INCREASING THE ACCURACY OF LOW-RESOLUTION COMMERCIAL SMART HEAT METER DATA AND ANALYSING ITS ERROR

Markus Schaffer[1], Daniel Leiria [1],J. Eduardo Vera-Valdés[2], and Anna Marszal-Pomianowska[1]
[1]Department of the Built Environment, Aalborg University, Aalborg, Denmark
[2]Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark

Abstract

Recent research has demonstrated the fundamental potential of smart heat meter (SHM) data. However, it has also been shown that the usability of the data is reduced because SHM energy measurements are commonly rounded down (truncated) to kilowatt-hour values. This study therefore investigates, for the first time, the error introduced by truncation using a high-resolution dataset. Furthermore, a method is developed to reduce the loss of information in the truncated data by combining smoothing with a ruleset and scaling approach (SMPS). SMPS is shown to increase the pointwise accuracy and correlation of the truncated data with the full-resolution data.

Introduction

In the last years, the field of the built environment has been witnessing a transition as a large amount of data from the building stock has become available. Consequently, the debate in the building sector has moved from ”whether” to ”how” to use the gathered data. The district heating (DH) sector is no stranger to this transition, with smart heat meters (SHMs) (remotely readable heat meters) being mandatory for every building within the European Union (EU) connected to district heating from 2027 on (European Parliament, 2018). Already yet, data measured by SHMs are widely available for large shares of the building stock, and their high potential for a large variety of purposes has become evident from current research (e.g. do Carmo and Christensen, 2016; Gianniou et al., 2018a,b; Kristensen et al., 2018; Calikus et al., 2019; Kristensen et al., 2020; Leiria et al., 2021). However, research also clearly indicates that the low transmitted measurement resolution from commercial SHMs is a hurdle for its use (Kristensen et al., 2018; Hedegaard et al., 2020; Hauge Broholt et al., 2022; Leiria et al., 2022).

Commercial SHMs transmit all data at a significantly lower resolution than the actual measurement resolution (Figure 1) to reduce the bandwidth required (Schaffer et al., 2022) and to match the billing model used by the DH utilities. This means that the combined space heating and domestic hot water (DHW) usage is transmitted as cumulative values rounded down to the greatest integer value less than or equal to the cumulative value (Kristensen et al., 2018; Schaffer et al., 2022). For example, any value between 1.0 and $1.\overline{9}$ is transmitted as 1.0. This can therefore also be described as truncating the value to integers, i.e. truncating the decimal values. Both terms, truncation and rounding, will be used interchangeably in the remainder of this paper. This low transmitted resolution introduces a considerable relative uncertainty for data on a high granularity (e.g., small-scale customers, such as apartments or single-family houses) where the heating demand is low. As a result, this requires decreasing either the spatial or the temporal granularity/resolution of the data (Kristensen et al., 2018), which might reduce the value of the data and knowledge gain. No similar research efforts attempting to address this or a similar problem through approaches other than reducing granularity or temporal resolution could be identified. It is assumed that this is because it is a specific problem for an emerging area of research.

For this reason, this study analyses, for the first time, the truncation error based on a high-resolution SHM dataset. An approach is then developed to increase the usability of truncated data by partially recovering the true underlying trend of the data. The results of this new approach are evaluated with high resolution sub-meter data (ground truth). The knowledge gained about the error, combined with the developed method, aims to increase the value and usability of commercial SHM data for research and industrial applications in the building and DH sectors.

Method

As noted above, measured energy consumption is transmitted by commercial SHMs as cumulative values in kilowatt-hours rounded down to integers, and thus has an accuracy of $-1.0\,\mathrm{kWh}$ (Figure 1). However, the most common analyses do not use cumulative values, but hourly energy use calculated as a first-order difference (the difference between a value and its predecessor). In the remainder of the paper, energy use refers to non-cumulative data. Calculating energy use reduces the accuracy of the data to $\pm1.0\,\mathrm{kWh}$. However, not only does this reduce the theoretical accuracy, but the truncation can also obscure the actual pattern of energy use by introducing on/off like patterns (Figure 1 bottom right).

