

## RESEARCH

# How unique is weekly smart meter data?

Dejan Radovanovic <sup>1\*</sup>, Andreas Unterweger <sup>1</sup>, Günther Eibl <sup>1</sup>, Dominik Engel <sup>1</sup> and Johannes Reichl <sup>2</sup>

---

\*Correspondence:

[dejan.radovanovic@fh-salzburg.ac.at](mailto:dejan.radovanovic@fh-salzburg.ac.at)

<sup>1</sup>Center for Secure Energy Informatics, Puch bei Hallein, Austria

Full list of author information is available at the end of the article

## Abstract

In many countries, energy consumption data is collected through smart meters in 15-minute intervals. Prior work has shown that one year's worth of this data is sufficient to extract sensitive information about households. In this **short paper**, we break down energy consumption data from a novel dataset into **one-week snippets**. Using **off-the-shelf algorithms**, we assess whether it is possible to clearly identify (i.e., fingerprint) an individual household only by its energy consumption from a **one-week period**. More generally, we ask whether an attacker can distinguish one household from a group of others by its energy consumption from only one week's worth of data. We find that a small number of households exist for which the weekly consumption is so unique that it can be distinguished almost always amidst weekly data from dozens of other households. Furthermore, a large number of households can be distinguished with surprisingly high accuracy and an order of magnitude better than guessing. We discuss the potential impact of these findings on the privacy of smart meter datasets with respect to de-anonymization and re-identifiability.

**Keywords:** smart meter; privacy; data analysis; attack

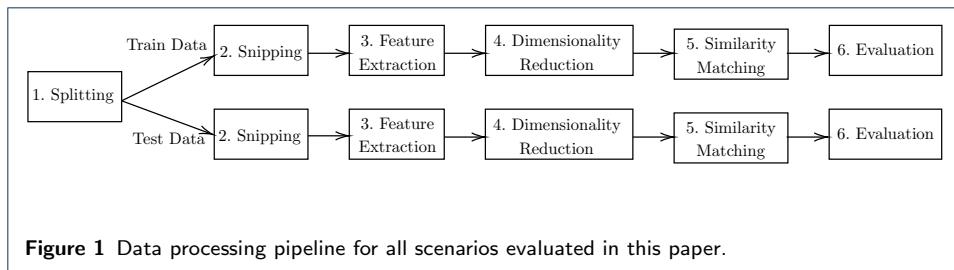
## Introduction

Conventional electricity meters are replaced by smart meters in many countries around the world. Smart meters collect and transmit energy consumption data, also referred to as load data, in regular intervals, e.g., 15 minutes in the European Union [1]. Aside from its intended uses such as billing, the collected energy consumption data of a household has been shown to be a privacy concern [2, 3, 4, 5].

Several privacy attacks based on the total consumption data of a household exist. Detection of human behavior using non-intrusive appliance detection methods (or variants) was studied in [6, 7, 8]. Occupancy was detected in [9, 10, 11, 12], holidays were detected in [13]. Approaches to detect air conditioners in low-resolution energy consumption data were proposed by [14]. Methodologies for the detection of swimming pools from 15-minute consumption data were introduced in [15] and [16].

While these privacy attacks primarily aim at an extraction of behaviors or appliances, the goal of this paper is to identify a household based on its **weekly load signature**. This enables an attacker to find a single household in a different, anonymized dataset based on only **one week of captured load data** and to break privacy by combining information from different datasets.

Consider the following scenario: There already exist datasets about consumer appliances and behavior, such as the ones described in [17, 18, 19, 20]. These datasets are anonymized and associate months' or years' worth of load data of a household with corresponding, privacy-critical information, e.g., income, appliance use, habits etc. If an attacker were to capture only one week's worth of load data



from a household or smart meter elsewhere, this attacker could find the household within one of the existing datasets and link the available information. This would undermine the corresponding individuals' rights to privacy according to the GDPR [21], despite the load datasets being anonymized under the implicit assumption that consumption data, by itself, is insufficient to identify or link households.

Thus, in this paper, it is assessed whether **one week's worth of energy consumption data** from a household is enough to find other weekly consumption data of the same household in a larger, anonymized dataset. The households for which this is possible can be linked and thus be de-anonymized. In addition, they can be distinguished from other households. It is assessed for how many and which households this is possible, and which accuracy can be achieved with **off-the-shelf algorithms**.

This paper is structured as follows: In the [Methodology](#) Section, the dataset and the methodology to distinguish households are described. In the [Results](#) Section, different scenarios are evaluated and the results on how well households can be distinguished in each scenario are presented and analyzed. Finally, the paper is concluded in the [Conclusion and Outlook](#) Section and an outlook to future work is given.

