

Steven Lynch
Technological University Dublin, stvenlynch.irl@gmail.com

Luca Longo
Technological University Dublin, luca.longo@tudublin.ie

Recommended Citation

Lynch, S. & Longo, L. (2017). On the Relationship Between Sampling Rate and Hidden Markov Models Accuracy in Non-intrusive Load Monitoring. *Irish Conference on Artificial Intelligence and Cognitive Science*, December, Dublin, Ireland. doi:10.21427/dvf8-nn35

On the relationship between sampling rate and Hidden Markov Models Accuracy in Non-Intrusive Load Monitoring

Steven Lynch, Luca Longo*

School of Computing, Dublin Institute of Technology
stvenlynch.irl@gmail.com, *luca.longo@dit.ie

Abstract. Providing domestic energy consumers with a detailed breakdown of their electricity consumption, at the appliance level, empowers the consumer to better manage that consumption and reduce their overall electricity demand. Non-Intrusive Load Monitoring (NILM) is one method of achieving this breakdown and makes use of one sensor which measures overall combined electricity usage. As all appliances are measured in combination in NILM this consumption must be disaggregated to extract appliance level consumption. Machine learning techniques can be adopted to perform this disaggregation with various levels of accuracy, with Hidden Markov Model (HMM) derivatives offering among the most accurate results. This work investigates how sensor sampling rate affects disaggregation accuracy obtained through HMM. Derived re-sampled data was passed through HMM and the resulting accuracy compared with the sampling rates. Correlation was observed and statistically verified. Two distinct groups of appliances were later identified, one which was highly correlated and another in which correlation was not observed.

1 Introduction

A detailed understanding of their energy consumption can empower consumers to modify their energy usage. This can lead towards a reduction in over-reliance on peak-time energy as well as reduced energy usage overall. Such modifications in energy usage patterns would have the effects of reducing the current high requirement for reserve and spin-up plants and under-utilised consumption. Similarly, they can potentially reduce the overall number of power plants in general, leading to major system reliability, economic, and environmental benefits. To understand usage patterns, modifications of energy must be measured. Either the load on each appliance is individually measured and recorded before this information fed into a central monitoring system, or the consumption pattern as a whole can be measured and interrogated to determine load usage by individual components. The former method is achievable with current technology but is prohibitively expensive as it requires that measurement and information transmission components be attached to every device in the system. The latter method, Non-Intrusive Load Monitoring (NILM) or Energy Disaggregation,

requires measurement at only a single point in the system. Unfortunately, the problem of working out which appliances are drawing power from the system at any point in time is computationally complex and has not yet been fully solved. Various modeling approaches have been proposed for tackling such a problem. However, previous research haven't focused on the investigation of the impact of appliances sampling rates on model accuracy. The aim of this research is to analyse one of the most commonly used data sets for comparing Non-Intrusive Load Monitoring techniques and ascertain how sampling rate affects model accuracy. This analysis will be performed by building a Hidden Markov Model across appliance channels responsible for the highest amount of power usage. The specific question being research is: *"Is the relationship between sampling frequency for feature extraction and selection in Non-Intrusive Load Monitoring and model accuracy significant?"*

The reminder of this paper is organised as it follows. Section 2 presents related work on load monitoring in general, intrusive and non-intrusive monitoring subsequently. It then reviews researcher works that have employed Machine Learning for load monitoring. An experiment is designed in section 3 to investigate the relationship between the rate at which power observations are sampled and the accuracy of Hidden Markov Models built on those samples. Section 4 presents the results of such investigation while section 5 critically discuss the findings. Eventually, section 6 concludes this research highlighting its impact to the body of knowledge and setting future avenues for research.

2 Related Work

Significant reductions in domestic energy use have been shown to be achievable through the utilisation of detailed and granular energy consumption feedback mechanisms [6]. A detailed review of 57 separate studies performed globally between 1974 and 2010 suggested an average reduction in energy usage of between 5% and 14% per household is achievable by providing usage information to the consumer in addition to the standard monthly bill [7]. During peak energy usage periods even more significant savings were observed of between 10% and 18%. Through informing the consumer overall energy use can be decreased and the variance between peak and off-peak demand undergoes slight normalisation,which results in a more economical and potentially more stable power grid. Feedback at the appliance level results in roughly double the energy savings of real time feedback at the aggregate level. Intrusive Load Monitoring (ILM) relies on directly observing the consumption for each appliance at the point which that consumption occurs. It is a bottom-up approach where devices are individually measured before, ideally, being presented to the consumer though a single interface [5] ILM therefore requires a distributed network of sensors throughout the home. This method of load monitoring can lead to highly accurate results. However there are a number of major drawbacks, most notably the higher cost associated with the volume of sensors required and the maintenance

