Accepted at the 29th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2016).

Investigating the Switch Continuity Principle Assumed in Non-Intrusive Load Monitoring (NILM)

Stephen Makonin
Engineering Science, Simon Fraser University
8888 University Drive, Burnaby, BC, V5A 1S6, Canada
Email: smakonin@sfu.ca

***Abstract*—Non-Intrusive Load Monitoring (NILM) researchers have always assumed the switch continuity principle (SCP), which assumes that only one appliance ever changes state at any given point in time. However, SCP cannot be relied upon 100% of the time, especially when unsupervised NILM is used to guess what appliances might be in a house. This principle breaks under certain conditions based on the data capture choices made: number of occupants, number of appliances, measurement unit, measurement precision, and sampling frequency. This paper identifies and explores the conditions under which SCP can and cannot be assumed. This is done through empirical tests performed on two of the most popular datasets used for NILM.**

***Index Terms*—NILM, disaggregation, data analysis, smart meter, load monitoring**

I. Introduction

Non-Intrusive Load Monitoring (NILM), or just simply disaggregation, is a growing research field that began in 1985 with a report [1] written by George W. Hart (at MIT) for Electric Power Research Institute (EPRI). NILM is used to discern what electrical loads (e.g., appliances) are running within a home/building using only the aggregate power meter. Why? To help occupants understand how they and their appliance use energy so that they could conserve to either save money, the environment, or both. Mathematically speaking, disaggregation is the inverse problem of aggregation; e.g., $x = sum(\mathbf{V})$ where x is known and the vector $\mathbf{V}$ is not – but where we have some sort of probabilistic model or feature set to help choose the best values for $\mathbf{V}$ given x.

Hart [2] defined the *switch continuity principle* (SCP) of having the property that "[i]n a small time interval, we expect only a small number of appliances to change state in a typical load" and "we begin with the switch continuity principle as the foundation" for NILM. He continued to write, "we expect the number of appliances which change state to be usually zero, sometimes one, and very rarely more than one". Through empirical methods he found that "[s]imultaneous events, or nearly so within 2-3 seconds, accounted for 4% of the events in one field test..., but this will vary considerably, depending on the appliance inventory and usage". The SCP is a necessary assumption to having NILM algorithms that can actively tune general appliance models into specific models that are house specific [3] (Section II). This is also known as a semi-supervised learning approach and sometimes inaccurately referred to as unsupervised learning (see Section II-A).

The caveat Hart identified in the SCP, appliance inventory and usage can be more specifically defined as: number of occupants, number of appliances, measurement unit, measurement precision, and measurement sampling frequency. The *number of occupants* refers to the the number of occupants that reside within the house. The *number of appliances* refers to the number of appliances that are being disaggregated rather than the the total amount of appliance in the house. The *measurement unit* refers to what measurement is used (e.g., current, power). The *measurement precision* refers to how precise the measurement reading is (e.g., deci-Amps, whole Amps, kilo-Watts). The *measurement sampling frequency* can also mean the sampling rate (e.g., 1Hz, per minute, 100kHz).

Hart's 1992 initial definition of the SCP was at a time when datasets did not exist for NILM researchers. Since that time other NILM researchers have relied only on this principle and have not investigated it further on other data to see if it holds true. Through an empirical data analysis experiment (Section III) on popular datasets (AMPds R2013 [4], AMPds v2 [5], REDD [6]) used by NILM researchers, I show evidence (Section IV) that SCP may not hold true as the percentage of simultaneous appliance state switching has increased since Hart's initial tests. I expand on initial experiments done in my PhD Thesis [7].

II. Unsupervised NILM

A. Defining Learning Types

In the artificial intelligence and machine learning fields there are three main types of learning: supervised, unsupervised, and semi-supervised. Supervised learning uses labelled priors/data to build a model for inference or prediction. Unsupervised learning builds a model without labelled priors/data. Semi-supervised learning builds models using a small amount of labelled data with a large amount of unlabelled data.

If we focus on learning for NILM we would define the following two terms. *Supervised NILM* uses sub-metered appliance data to build a model to then disaggregate. *Unsupervised NILM* uses general appliance models and actively tunes them to specific house appliances. Unsupervised learning is not equivalent to unsupervised NILM because the general appliance models used for training are considered labelled data. This means that unsupervised NILM is a semi-supervised learning problem.