## Experimental Setting & Methodology

In this section, we describe how to distinguish one household from a group of others by its energy consumption from only **one week's worth of data**. We first describe our novel dataset before providing a step-by-step description of our methodology.

### Data

The used data stems from a field test collecting electricity consumption profiles of 1,589 suburban households from Upper Austria via smart meters between 1st of May 2017 until October 15th 2018. The field test aimed at testing various incentive schemes for motivating electricity consumers to shift loads towards times of high renewable production. The study design as well as the procedure of recruiting was done in accordance with the Austrian data protection agency, and the respective communication can be requested from the data protection agency under reference number DSB-D036.500/0005-DSB/2017.

The used data contains consumption values in 15-minute intervals as well as household and demographic data. To put the characteristics of the suburban households into perspective, the following statistics are illustrated: The yearly average energy consumption per household is 5163 kWh, with the median being 4256 kWh. The

average household size in square meters is 138 (mean) and 130 (median), respectively with 2.8 and 2 residents per household, respectively.

Since a large fraction of households do not have smart meter data from the whole time range, only data from a common time frame of 52 weeks or about one year (October 1, 2017 to September 29, 2018) is used. This ensures the presence of all seasons as well as comparability. Similarly, only households with one smart meter<sup>[1]</sup> are used, limiting the maximum number of used households to 721.

## Methodology

From the available 721 households, subsets of 50 and 200 are sampled randomly for the two evaluation scenarios with 25 and 100 households, respectively. It is important to note that the problem at hand is a multi-class classification problem: the goal is to identify a single household among a group of households based on measurements from one single week. The difficulty of such a multi-class classification increases with the size of the group to which the household belongs. This is reflected by the low accuracy reached by random guessing which is only 4% and 1% for the cases with 25 and 100 households, respectively.

Fig. 1 depicts the proposed data processing pipeline for both scenarios. The pipeline consists of the following steps:

- 1 **Splitting:** The households are split into two distinct sets – a training and a test set. The training set is only used to learn data characteristics which are used to evaluate the test set. This strategy ensures generalizability by avoiding to evaluate the specific characteristics of one large dataset. The assignment of households to the training and test set is performed randomly so that both sets are of equal size (50/50 split), which is the most difficult setting and allows to assess the generalizability of the proposed approach [22]. All subsequent steps are performed on the training and test set separately.
- 2 **Snipping:** The energy consumption data of each household is separated into one-week snippets, i.e., seven days' worth of 15-minute energy consumption values. For 100 households, this yields a total of 5200 separate one-week consumption snippets.
- 3 **Feature Extraction:** For each one-week snippet, 787 off-the-shelf numerical characteristics (statistical features, e.g., mean and standard deviation as well as frequency-transform features, e.g., Fourier and Wavelet coefficients) are computed using *tsfresh*<sup>[2]</sup>. These time-series characteristics (extracted features) are generic and not specific to energy consumption data. Any feature that could yield an undefined expression (not a number), e.g., when dividing zero by zero, is removed. Effectively, 751 numerical characteristics (features) are computed for each one-week snippet.
- 4 **Dimensionality reduction:** The next step differs between the training and the test sets. For the training set, a Principal Component Analysis (PCA) [23] is performed for all snippets at once. This reveals which snippet characteristics

---

<sup>[1]</sup>In this geographic region, it is typical for households with electric water heating, electric vehicle charging facilities and other special consumers to have additional smart meters.

<sup>[2]</sup><https://tsfresh.readthedocs.io>, accessed on April 3, 2022

(features) are likely to be relevant for distinguishing snippets of different households. In the test set, for each weekly snippet, the  $p$  dominant characteristics determined during training are extracted. Thus, each weekly snippet is reduced to a  $p$ -dimensional numerical feature vector which represents a fingerprint of the corresponding weekly consumption and possibly the household. The reduction to  $p$  characteristics is also performed for each weekly snippet of the training set, but for verification purposes only (see below).

- 5 **Similarity matching:** For each weekly snippet, the  $k$  most similar snippets are determined using the previously extracted  $p$  characteristics. Households are considered more similar, the smaller the Euclidean distance between the  $p$ -dimensional numerical feature vectors is. Note that using Cosine similarity or Manhattan distance instead of Euclidean distance does not change the results significantly.
- 6 **Evaluation:** For each weekly snippet and its  $k$  most similar snippets, the corresponding households are revealed to assess performance. If the household corresponding to a given weekly snippet is the same as the household which occurs most often in the  $k$  most similar snippets, the matching is considered successful, i.e., snippets from the correct household have been found. Otherwise matching is considered to be unsuccessful. For all weekly snippets of a household  $h$ , the number of successful matches  $s_h$  and unsuccessful matches  $u_h$  determine the overall per-household matching accuracy  $acc_h = \frac{s_h}{s_h + u_h}$ . If multiple candidates occur equally often, one is chosen at random.