of the distributed network of sensors. Non-Intrusive Load Monitoring (NILM) is the top-down approach to energy consumption measurement. The total combined power consumption is measured and through the use of disaggregation techniques an attempt is made to split that value into its constituent parts [9]. Current state of the art approaches are able to disaggregate the combined load signal with an accuracy in the region of 80- 90 % [10]. As NILM does not require that observations be made at the appliance level, there is a significant reduction in the physical hardware required over ILM. Only a single measurement point is required. This negates the requirement for a distributed network of sensors, resulting in a significant initial outlay cost to the consumer who does not need to maintain the sensor network [6]. NILM is formally described in equation 1.

$$P(t) = p_1(t) + p_2(t) + \ldots + p_n(t) \tag{1}$$

where $P(t)$ is the *known* total overall power consumed at time t and $p_i(t)$, the *unknown* power p consumed by each appliance i at time t in a system with n appliances. Multiple devices may be active at any given point in time. Electrical loads tend to exhibit unique energy consumption patterns over time depending on their state. These patterns are known as *load signatures*. A single appliance may have multiple load signatures, for example the load signature of a microwave operating at high power will differ from the same microwave operating on the defrost setting. Different models of the same appliance type will tend not to exhibit the same load patterns. When enough granularity is introduced into the observations, different instances of the same devices, that means multiple microwaves of the same model, will tend to exhibit different load conditions under the same parameters. This uniqueness, while useful in discriminating between different light-bulbs in the same home also tend to make model generalisation more complex [8]. A number of data sets comprising appliance level and aggregated domestic energy consumption such as REDD [12] and SMART [1] exist in the literature. Each of these data sets contain power observations at the individual appliance and whole house level. The REDD dataset is the most commonly cited at present [12]. It spans multiple homes and it contains both high and low frequency observations.

2.1 Machine learning for non-intrusive load monitoring

Both supervised and unsupervised machine learning techniques have been utilised in the search of an optimised NILM solution. Supervised machine learning relies on a priori knowledge, of which there are two levels with respect to NILM. The total number of appliances in the system as well as a name, or some other designation, for each appliance is the most basic level of a priori knowledge. The second level is knowledge of the consumption patterns of each appliance in the system as opposed to only the combined usage of all appliances [2]. It is at the second level of knowledge that the split between supervised and unsupervised is most relevant to NILM techniques. With sufficient feature selection performed classification accuracies in the region of 90% has been obtained using Naive Bayes, J48 Decision Trees and Bayesian Networks which rises to 95%

accuacy [15]. Similarly, the application of ensemble methods such as random forest or LogitBoost proved useful in building highly accurate models [14]. In [13], an attempt to implement a supervised NILM technique on board a large US Coast Guard ship was performed. In this context, even with a highly trained, regimented, and dedicated workforce, the accurate collation of the required a priori knowledge proved both difficult and highly time consuming. Existing historical data sets such as the REDD [12], or SMART [1] can be harnessed to build NILM solutions. The major limitation of these models is that they tend to perform poorly when appliances which were not present in the historical dataset are introduced to the model. As discussed above, different instances of the same appliance will tend to have different load signatures. Therefore even if all of the appliances in the implemented system are also present in the historical dataset, unless they were *the appliances* from the historical dataset a loss of accuracy is to be expected. As shown in [3], while historical datasets are invaluable for comparing different NILM methods and algorithms which may later be deployed, actual models built in such a supervised manner tend to function with limited success once deployed to real world situations. When available dataset do not contain labelled information, then unsupervised machine learning methods have been employed. For example, in an analysis by [4], Hidden Markov Models (HMM) and variants of the HMM were discovered to be the highest performers. These variants include Hierarchical Dirichlet Process HMM (HDP-HMM), factorial Hidden semi-Markov model and Conditional (FHMM), Hidden Markovs as Bayesian networks, and Additive Factorial Approximate Maximum A-Posteriori (AFAMAP) [11]. Each of the HMM variant models relies on the accuracy of the underlying base HMM in order to achieve optimal results. A major drawback of purely unsupervised machine learning techniques is that their output is not in a form that is easily digestible. The model may, with an acceptable level of accuracy, correctly classify all of the power use across the system. However, if the user cannot be made aware of what specific appliance is using energy, at any given point in time, they are unable to modify their consumption behaviour and realise energy savings. Semi-automatic machine learning techniques offer a solution to this problem [2]. An unsupervised method can be used to identify different appliances and apply a temporary label to them after which a user manually modifies the labels so that they are understandable by them and useful in NILM feedback [5].