Unsupervised NILM is a difficult problem for three main reasons. Firstly, different sampling rates can lead to feature loss (high-frequency *vs* low-frequency sampling). Secondly, there are vastly different power signatures amongst the same type of appliances. This means that there is no "one general model" for one type of appliance. Thirdly, learning can only happen when an appliance state change occurs, and in the case of unsupervised NILM, only when the state change of a single appliance occurs, which can be very rare.

B. Related Works

Kim et al. [8] used a combination of four factorial hidden Markov model (HMM) variants to provide an unsupervised learning technique. Some factorial HMMs modelled load-state durations while others modelled time-of-day usage. Emission probabilities used Gaussian distributions, which were prone to over fitting. They were not able to use the Viterbi algorithm to infer load states because of the intractability of some factorial HMMs and instead used simulated annealing (SA) [9]. They achieved classification accuracies of between 69%–98% (for 10 homes) using their modified f-score accuracy measure. Their results seem to suggest that the accuracy of the disaggregator quickly decreases as more appliances are added for disaggregation. Such a disaggregation requires a high degree of computational power to disaggregate.

Recently, Johnson et al. [10] considered using the factorial variant of a hidden semi-Markov model (HSMM) [11], [12] because it provided a means of representing state durations in a load model. They introduced the idea of *change-point detection* as a way to rule out observations that would not present a learning opportunity for active tuning or infer a state change. This allowed them to reduce the computational complexity. Although the authors claim this was unsupervised, they were incorrect, as labelled data from the same dataset was used for training and than testing. Their models are specific to a given dataset or house. Rather then having their algorithm run on the entire dataset, specific hand-picked segments of data were used, which does not constitute a real-world scenario.

Guo et al. [13] used explicit duration difference HMMs (EDHMM-diff) and a modified forward-backward algorithm to disaggregate a fridge and clothes dryer. They down sample from 1–3 second sampling to 30 seconds sampling using the REDD dataset [6]. Using a difference model allowed them to actively tune their general appliance models using a detect and re-estimate approach for the difference in spikes in $\Delta y = y_{t-1} - y_t$, where y is the observed aggregate power signal at the previous $t-1$ and current time t. The authors claim that the REDD dataset provides "no reliable ground truth" – this makes it hard to verify the claims of their disaggregator. They also model the fridge and clothes dryer as having just 2 states, OFF and ON. This is not usually the case for clothes dryers, which can have 3–4 states: OFF, STANDBY, DRY, and COOL_DOWN.

The authors above assume Hart's SCP. Section IV shows that the amount of data in datasets that give opportunities for unsupervised learning is very small. Therefore, testing accuracy claims with datasets bring a high amount of skepticism. Parson [3, Fig. 8], on the other hand, has shown how actively tuning the general models of fridges and freezers during the night can better ensure that the SCP can be relied upon; especially for these type of appliances that consume low amounts of power (100–200W) and are cyclical in nature.

III. Experimental Setup

A. Experiment Definition & Algorithm

The experiment's definition is simple. Collect statistics on the amount of times state changes occur and the number of appliances that change state within a given dataset.