High-resolution dataset description

A high resolution dataset is used to analyse the truncation error and to test the proposed algorithm's accuracy. The dataset, previously used by Marszal-Pomianowska et al. (2019) and Leiria et al. (2022), consists of 28 single-family houses in Denmark that have been renovated to Nearly Zero Energy Building standard (between 2012 and 2020). The apartments are equipped with radiators and underfloor heat-

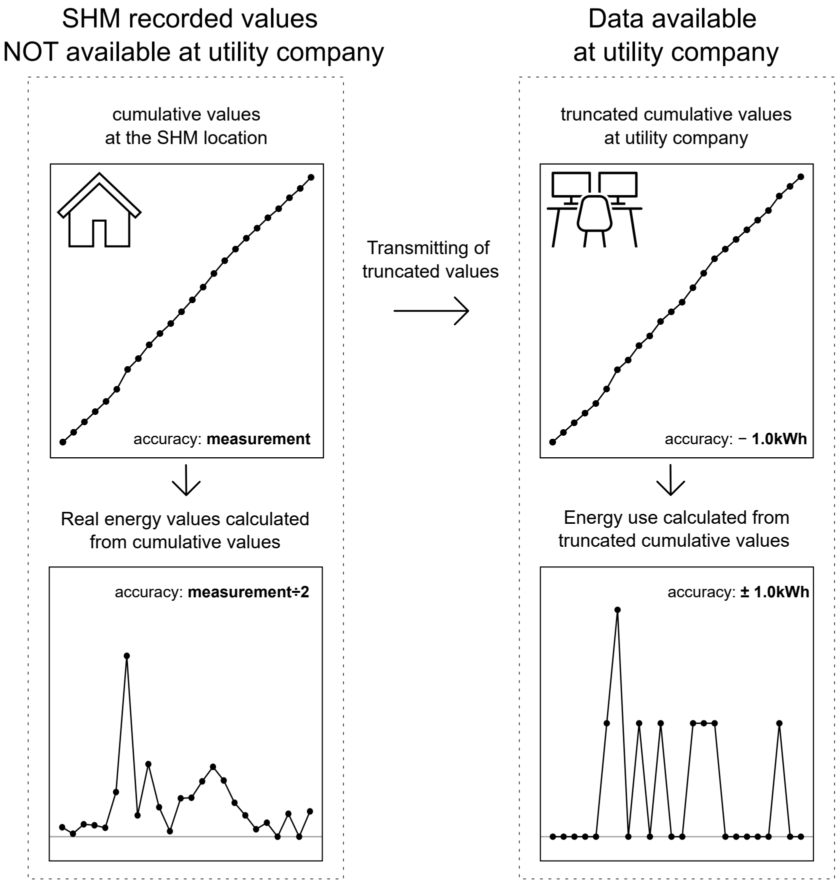

Figure 1: Measurement and data transmitting process of commercial SHMs, including its impact on the accuracy.

ing and have a floor area of between 97 m^2 to 112 m^2. The energy use for space heating and DHW were recorded with a temporal resolution of one hour. To mimic the process of commercial SHMs (Figure 1), the energy for space heating and DHW were first combined, then the hourly high resolution measurements were accumulated and truncated to integer values, before the hourly energy use was calculated based on the truncated cumulative data.

As ∼25 % of the data were missing, it was decided to impute them. Since the scope of this work is to assess and reduce the truncation error in SHM measurements, it is not essential that the imputed values are exact, but that their pattern is plausible, i.e. the values are within a realistic range and the pattern follows the existing data. Thus, gaps up to 48 values, representing 99 % of all gaps, were imputed using the mean of 24 h and 48 h leading and lagging values, i.e. the mean of the value at the same time step two days before, one day before, one day ahead and two days ahead (in the original full resolution data). The resulting data were visually assessed (Figure 2) and it was concluded that this imputation produced plausible results. As long missing gaps (>48 values) remained, the data were split at the missing values, with the remaining sequences having to be at least seven full days long (186 values). This left 166,392 data points in 83 sequences (length between 19 and 183 days; mean of 84 days).