## Results

The methodology described in the [Methodology](#) Section is applied to two different scenarios: subsets of 25 and 100 households representing a small and a medium-sized residential area, respectively. Before discussing the results, we describe the parameter selection for our proposed method for the two scenarios.

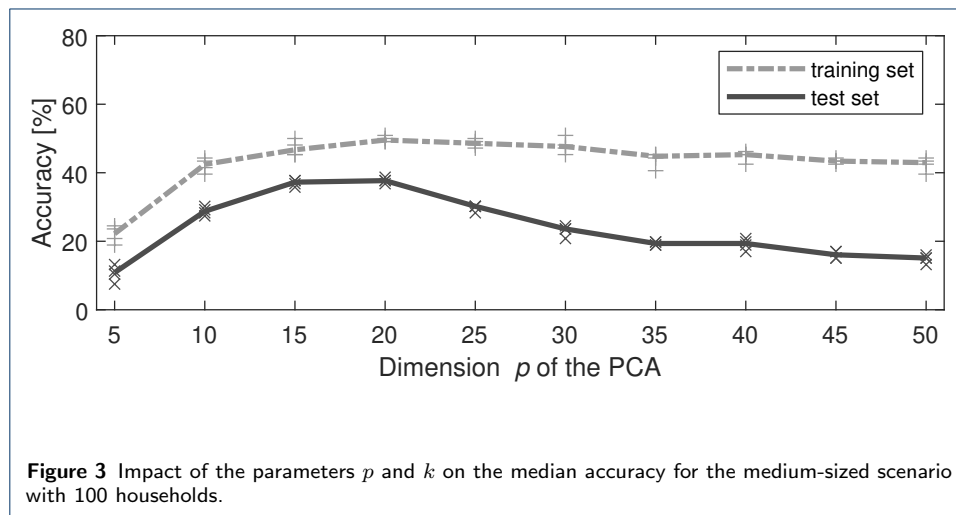
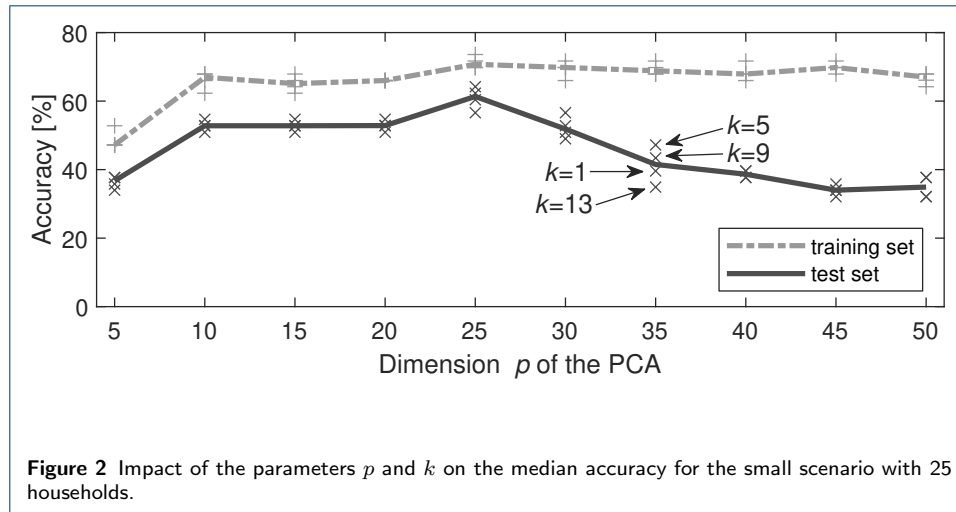
### Parameter Selection

The selection of parameters is described and demonstrated for the small residential area consisting of 25 households. The medium-sized residential area consisting of 100 households is derived analogously.

Only two parameters need to be selected –  $p$ , the number of dimensions used for the fingerprint (step 4 in [Figure 1](#)), and  $k$ , the number of neighbors considered during matching (steps 5 and 6).  $p$  is varied between 5 and 50, and  $k$  is varied between 1 and 15. [Figure 2](#) illustrates the median accuracy (Y axis) over all weeks of all households with respect to the different values of  $p$  and  $k$ .

As can be seen, the dependency of the accuracy on  $p$  (X axis) is much more pronounced than on  $k$  (pluses for the training set and crosses for the test set, respectively) for any particular value of  $p$ . For the sake of visibility, only  $k = 1, 5, 9, 13$  are depicted as single points. The fact that all four points are very close to each other for all values of  $p$  shows that the dependency on  $k$  is weak. Thus,  $k = 1$  neighbor is chosen for sake of simplicity.