The literature seems to focus on evaluating and improving different machine learning techniques. Little research was found on the design choices made at the data capture stage with respect to NILM. In the creation and dissemination of such data sets it is necessary to choose appropriate sampling rates for both data transmission and storage reasons. Do these design choices have an impact on model accuracy? This research has potential to inform both the future collection of new comparative data sets as well as the design choices made for in-situ NILM mains monitoring.

3 Design and Methodology

A secondary research study has been designed for the investigation of the relationship between the rate at which power observations are sampled and the accuracy of Hidden Markov Models (HMM) built upon those samples. The REDD dataset has been chosen for such an investigation [12]. A note that the specific accuracy of the models themselves is not particularly relevant, how the accuracy changes across the different sampling rates is. Additionally, as HMM models are the foundation of factorial Hidden Markov models (fHMM), it is not necessary to perform this evaluation at the fHMM level. An HMM evaluation is sufficient. The whole house disaggregation itself will not form part of this experiment, but it is expected that whole house disaggregation will benefit from this research. The REDD is unlabelled with respect to true activation state of each appliance at any given point in time. It requires that an approximation of ground truth be derived per appliance to test model accuracy. The research hypothesis set is:

H_0: *"Feature Sampling Frequency and Model Accuracy are linearly related"*.

The research hypothesis has been be tested with a statistical confidence level of 95% and the CRISP-DM process for data mining has been followed.

Data Understanding The REDD dataset is made available for general research and its compressed size is 1.58GB. It consists of three separate sets of data which vary in their levels of completeness.

A Low frequency data containing average power readings for the mains power sampled at 1Hz and individual circuits at $^1/3$ Hz;
B Higher frequency data created by means of lossy compression;
C Raw data sampled at a very high frequency.

The raw data provided is incomplete spanning only 90 minutes of observation for 2 houses. The high frequency compressed data set contains whole house aggregated data and no observations at the individual appliance level. The low frequency data set, which consists of both appliance level and aggregated level recordings, was used for the purpose of this research. Data related to house 3 and house 5 far outperform all of the other houses with regards to synchronisation of observations across both mains and appliance. For house 5, mains observations always exists when appliance observations exists while for house 3 aggregated observations exists for one of the days without appliance level observations. The low frequency data set contains average power readings for both the two power mains (US domestic circuitry comprises two 110V mains lines) and the individual appliance circuits of the house. The data is logged at a frequency of about 1 second for a mains and 3 seconds for the appliances. Each appliance file contains UTC timestamps and power readings (recording the apparent power of the circuit) for the channel. There are 104 unique appliance channels contained within the REDD dataset, 44 of which exists in houses 3 and 5. Appliances in house 5 each consist of 404,107 observations with 1,427,284 observations in each of its aggregated mains channels. For house 3 there are 80,417 readings for each appliance and 302,122 observations in both aggregated mains channels.

Data Preparation For maximum replicability, a decision was made to focus on house 3 and 5, the only houses present in both the low-frequency and the high-frequency data sets. This decision reduced the number of potential channels to 44. Calculation alone for all 44 channels in house 3 and 5, with available computational power, would have taken close to 2 months to perform and was deemed excessive. Another decision was made to focus on high power usage appliance channels. While the low frequency appliance observation rate has been originally described as one sample every 3 seconds, in practice this is not the case. The duration of each sample is not explicitly stated in the REDD dataset description, but only the starting time of each observation is recorded. It can be extrapolated that the start time of a subsequent observation corresponds to the end time of the current observation. The low frequency data set, while interspersed with gaps, follows a rough pattern of 15 recording alternating between 3 and 4 seconds apart, followed by a 16^{th} recording after an 8 or 9 second lag. Based on this, lack of observations spanning more than 10 seconds are considered gaps. Gaps are not included when re-sampling as the period over which the average power reading is being measured is unknown and can therefore have a significant negative impact on the re-sampling calculation. Observations which correspond to gaps are thus dropped at the point prior to re-sampling. Because the goal is to investigate how sampling rates affect model accuracy, the low frequency channels being tested were resampled to different sampling rates. After initial tests on 6 sampling rates, a further 26 sampling rates were added to provide a good spread of experimental data and granularity of results. Sampling rates ranged from an observation every second through to one ever 6 minutes, with more focus placed on the higher granularity resamples. Down-sampling to a rate of 1 sample every second is achievable but no more information can be attained at higher frequency sampling rates. Importantly, while a single span may contain multiple observations, so to may a single observation exists across multiple spans. Re-sampling was performed using a custom algorithm.