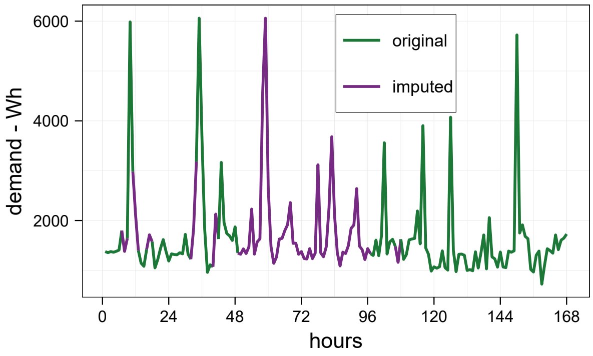

Figure 2: Example of one week of hourly energy use data with imputed values demonstrating the suitability of the chosen imputation approach. The large imputed gap in the middle is 48 *values long.*

Analysis of rounding error

The error introduced by truncating the cumulative values was expected to be uniformly distributed with limits 0 kWh and 1.0 kWh. Applying a one-sample Kolmogorov-Smirnov test for each month separately showed that this assumption can be considered plausible for the winter months (March, April and October - December) but not for the summer months (May - September) (Figure 3). Further analysis of the truncation error in the energy use data revealed an unexpectedly high number of 0 kWh errors during the same

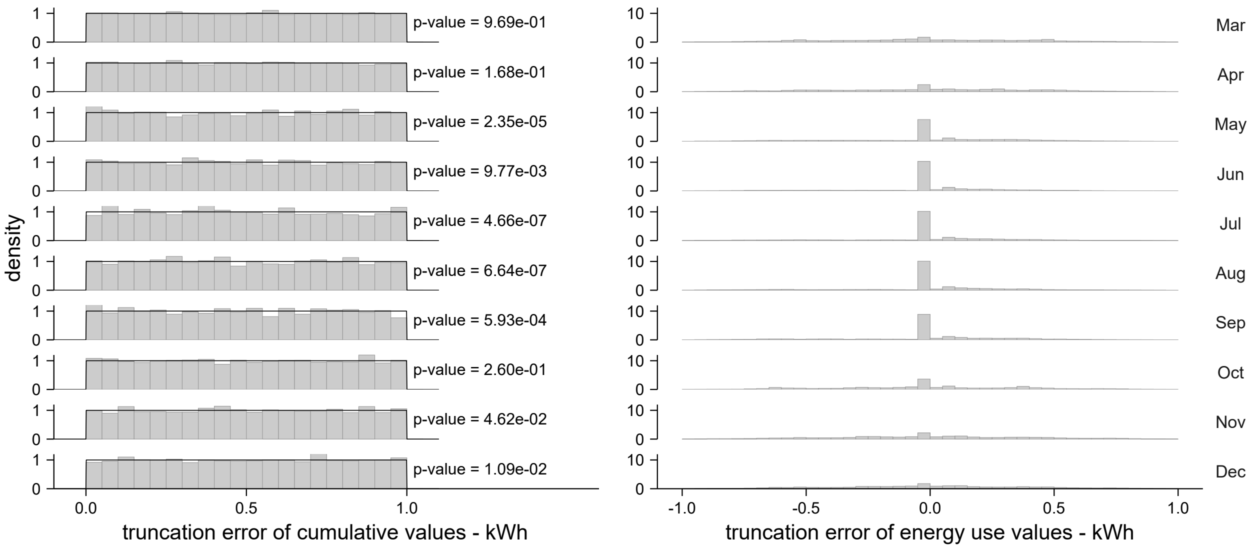

Figure 3: Analysis of the truncation error based on the high-resolution dataset used. The denoted p-values are from a one-sample Kolmogorov-Smirnov test against a uniform distribution with limits 0 kWh *and* 1.0 kWh*.*

summer months. This unexpectedly high number of 0 kWh errors for energy use was traced back to the hours with actual energy use of 0 kWh. Thereby, two (or more) times, the same truncated cumulative values are transmitted by commercial SHM. Consequently, these data points have the same truncation error in the cumulative data, which shifts the distribution of the truncation error away from a uniform distribution, explaining why the error is not uniformly distributed for some months. At the same time, the repeated transmitted values correctly lead to 0 kWh for the energy use, which explains the high number of hours without error. This was also confirmed by analysing the frequency of hours with 0 kWh per day in the dataset used, as shown in Figure 4, which also shows that the truncated data can still be used as a proxy to identify months/days with a high number of hours with 0 kWh energy use.