However, the effect of  $p$  is significant: [Figure 2](#) shows that  $p = 5$  dimensions are too small since the accuracies of both, the training and the test set are small. This

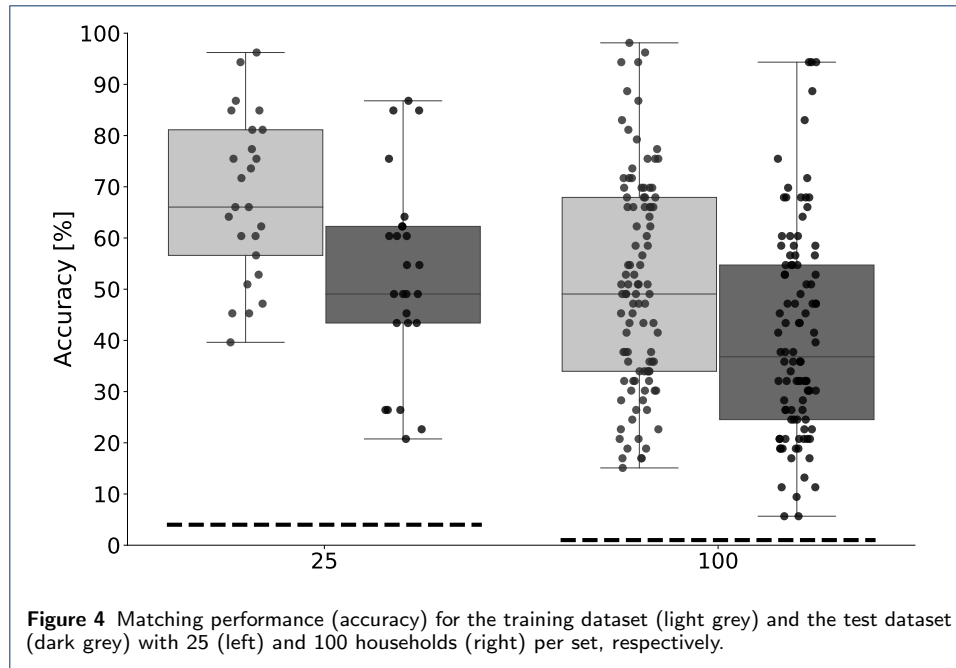


indicates that not enough of the available information is used to distinguish different households. For  $10 \leq p \leq 20$ , the accuracies of both, the training set (dash-dotted, light grey line) and test set (solid, dark grey line), increase. For larger values of  $p$ , the training accuracy stays high, but the test accuracy drops compared to the training set. This indicates that the features learned during training are mostly specific to the households of the training set and do not generalize to the households of the test set.

Thus, a value of  $p = 25$  is used for the small scenario with 25 households. Analogously, a value of  $p = 20$  is used for the medium-sized scenario with 100 households as can be seen from Figure 3. The value of  $k = 1$  is used for both scenarios.

### Matching Performance

In this section, the matching performance achieved with the parameters selected in the previous section is assessed. Figure 4 depicts the per-household matching accuracy for the training set (light grey) and the test set (dark grey) for both scenarios (25 and 100 households per set, respectively). The black dots illustrate the individual per-household matching accuracy.



The overall accuracy within the test set (dark grey) is surprisingly high, considering the simplicity of the approach and the difficulty of the corresponding classification problem with 25 and 100 classes, respectively. For reference, guessing the correct household (class) randomly is expected to yield an accuracy of  $acc_{h,rand} = 1/n_{Test}$ , i.e., 4% and 1% for a 25 and a 100-household-sized set, respectively. This reference (guessing) accuracy is depicted as thick dashed lines in Figure 4.