1. set a variable *timeInSpan* to 0 for each observation;
2. calculate the span for each observation;
3. calculate the number of spans in which each observation exists as the difference between the start and end span;
4. tag the observations that exists entirely within a single span and set their *timeInSpan* to the length of the observation;
5. duplicate observations that exist in multiple spans; calculate the length of time spent in the either the 1st or final span and recorded it in *timeInSpan*;
6. a new span observation is added for each span crossed, for observation in more than 2 spans (except 1st and last span, already handled in 5); for each new observation, set the *timeInSpan* to the full length of the span;
7. merge sll of the observations from steps 4, 5, 6 into a single structure;
8. calculate an average voltage, reading weighted based on *timeInSpan* for each span (spans with no observations are considered gaps and not used in subsequent models; for spans with both gaps and observations, the weighted average of the observation is considered to be the average observations for the duration of the span).

Data issues Hidden Markov Models (HMMs) rely on the sequential nature of the data. HMMs expect to be able to draw assumptions on the current state of the system based on the previous state of the system. As such the model must know where there is a break in the sequence. Sequential data without gaps is considered to be a run, with each run being considered to be an entirely new sequence on to which testing or training may be performed. It is to be expected that the threshold of temporal gaps allowed within a run will have a bearing on experimental results. As processing and model building of all of the required HMM for a single channel took in the order of 30 hours a single threshold was used throughout. The threshold considered as an acceptable gap within a run was taken as 60 seconds. Short experiments with differing thresholds at a 1 second sampling rate were carried out prior to making this decision. Table 1 details the number of runs as a result of differing acceptable gap thresholds across both considered houses. As the threshold reduces model run time increased significantly. It was assumed that, under practical scenarios, most appliances were to be active for longer than a minute, thus 60 second gaps were tolerable.

	Acceptable Gap Threshold (in seconds)													
	1s	10s	20s	30s	40s	50s	**60s**	70s	80s	90s	100s	120s	180s	240s
House 3	812	812	252	223	197	183	**165**	146	127	111	88	70	50	36
House 5	267	264	52	38	32	26	**24**	23	22	22	22	20	14	12

Table 1: Number of runs, per house at differing acceptable gap thresholds.

Adding constant noise Hidden Markov models work by calculating the probability of a change in state. It was observed that many of the appliances being measured remain in a constant state for extended periods of time. This resulted in runs where zero variance was observed in the data over the course of many days. It proved technically impossible to build a model under such circumstances. In order to counteract this, variance was added to the data in the form of $^1/_2$ Watt $^1/_2$ Hz wave overlaid across each dataset. This new noise, while insignificant for the overall power usage and thus having a minimal effect on the results, introduced the required variance to allow the HMM to function.

Modeling According to the literature, the observed states are known to follow a Gaussian distribution. As such a *Gaussian HMM* was used. Tests were performed across each of the appliances at 2 second and 10 second sampling rates. In order to determine the optimal number of expected states using the Akaike Information Criterion (AIC) . AIC is a measure of the relative quality of statistical models for a given data set. It provides a relative estimate of the information loss of a given model. The optimal number of states was determined as 3. As a probabilistic model which relies on local maximums for optimisation, the same data is to be expected to return a slightly different model depending on the initial seed provided to the model. Seeds for this experiment were chosen randomly. 15