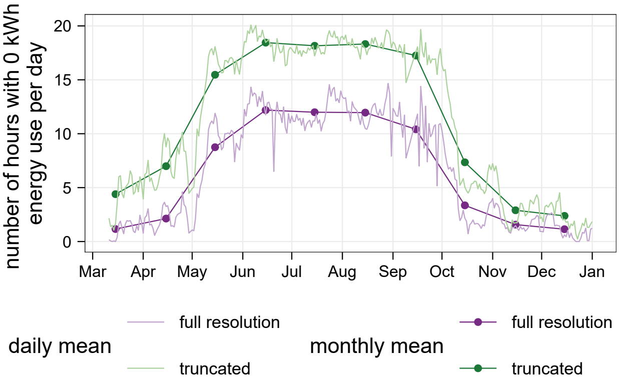

Figure 4: Frequency of hours with 0 kWh *energy use per day, as a daily and monthly average.*

An analysis of the error distribution when removing hours with an actual energy use of 0 kWh (Figure 5) showed that in this case the truncation error in the cumulative values for all months could be considered as coming from a uniform distribution (limits: 0 kWh and 1.0 kWh). The number of 0 kWh errors for the energy use was thereby drastically reduced to an expected level. The distribution then resembled approximately a normal distribution for the winter months, with the majority of errors between −0.5 kWh and 0.5 kWh. Consequently, it appears that the error introduced by truncating the energy use in winter is a combination of an approximately normal distribution when actual energy use is not zero and zero error when actual energy use is zero. For the summer months, no clear theoretical distribution could be identified, but the majority of errors are still between −0.5 kWh and 0.5 kWh. It is hypothesised that this is related to the different energy use, heating and DHW in winter and mainly DHW in summer, but this could not be definitively confirmed.

Recovering algorithm - SPMS

The issues caused by truncation described above, in particular the masking of the energy use pattern, inspired the authors to approach this as a smoothing problem. The three main conceptual points of the developed method, which can be summarised as *Smooth - Pointwise Move - Scale* (*SPMS*), are outlined below:

1) *Smooth:* Smooth the truncated energy use
2) *Pointwise Move:* Ensure that some chosen pointwise accuracy (maximum deviation from the truncated data) is obeyed and that obtained values are positive
3) *Scale:* Ensure that the resulting data over a specified period sums up to the same amount as the truncated data

The idea behind the first step, smoothing, is to even out the on/off patterns caused by the truncation (Figure 1 bottom right). In principle, any smoothing technique can be considered. However, given the expected large amount of data, it should be computationally efficient. The second step ensures that SPMS does not lead to values that are known to be

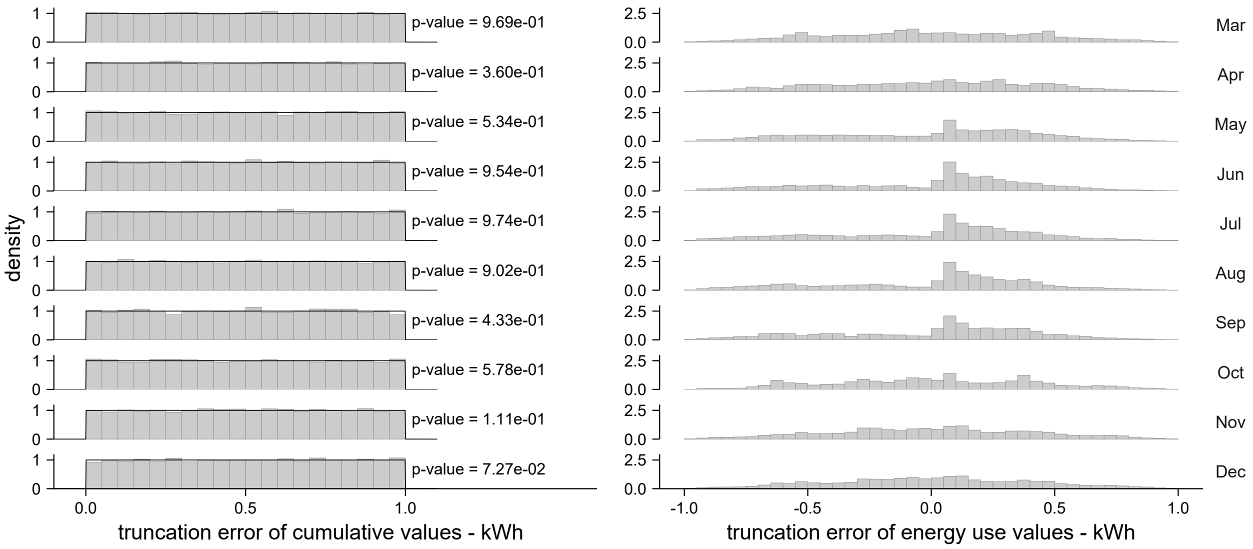

Figure 5: Analysis of the truncation error based on the high-resolution dataset used, when all hours with real 0 kWh *energy use are excluded. The denoted p-values are from a one-sample Kolmogorov-Smirnov test against a uniform distribution with limits* 0 kWh *and* 1.0 kWh.