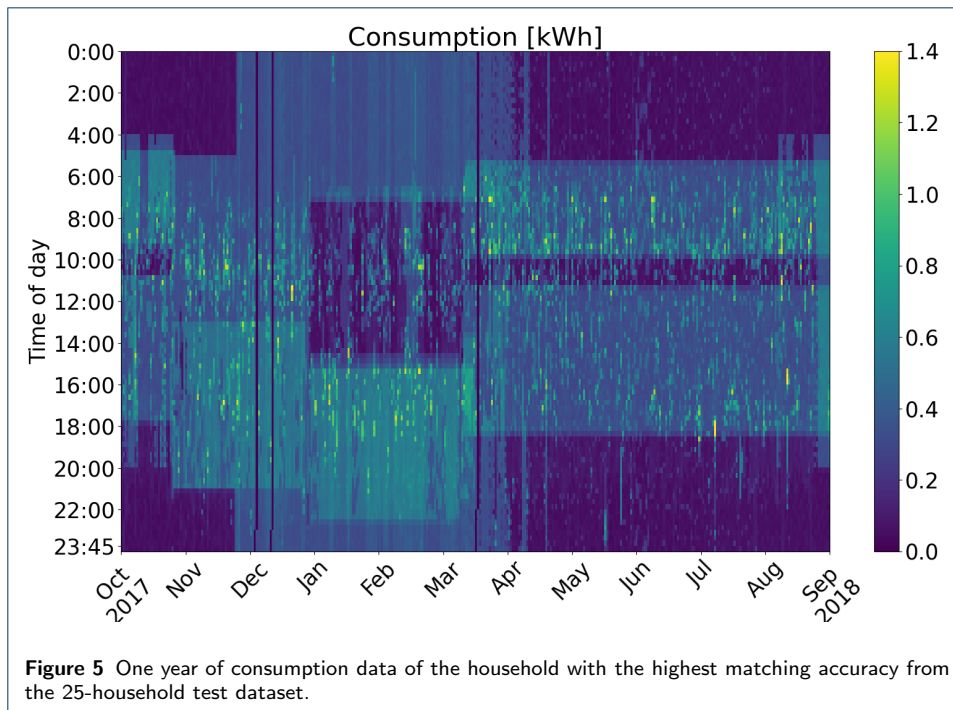
Compared to random guessing, the median accuracy of the proposed methodology is between roughly 16 and 35 times higher on average for the small and medium-sized residential areas, respectively. Note that the difficulty of the problem increases with the number of households. This explains why the accuracy is lower for the case of a medium-sized residential area compared to the small residential area. Yet, the performance of the proposed methodology is significantly better in the medium-sized case relative to guessing.

The black dots in Figure 4 depict the matching accuracies of the individual households. For some households, the accuracy is nearly 100% which implies that the corresponding household can be identified based on an **arbitrary single week** of a year. This is surprising as one would expect the seasonal differences to have a significant impact on the consumption patterns throughout the weeks of a year.

The subsequent privacy implication is that some households exist which can be identified very easily from a single, **arbitrary week's** worth of energy consumption with an approach that uses **off-the-shelf algorithms**. A number of other households cannot be detected well, i.e., they have a low matching accuracy. However, the matching accuracy for these households is still much better than guessing.

### Extreme Households

The question arises what makes the identification of a household easier or harder, i.e., why the matching accuracy is relatively high or relatively low, respectively. As a first attempt to answer this question, a preliminary descriptive analysis is provided.

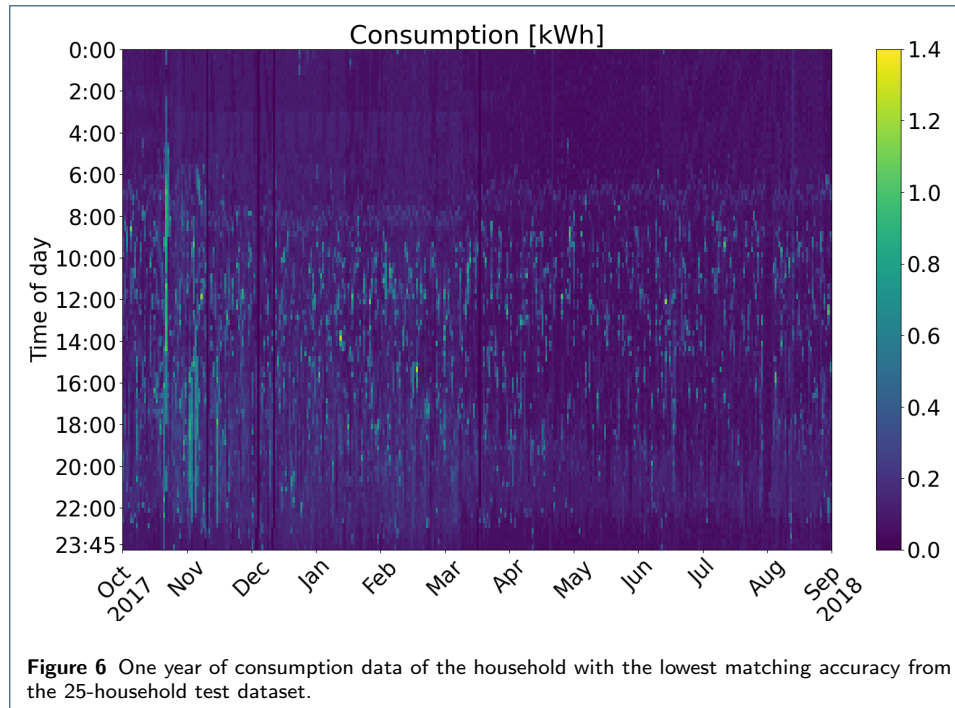


Based on the matching results, the most extreme households, i.e., those with the highest and the lowest matching accuracy, are visualized. The consumption data of a whole year of a household is illustrated as a heatmap. The X axis denotes the days of the year from left to right, the Y axis denotes the time of day from top to bottom in intervals of 15 minutes. The color of each 15-minute interval depicts the associated energy consumption in kWh. Dark (purple) represents 0 kWh and bright (yellow) represents 1.4 kWh.