models with different initial seeds were created for one of the data sets to further understand the scale of these differences. These differences were not found to be significant. The results across all 15 models were found to be broadly in line with one another. In order to allow somewhat for seed difference, but yet complete the experiment in a reasonable time frame, 3 models were created for each data set, each with a different initial seed. For each channel tested 3 differently seeded models were built, fit, and trained across the 32 sampling rates. The data was split into training and test sets at a 60:40 ratio. Due to the sequential nature of the data random sampling was deemed inappropriate as it would de-sequentialise the data. To overcome this the data was initially grouped into sequential *runs* Runs longer than 2 hours in duration were split into multiple runs with a maximum length of 2 hours each. Random sampling , without replacement was performed of the runs until the training data set comprised 60 % of the runs. The remaining 40% was considered the test data set.During the training phase the Viterbi Algorithm identifies the hidden states which are most likely to have generated the observed states. This is based on the knowledge of the number of expected states which was passed, which for the purpose of this experiment was 3. The Baum-Welch EM algorithm identifies the local maximums of the model parameters iteratively so as to optimise the model. Testing a HMM model involves feeding data not previously used in creating that model. This is crucial to fairly assess the model accuracy. The depmixS4 package does not allow for data absence during the fit phase to be added as test or validation data. To overcome this limitation new models were built for the test phase which were fit using the 3 parameters of the training models that describe a complete HMM (the state transition probability matrix, the emission probability matrix, the initial state distribution). State probabilities were assigned by the model to each observation in the training data set. The state with the highest probability per observation, being the most likely state, was considered to be the models assertion. Within the REDD dataset the absolute ground truth state per observation channel is unknown and must be extrapolated from the data. The derived metric used as the ground truth in this experiment is that a channel is considered active at any point in which power consumption exceeds 10% of maximum usage for that channel. Otherwise it is considered inactive. For each observation in the test data there now exists a calculated ground truth of state and the models assertion of state. Comparison of these states was used to determine the accuracy of each model. This resulted in 96 models per observed channel (3 randomly seeded models for each of the 32 sampling rates = 96).

Evaluation A determination of which states derived from the HMM corresponded to which ground state was performed by finding the median power value of each state observed by the model. High value corresponded to ground states of *On* while low values corresponded to ground states of *Off*. As 3 state HMMs were used but only 2 states (*On* or *Off*) were used as the calculated ground truth 2 of the HMM states would be classified as either *On* or *Off*. This resulted in a merging of 2 of the HMM determined states.

4 Results

Precision, Recall and F1 score were plotted for all 8 of the appliance channels measured. The F1 value for each appliance was then plotted along with a Loess curve as an indicator of potential correlation between sample length and F1 accuracy. As shown in figure 1 there appears to be weak correlation overall. A Pearson's product-movement correlation tests was performed to evaluate the statistical significance of the correlation. The test resulted in a correlation, at the 95% confidence interval, between -0.303 and -0.065 with a p-value of 0.00285. As the 95 % confidence band never intersects with a correlation value of zero and the p-value was less than 0.05 there is evidence to accept the null hypothesis. This means that sampling rate and model accuracy are related across employed dataset. The tested appliance channels appear to fall into two categories.

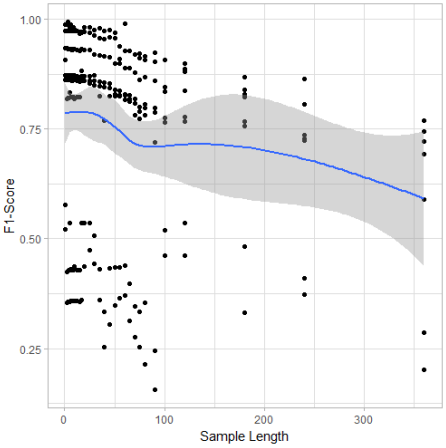
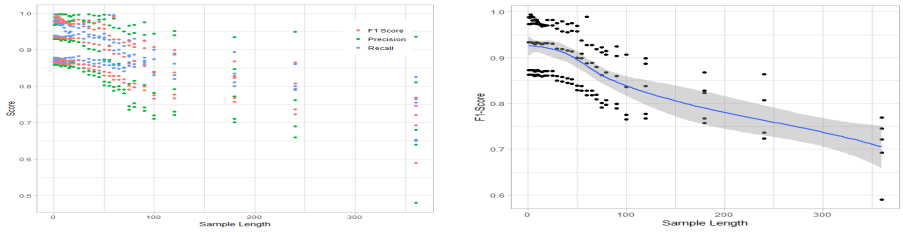


Fig. 1: F1 score across all measured appliance channels with a Loess curve.