impossible, i.e. deviating by more than ±1.0 kWh from the truncated values or being negative. However, based on the analysis of the distribution of the truncation error of energy use, the optimal allowed pointwise deviation is expected to be smaller than the theoretical maximum of ±1.0 kWh. In addition, it is not known whether the optimal allowed pointwise deviations might differ between periods with a high and low number of hours with 0 kWh energy use. This will be investigated as part of the case study analysis. The third step of SPMS ensures that the cumulative trend of the data is followed. This is done by scaling the obtained values uniformly so that, over a defined period of time, the recovered data sums up to the same amount as the truncated data. The length of the chosen period over which this is done (e.g. a day, a week) represents a trade-off between the acceptable deviation in cumulative values per period and the acceptable relative error per period introduced by truncating the cumulative values (i.e. a longer period with high cumulative energy has a smaller relative truncation error). Steps two and three are performed in a loop until the conditions of both steps are fulfilled.

Case study

With the proposed algorithm defined, the next step was evaluating the algorithms' performance. Therefore, the high-resolution dataset, is used. In the following, the different tested SPMS settings and the evaluation criteria are outlined.

SPMS settings

As mentioned above, any smoothing technique can be used for SPMS. For this work, the focus fell on a moving average (MA) and regression using Fourier basis functions (RFB). RFB was applied to the data structured in days, i.e. each day was treated as a separate sequence for smoothing, while MA was applied to the whole sequence. For both smoothing techniques, various settings were varied to control the smoothing. For RFB, the number of basis functions was changed from 3 to 23 (accounting sine and cosine pairs plus the constant) to evaluate different degrees of smoothing. For the MA the following parameters were analysed:

- Alignment of value within the window (window alignment): left, centre, and right
- Weighting: simple (equal), linear, exponential
- Window size: 2 to 6 values

In addition to these smoothing related parameters, the maximum pointwise deviation was also tested as it was expected, based on the rounding error analysis, that the optimal deviation would be ≤ 1.0 kWh and ≥ -1.0 kWh. The maximum allowed pointwise deviation was varied symmetrically from ±0.1 kWh to ±1.0 kWh in steps of 0.1 kWh.

For step three of SPMS, one day was chosen as the period over which the newly recovered data must sum up to the same amount as the truncated data. Firstly, because the cumulative energy consumption per day (mean =20.66 kWh) is high enough that the truncation error is relatively small; secondly, because the data usually has a daily trend; and thirdly, because a day is a 'natural' unit of time, which simplifies processing.

Evaluation criteria

To evaluate the accuracy of SPMS and find its optimal setting, the newly calculated energy use was compared against the real measured ones with full resolution. Each day was separately evaluated. The Normalised Root-Mean-Square Error (NRMSE) was used to assess the pointwise accuracy (equation 1).

$$NRMSE = \frac{\sqrt{\frac{\sum_{t=1}^{n}(p_t - o_t)^2}{n}}}{\bar{o}} \tag{1}$$

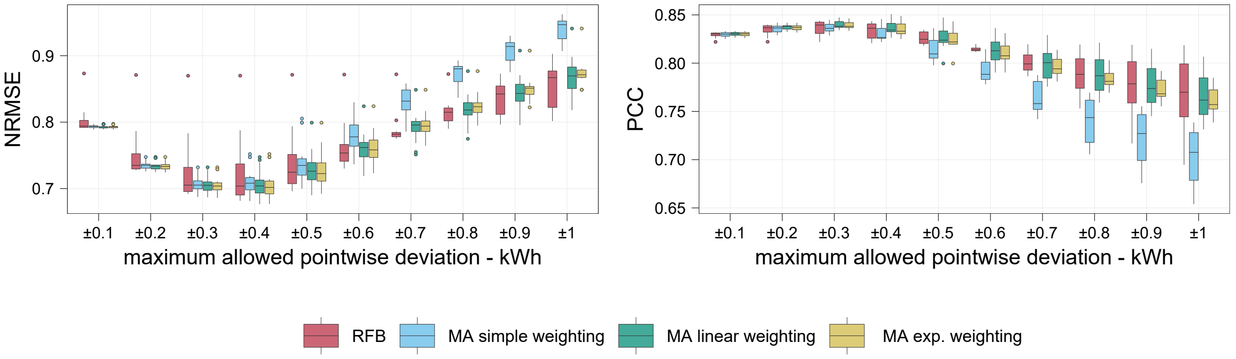

Figure 6: Distribution of mean results for each smoothing method as a function of the maximum allowed pointwise deviation from the truncated data. Based on the whole available data period (MA = moving average, RFB = regression using Fourier basis functions).