Figure 5 shows the energy consumption for the household with the highest matching accuracy within the test set of the small residential area. One can see that the consumption is quite regular, i.e., the consumption barely changes between weeks of the periods from April to November, and December to March, respectively. Note that the apparent one-hour time shifts in March and October are mostly likely due to daylight saving time.

The regular rectangular areas might be from a pool pump as proposed in [15]. While this pattern is not the same throughout the whole year, it is comparatively regular over periods of multiple weeks. This suffices as the proposed methodology only needs to find one of the few similar weeks. The identifiability seems to be related to periodic behavior due to the dominance of Fourier and Wavelet features but requires further investigation in future work.

Figure 6 shows the household with lowest matching accuracy within the test set of the small residential area. While its consumption is comparatively regular over the year, it does not show any remarkable features which appear over multiple consecutive weeks. Thus, with the proposed methodology, any given week of this household shares more similarities with weeks from other households than it does with weeks from the same household.



The extreme households of the medium-sized residential area exhibit similar characteristics to those of the small residential area described above. For the sake of completeness, the corresponding heatmaps are visualized in Figures 7 and 8.

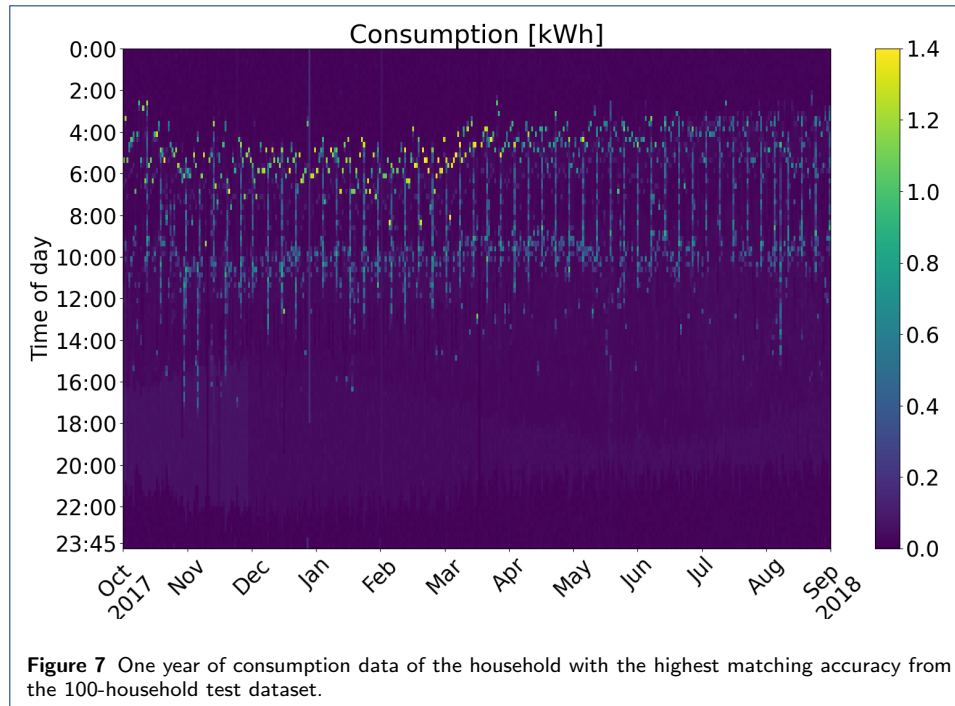
Note that this analysis is a first attempt of an explanation. Future analyses might offer further insight into the relevant household-specific characteristics which impact matching accuracy.

## Conclusion and Outlook

This paper shows that the consumption data of a single week can, in principle, be used to identify a household in a larger database of load profiles. Even if these consumption data are anonymized, it is possible to de-anonymize a relatively large portion of households with surprisingly high accuracy, i.e., about an order of magnitude better compared to guessing. This implies that the anonymization of existing and currently collected energy consumption data may be insufficient to protect the privacy of the individuals living in the corresponding households. A preliminary analysis was performed to determine which households are especially susceptible or immune to this type of attack. Future work will need to investigate this in more depth.