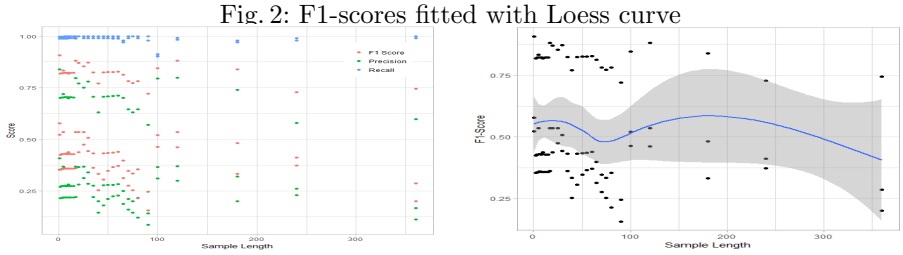
Category A has broadly similar precision and recall within each model while category B tends to have high recall but very low precision. Both the categories were investigated and statistically tested. Category A includes appliances which have broadly similar precision and recall at a given sample rate. These appear to have an F1-score above 80% at higher sampling rates and to show a reduction in accuracy which is more closely correlated to sample length than the total data set as shown in figure 2. Yet their overall F1-score at the 360s sampling rate is in the 60% to 80% band. A Pearson's product-movement correlation tests was performed on the category A appliances. With a p-value $< .0001$ the correlation was between -0.761 and -0.592 at a 95% confidence interval suggesting to accept the hypothesis.

Category B includes appliances which have high recall but very low precision at with a given sample rate. Their F1-score appears to vary much less with respect to sample length than was found in the category A appliances. These appliances have a lower F1-score than is found in the Category A appliances and would appear to have significantly less correlation between sample length and F1-score (as per fig. 3, from the Loess curve). The Pearson's correlation tests returned a p-value of 0.2587. Within a 95% confidence interval the true correlation was found to be between -0.309 and 0.086. For category B appliances, the null hypothesis was rejected, showing how sample rate and accuracy of models are not related.



(a) Evaluation metrics (b) F1-score fitted with a Loess curve

Fig. 2: F1-scores fitted with Loess curve



(a) Evaluation metrics (b) F1-score fitted with a Loess curve

Fig. 3: Category B appliances.

5 Discussion of findings

The results of the experiment have shown, with a confidence level of 95%, that there is weak correlation between sample length and model accuracy as a whole across all of the tested appliance channels. Up to the limits of the sampling rates, which could be derived from the REDD dataset, higher sampling will yield more accurate results. While strong correlation was observed in category A appliances, in order for the model to understand which appliances were category A appliances the model either labelling or a higher level of user feedback would be required. As discussed above this is sub-optimal in a domestic setting. All appliances must therefore be considered in their entirety. Category A appliances would seem to be appliances which have a finite numbers of states. The appliance channels categorised in this research as category B present as devices having continuously variable state characteristics. The move towards more devices which are more energy efficient tends towards the creation of appliances with more variable characteristics in terms of consumption patterns. As the ratio of category A to category B appliances within the home shifts over time, it can be expected that this shift will have implications for these findings. However, based on the REDD dataset, it would appear that the higher the sampling rate the higher the accuracy of predictions which can be obtained through the use of HMM and by extension fHMM. As this research was performed on HMM, which is the lowest level building block for all of the fHMM models currently being researched or implemented with respect to NILM, this research has a high level of applicability. Only the appliances found to have the highest consumption within 2 houses were tested. It is possible that different conclusions could have

been reached had each appliance channel undergone testing. 32 sampling rates were chosen for this experiment. More sampling rates, especially at above 60 seconds would have provided more granular results. It was observed that HMM returns slightly different results under different seed settings. Three HMMs were built per sampling rate for each appliance channel. The effect of seed variance could have been further reduced by building more models per sampling rate.