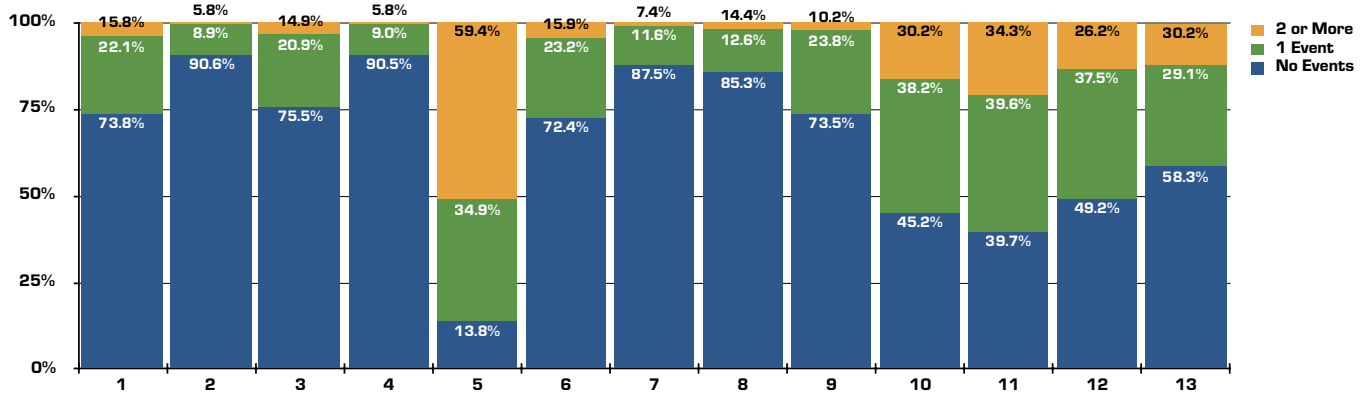
The procedure used in this experiment and to collect statistics is listed in Algorithm 1. Once the dataset is loaded (line 1), each appliance's sub-meter data is used to create a histogram. This histogram is then quantized – meaning that the histogram is algorithmically examined to determine the appliance's states. Significant peaks or spikes in the histogram are treated as distinguishable states – one state corresponding to one peak (lines 2–4). This method of determining appliance states has been used and discussed in previous NILM experiments [7], [14], [15]. Next, for each reading within that dataset (line 5), the state of each appliance is determined for that reading and stored in a list/vector (line 6). The current vector of appliance states is compared to the previous vector of appliance states. If these two vectors are different then one or more appliances have changed state (line 7). If so, the number of appliances that have changed state is recorded (line 8) and resulting statistics appear in Table I. Finally the current vector is stored as the previous vector for comparison to the next dataset reading (line 10). The process is repeated until all readings in the dataset are processed.

Algorithm 1 Calc-Dataset-State-Change-Statistics

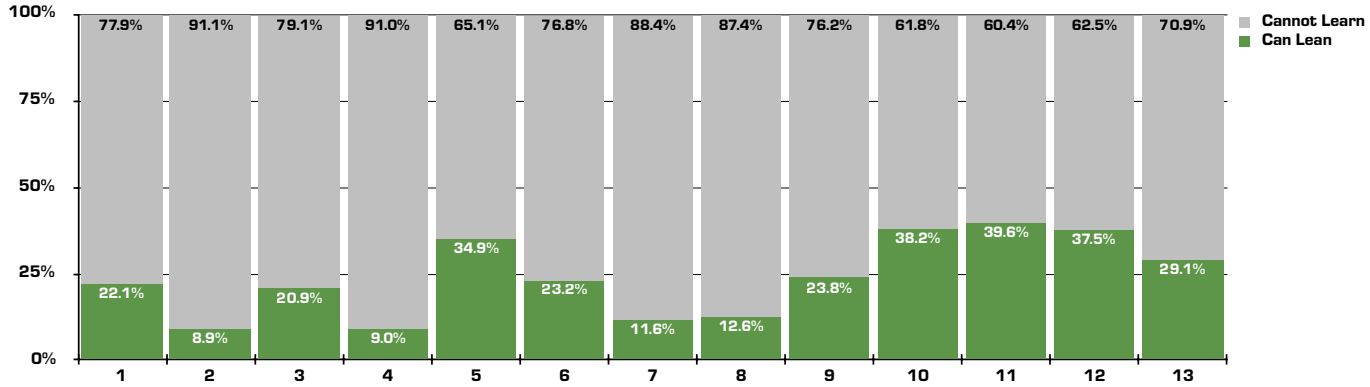
```
1:  data ← load_dataset(dataset_name)
2:  for each appliance do
3:      models[appliance] ← create_model(appliance)
4:  end for
5:  for each reading ∈ data do
6:      states_t ← calc_current_states(reading, models)
7:      if states_t ≠ states_{t−1} then
8:          record number of state changes
9:      end if
10:     states_{t−1} ← states_t
11: end for
```

B. Experimental Datasets

In Hart's tests, multiple loads switching states accounted for 4% of the reading collected at a sampling rate of 2–3 seconds. Since 1992, homes have more energy efficient, multi-state appliances. It would also be interesting to see if this principle holds true for even lower frequency sampled datasets. For the purpose of this test three popular datasets that NILM researchers used were chosen: AMPds R2013 [4], AMPds v2 [5], and REDD [6]. AMPds R2013 is a dataset



(a) Amount of Simultaneous State Changes (Events)



(b) Unsupervised Learning Opportunities

Fig. 1. Stacked bar charts of simultaneous appliance state changes data listed in Table I Chart (a) shows the percentage of the dataset where no state changes occur, one state change occurs, and more than one state change occurs. Chart (b) shows what percentage of the dataset where unsupervised learning can occur. Tests 1 and 2 use data from the AMPds R2013 dataset. Tests 3 to 8 use data from the APMds v2 dataset. The remaining tests use data from each of the six houses in the REDD dataset.