Similarly, several other questions are not tackled yet: How far does the accuracy decrease in an even bigger dataset, e.g., one representing a city? Can the proposed methodology be generalized to other datasets and/or different kinds of areas, e.g., rural vs. urban areas? How does the time granularity influence the de-anonymization probability, e.g., one day's worth of data vs. one month's worth of data? Finally, the proposed methodology is relatively simple and uses only off-the-shelf algorithms. It might easily be improved by using more sophisticated models like time-series-specific auto-encoders and deep neural networks.





#### Funding

Funding from the Federal State of Salzburg, the Austrian Research Promotion Agency (FFG project number 881165) and the European Union's Horizon 2020 research and innovation programme under the PEAKapp project grant agreement number 695945 is gratefully acknowledged.

#### Availability of data and materials

The data used in this paper can be requested from the Austrian data protection agency under reference number DSB-D036.500/0005-DSB/2017 .

#### Author's contributions

JR provided the data and performed data pre-processing. DR and AU designed and implemented the data processing pipeline. GE, AU and DR designed the test methodology and performed the evaluation. AU and GE wrote the Introduction Section. DR, AU and JR wrote the Methodology Section. GE, DR and AU wrote the Results and Conclusion and Outlook Sections. DE performed editorial work and provided valuable suggestions for improvements. The overall contributions are DR (40%), AU (30%), GE (20%), DE (5%) and JR (5%).

#### Competing interests

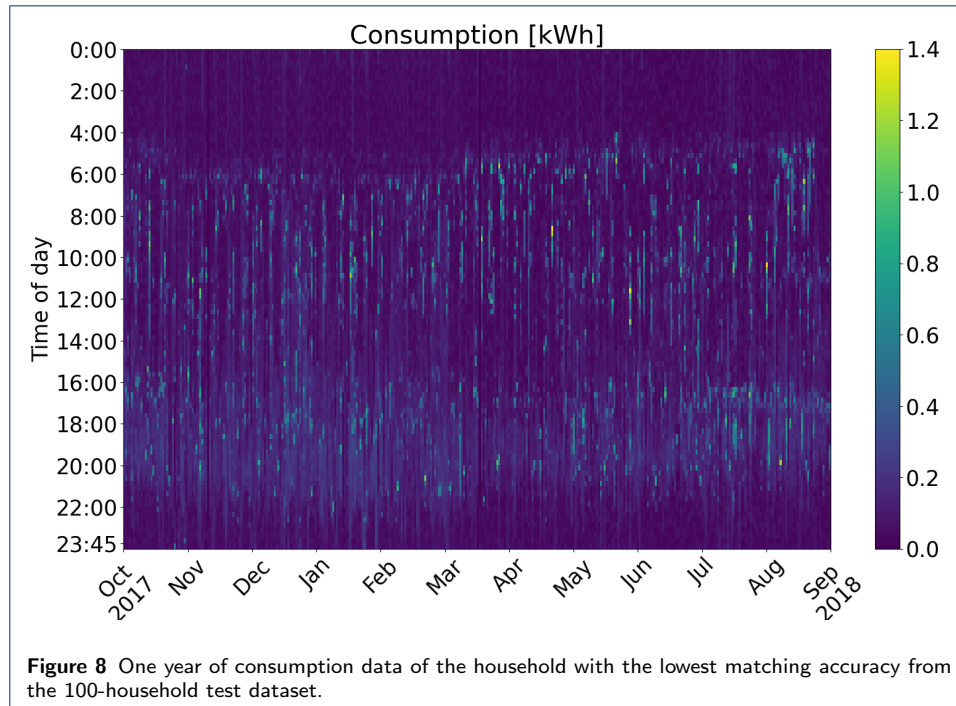
The authors declare that they have no competing interests.

#### Author details

<sup>1</sup>Center for Secure Energy Informatics, Puch bei Hallein, Austria. <sup>2</sup>Energieinstitut an der Johannes Kepler Universität Linz, Linz, Austria.

#### References

- European Commission: 2012/148/EU: Commission Recommendation of 9 March 2012 on preparations for the roll-out of smart metering systems (2012). <http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32012H0148>
- Kim, H., Marwah, M., Arlitt, M.F., Lyon, G., Han, J.: Unsupervised disaggregation of low frequency power measurements, pp. 747–758 (2011)
- Kolter, J.Z., Jaakkola, T.: Approximate inference in additive factorial hmms with application to energy disaggregation. *Journal of Machine Learning Research - Proceedings Track* **22**, 1472–1482 (2012)
- Fan, Z., Kulkarni, P., Gormus, S., Efthymiou, C., Kalogridis, G., Sooriyabandara, M., Zhu, Z., Lambotharan, S., Chin, W.H.: Smart grid communications: Overview of research challenges, solutions, and standardization activities. *IEEE Communications Surveys and Tutorials* **15**, 21–38 (2013). doi:[10.1109/SURV.2011.122211.00021](https://doi.org/10.1109/SURV.2011.122211.00021)
- Wang, Y., Chen, Q., Kang, C.: *Smart Meter Data Analytics: Electricity Consumer Behavior Modeling, Aggregation, and Forecasting*. Springer, ??? (2020)
- Lisovich, M., Mulligan, D., Wicker, S.: Inferring personal information from demand-response systems. *IEEE Security & Privacy* **8**, 11–20 (2010). doi:[10.1109/MSP.2010.40](https://doi.org/10.1109/MSP.2010.40)
- Molina-Markham, A., Shenoy, P., Fu, K., Cecchet, E., Irwin, D.: Private memoirs of a smart meter, pp. 61–66. *ACM*, ??? (2010). doi:[10.1145/1878431.1878446](https://doi.org/10.1145/1878431.1878446). <http://doi.acm.org/10.1145/1878431.1878446>