6 Conclusion

This research investigated the problem of providing more accurate feedback to end users of their power consumption patterns. Through the literature review it was identified that there are significant energy efficiencies to be wrought by the provision of granular feedback to the end user. A comparison was then made between the differing techniques of collecting data to feedback to the user: intrusive and non-intrusive methods. Following a discussion on their relative merits and drawbacks, non-intrusive methods were further examined. Hidden Markov Model (HMM) was identified as the building block of the state of the art approaches to Non-Intrusive Load Monitoring (NILM) and as such it was believed that any improvements in HMM would feed though to improvements in those methods. A well-known dataset (REDD) has been used as a baseline for comparing different machine learning algorithms. However, it was unclear if it had been optimally built with respect to the chosen sampling rate. Similarly, other datasets of the same ilk were found to sample at different frequencies but no evaluation of the different sampling frequencies which could be extrapolated from the REDD dataset was found. Therefore, the research question investigated focused on the relationship between sampling frequency for feature extraction and selection in Non-Intrusive Load Monitoring and HMM model accuracy. An experiment was designed for such an investigation, and Gaussian Hidden Markov Models were employed as modelling technique. Findings support the null research showing how sampling rate is linearly related to model accuracy, with a 95% confidence level. In particular, with higher sampling rates, model accuracy increases. As fHMM techniques are the current state of the art approach, this research can have a positive impact at the forefront of the NILM field. This project verified the correlation between sampling rate and model accuracy at the *building block* level of HMM. While mathematically this should apply to fHMM techniques such as the AFAMAP this has not been verified experimentally and is proposed as future work. A correlation was identified but no investigation was performed on an optimal sampling rate taking recording, storage, processing and financial cost into account. Having found a relationship, finding an optimal level is fundamentally the next step towards optimising NILM.

References

1. Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. Smart*: An open data set and tools for enabling research in sustainable homes. *SustKDD, August*, 111:112, 2012.

2. Karim Said Barsim and Bin Yang. Toward a semi-supervised non-intrusive load monitoring system for event-based energy disaggregation. In *2015 IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, pages 58–62. IEEE, 2015.
3. Nipun Batra, Amarjeet Singh, and Kamin Whitehouse. If you measure it, can you improve it? exploring the value of energy disaggregation. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, pages 191–200. ACM, 2015.
4. Roberto Bonfigli, Stefano Squartini, Marco Fagiani, and Francesco Piazza. Unsupervised algorithms for non-intrusive load monitoring: An up-to-date overview. In *Environment and Electrical Engineering (EEEIC), 2015 IEEE 15th International Conference on*, pages 1175–1180. IEEE, 2015.
5. Enrico Costanza, Sarvapali D Ramchurn, and Nicholas R Jennings. Understanding domestic energy consumption through interactive visualisation: a field study. In *The ACM Conf. on Ubiquitous Computing*, pages 216–225. ACM, 2012.
6. Sarah Darby et al. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 486(2006), 2006.
7. Karen Ehrhardt-Martinez, Kat A Donnelly, Skip Laitner, et al. Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities. American Council for an Energy-Efficient Economy Washington, DC, 2010.
8. Jon Froehlich, Eric Larson, Sidhant Gupta, Gabe Cohn, Matthew Reynolds, and Shwetak Patel. Disaggregated end-use energy sensing for the smart grid. *IEEE Pervasive Computing*, 10(1):28–39, 2011.
9. George William Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.
10. Christoph Klemenjak and Peter Goldsborough. Non-intrusive load monitoring: A review and outlook. *arXiv preprint arXiv:1610.01191*, 2016.
11. J Zico Kolter and Tommi S Jaakkola. Approximate inference in additive factorial hmms with application to energy disaggregation. In *AISTATS*, volume 22, pages 1472–1482, 2012.
12. J. Zico Kolter and Matthew J. Johnson. REDD: A public data set for energy disaggregation research. *SustKDD workshop on Data Mining Applications in Sustainability*, 2011.
13. Peter Lindahl, Steven Leeb, John Donnal, and Greg Bredariol. Noncontact sensors and nonintrusive load monitoring (nilm) aboard the uscgc spencer. In *IEEE AUTOTESTCON, 2016*, pages 1–10. IEEE, 2016.
14. Andreas Reinhardt, Paul Baumann, Daniel Burgstahler, Matthias Hollick, Hristo Chonov, Marc Werner, and Ralf Steinmetz. On the accuracy of appliance identification based on distributed load metering data. In *Sustainable Internet and ICT for Sustainability (SustainIT), 2012*, pages 1–9. IEEE, 2012.
15. Ahmed Zoha, Alexander Gluhak, Muhammad Ali Imran, and Sutharshan Rajasegarar. Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors*, 12(12):16838–16866, 2012.