TABLE I
COMPARING THE NUMBER OF LOAD SIMULTANEOUS ON EVENTS

Test ID	Dataset	Unit	Precision	Appliances	No Events	1 Event	2 Events	3 Events	4 Events	5 or More
1	AMPds R2013	Current	dA	19	386,962	115,823	19,212	2,317	214	15
2	AMPds R2013	Current	A	19	474,981	46,699	2,661	197	5	0
3	AMPds v2	Current	dA	19	793,652	219,257	33,724	4,115	411	40
4	AMPds v2	Current	A	19	951,124	94,236	5,456	366	17	0
5	AMPds v2	Power	W	19	145,344	367,387	321,554	154,776	49,087	13,051
6	AMPds v2	Power	daW	19	760,995	243,954	40,750	4,967	492	41
7	AMPds v2	Power	hW	19	919,597	121,857	9,077	640	27	1
8	REDD House 1	Power	W	10	346,950	51,178	6,762	903	578	376
9	REDD House 2	Power	W	8	232,928	75,386	7,536	845	120	24
10	REDD House 3	Power	W	12	169,983	143,861	50,053	10,157	1,646	449
11	REDD House 4	Power	W	11	169,754	169,594	69,020	16,642	2,700	365
12	REDD House 5	Power	W	14	38,096	29,032	8,766	1,385	134	37
13	REDD House 6	Power	W	11	112,108	55,891	21,221	2,624	227	120

that has a sampling rate of once per minute and contains one year's worth of data. AMPds v2 is also sampled at once per minute but contains two year's worth of data. REDD has a sampling rate of every 2-3 seconds and contains approximately one month's worth of data for six houses – the number of occupants is not known. AMPds has only one house, R2013 has three occupants and v2 has four occupants.

IV. FINDINGS & RESULTS

A series of 13 tests were performed on the three datasets identified above. Previously, I introduced five caveats to SCP: number of occupants, number of appliances, measurement unit, measurement precision, and measurement sampling frequency. For this analysis refer to Table I and Figure 1(a).

Although my tests were not able to test how the number

of occupants affects SCP, I think that is it easily understood that this can be the case – the more occupants the more simultaneous activity there is. For the number of appliances, when comparing real power tests between AMPds and REDD it is clear that AMPds (which has many more appliances) has a greater amount of simultaneous appliance state changes.

The measurement unit chosen has a curious effect on SCP. When we compare the measurement of current (Tests 3 and 4) *vs* real power (Tests 5, 6, and 7) we can see that Test 5 (in watts) has a very large percent of simulations events (59.4%) compared to Test 4 (in amps, 5.8%). However, the results are very close if we compare Test 3 (in deci-amps) and Test 6 (in deca-watts). This is due to how the meter rounds off measured numbers – there is more precision in the measurement of power than current due to the simple fact that $P = V \times I$ [16].

The lower the measurement precision, the lower the percentage of simultaneous events (see Test 1 *vs* Test 2, Test 3 *vs* Test 4, and Test 5 *vs* Test 6 *vs* Test 7) that have a positive effect on SCP. However, there would be a decrease in the accuracy of appliance consumption estimation. I did run tests that compared each of the six REDD houses using different precision (W *vs* daW) but found that the results were identical. This might be due to the fact that REDD only has about one month of data capture and that the houses used seem to be smaller. Another observation is the little difference in the results of one year of capture (Tests 1 and 2) *vs* two years of capture (Tests 3 and 4). This seems to suggest that the variations in appliance usage have been captured over these long periods of time.

Measurement sampling frequency has negative affects on SCP as the lower the sampling frequency, the more appliances will have changed their state within the sampling period. Again, no direct tests were possible, but this is a well understood fact.