8. Eibl, G., Engel, D.: Influence of data granularity on smart meter privacy. *IEEE Transactions on Smart Grid* **6**, 930–939 (2015). doi:[10.1109/TSG.2014.2376613](https://doi.org/10.1109/TSG.2014.2376613)
9. Chen, D., Barker, S., Subbaswamy, A., Irwin, D., Shenoy, P.: Non-intrusive occupancy monitoring using smart meters, pp. 1–8 (2013). doi:[10.1145/2528282.2528294](https://doi.org/10.1145/2528282.2528294)
10. Kleiminger, W., Beckel, C., Staake, T., Santini, S.: Occupancy detection from electricity consumption data, pp. 1–8 (2013). doi:[10.1145/2528282.2528295](https://doi.org/10.1145/2528282.2528295). <http://dl.acm.org/citation.cfm?doid=2528282.2528295>
11. Kleiminger, W., Beckel, C., Santini, S.: Household occupancy monitoring using electricity meters, pp. 975–986 (2015). doi:[10.1145/2750858.2807538](https://doi.org/10.1145/2750858.2807538)
12. Tang, G., Wu, K., Lei, J., Xiao, W.: The meter tells you are at home! non-intrusive occupancy detection via load curve data, pp. 897–902 (2015). doi:[10.1109/SmartGridComm.2015.7436415](https://doi.org/10.1109/SmartGridComm.2015.7436415)
13. Eibl, G., Burkhart, S., Engel, D.: Insights into unsupervised holiday detection from low-resolution smart metering data. *Information Systems Security and Privacy*, 281–302 (2018). doi:[10.5220/0006719704770486](https://doi.org/10.5220/0006719704770486)
14. Pathak, N., Lachut, D., Roy, N., Banerjee, N., Robucci, R.: Non-intrusive air leakage detection in residential homes, pp. 1–10. *ACM*, ??? (2018)
15. Burkhart, S., Unterweger, A., Eibl, G., Engel, D.: Detecting swimming pools in 15-minute load data, pp. 1651–1655. *IEEE*, ??? (2018). doi:[10.1109/TrustCom/BigDataSE.2018.00244](https://doi.org/10.1109/TrustCom/BigDataSE.2018.00244)
16. Ferner, C., Eibl, G., Unterweger, A., Burkhart, S., Wegenkittl, S.: Pool detection from smart metering data with convolutional neural networks. *Energy Informatics* **2**, 1–9 (2019). doi:[10.1186/s42162-019-0097-8](https://doi.org/10.1186/s42162-019-0097-8)
17. Azarova, V., Engel, D., Ferner, C., Kollmann, A., Reichl, J.: Transition to peak-load-based tariffs can be disruptive for different groups of consumers. *Nature Energy* **4**, 829–830 (2019). doi:[10.1038/s41560-019-0479-y](https://doi.org/10.1038/s41560-019-0479-y)
18. Kolter, J.Z., Johnson, M.J.: Redd: A public data set for energy disaggregation research, pp. 1–6 (2011)
19. Beckel, C., Kleiminger, W., Cicchetti, R., Staake, T., Santini, S.: The eco data set and the performance of non-intrusive load monitoring algorithms, pp. 80–89 (2014)
20. Barker, S., Mishra, A., Irwin, D., Cecchet, E., Shenoy, P., Albrecht, J.: Smart\*: An open data set and tools for enabling research in sustainable homes, p. 108 (2012)
21. European Parliament and Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>
22. Ojala, M., Garriga, G.C.: Permutation tests for studying classifier performance. *The Journal of Machine Learning Research* **11**, 1833–1863 (2010)
23. Jolliffe, I.: *Principal Component Analysis*. Springer, ??? (2002). doi:[10.1007/978-3-642-04898-2\\_455](https://doi.org/10.1007/978-3-642-04898-2_455)