The above results show that the low amount of simultaneous events (4%) that Hart found in his field test [2] was not achievable in the datasets I tested. The best that could be achieved was 5.8% in Tests 2 and 4. The next best was 7.4% in Test 7. Averaging the test results shows that 20.8% of the time simultaneous events (more than one load switching state) will occur. This may very well be due to the fact that houses contain more appliances that are multi-state and are more energy efficient.

Finally, Figure 1(b) shows that learning opportunities for unsupervised NILM algorithms are rare in these datasets only occurring on average 24% of the time. A learning opportunity is defined as a point in time were there is only one appliance that changes state. This means that unsupervised NILM algorithms would need to learn from datasets that have long-term captures of house and appliance data to be able to report credible testing accuracy and performance results.

V. Conclusions

This paper has presented experimental results that show Hart's *switch continuity principle* can still hold true, in some cases (small homes/apartments). The small amount of simultaneous events (4%) that Hart found in 1992 is no longer the case. I have found that it has increased to 20.8% on average especially when there are a large number of appliances in a home. With this increase in appliances the mix and profile of appliances within many homes has changed over the last 24 years. For example, the increase in multi-state, energy efficient appliances. The experimental results also show that in the case of unsupervised NILM algorithms only 24% (on average) of a dataset can be relied upon for learning where there is only one appliance that changes state. To truly test the robustness of unsupervised NILM, large long-running datasets are needed because the number of learning opportunities are rare. This rarity makes it hard to claim solid testing accuracy and performance of unsupervised NILM algorithms.

References

[1] G. Hart, "Prototype Nonintrusive Appliance Load Monitor," MIT Energy Laboratory and Electric Power Research Institute Technical Report, Tech. Rep., September 1985.

[2] ——, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[3] O. Parson, S. Ghosh, M. Weal, and A. Rogers, "An unsupervised training method for non-intrusive appliance load monitoring," *Artificial Intelligence*, vol. 217, pp. 1–19, 2014.

[4] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajic, "AMPds: A Public Dataset for Load Disaggregation and Eco-Feedback Research," in *Electrical Power and Energy Conference (EPEC), 2013 IEEE*, 2013, pp. 1–6.

[5] S. Makonin, B. Ellert, I. V. Bajic, and F. Popowich, "Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014," *Scientific Data*, vol. 3, no. 160037, pp. 1–12, 2016.

[6] J. Kolter and M. Johnson, "Redd: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, 2011.

[7] S. Makonin, "Real-time embedded low-frequency load disaggregation," Ph.D. dissertation, Simon Fraser University, School of Computing Science, 2014.

[8] H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements." in *SDM*, vol. 11. SIAM, 2011, pp. 747–758.

[9] R. Kohlenberg, T. Phillips, and W. Proctor, "A behavioral analysis of peaking in residential electrical-energy consumers1," *Journal of Applied Behavior Analysis*, vol. 9, no. 1, pp. 13–18, 1976.

[10] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-markov models," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 673–701, 2013.

[11] V. Barbu and N. Limnios, "Hidden semi-markov model and estimation," in *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*, ser. Lecture Notes in Statistics. Springer New York, 2008, vol. 191, pp. 1–48.

[12] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 12 1966.

[13] Z. Guo, Z. J. Wang, and A. Kashani, "Home appliance load modeling from aggregated smart meter data," *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 254–262, 2015.

[14] S. Makonin, F. Popowich, I. Bajic, B. Gill, and L. Bartram, "Exploiting hmm sparsity to perform online real-time nonintrusive load monitoring," *Smart Grid, IEEE Transactions on*, vol. PP, no. 99, pp. 1–11, 2015.

[15] S. Makonin, I. V. Bajic, and F. Popowich, "Efficient Sparse Matrix Processing for Nonintrusive Load Monitoring (NILM)," in *NILM Workshop 2014*, 2014.

[16] S. Makonin, W. Sung, R. Dela Cruz, B. Yarrow, B. Gill, F. Popowich, and I. V. Bajic, "Inspiring energy conservation through open source metering hardware and embedded real-time load disaggregation," in *Power and Energy Engineering Conference (APPEEC), 2013 IEEE PES Asia-Pacific*. IEEE, 2013, pp. 1–6.