Where n is the number of data points per day (24), p_t is the new value at time instance t, o_t is the real observed value at time instance t, and $\bar{o}$ is the mean of the actual observed values. It is to be noted that for days where $\bar{o} = 0$, i.e. no energy use is recorded for the whole day, the NRMSE is not defined and such days were not considered for evaluation. These days are expected to occur mainly during the summer period, when DHW drives the energy use, and consequently, no energy is used if occupants are, for example, not at home. Additionally, the Pearson correlation coefficient (PCC) was used to evaluate the agreement between the trend of the calculated data and the original data. The PCC was calculated separately for each day, excluding days where it was not defined, e.g. where the standard deviation was zero.

Results

First, the maximum pointwise deviation over the whole data period was analysed. Figure 6 shows the distribution of the mean results (mean over all days) for each smoothing method with its different settings. It can be seen that the maximum pointwise deviation has a significantly more decisive influence than the smoothing methods. A maximum pointwise deviation of ±0.3 kWh or ±0.4 kWh leads to the most favourable results for both the NRMSE and the PCC. The same evaluation was done for the results split based on the monthly average frequency of hours with 0 kWh energy use per day (Figure 4), whereby only the truncated data was considered (as the high-resolution data would typically be not available) and 10 hours were used as the threshold. Summer months and winter months were consequently analysed separately. The results showed overall the same trend for both periods as the results for the whole period (Figure 6). Again, a maximum deviation of ±0.3 kWh or ±0.4 kWh gave the best results. This demonstrates that one maximum pointwise deviation can be used for both summer and winter periods. Additionally, it was found that the magnitude of the NRMSE decreases significantly in winter (0.38 to 0.55) while it increases in summer (1 to 1.4). This was expected as the average energy use was lower in summer than in winter. The PCC remained for both periods similar to that obtained for the whole period (Figure 6).

Overall, it was found that MA with a linear weighting, a centre-aligned window with a length of 5, and a maximum pointwise deviation of ±0.4 kWh leads to one of the best results for both periods and the best result for the whole data period (Figure 7). This maximum pointwise deviation of ±0.4 kWh is also supported by the analysis of the error introduced by truncation (Figure 5) which showed that the majority of errors lays between ±0.5 kWh. With this setting, SPMS decreased the NRMSE by 0.20 and increased the PCC by 0.03 compared to the truncated data.

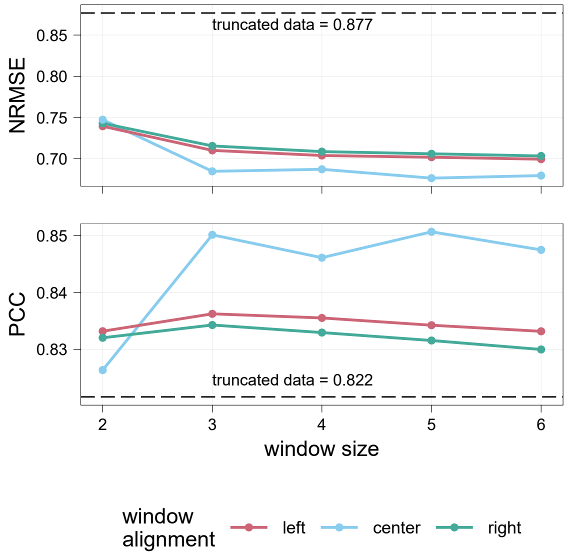

Figure 7: Mean over all days, for MA with a linear weighting and a maximum allowed pointwise deviation of ±0.4 kWh. *Based on the whole available data period.*

Two exemplary days were chosen to compare the energy use obtained from SPMS with optimal settings with the energy use of the original high-resolution data and the truncated data (Figure 8). It is evident that SPMS successfully reduces the on/off pattern of the truncated data and yields energy use curves close to the original high-resolution data.

However, the influence of the truncated data on the results from SPMS is also evident.

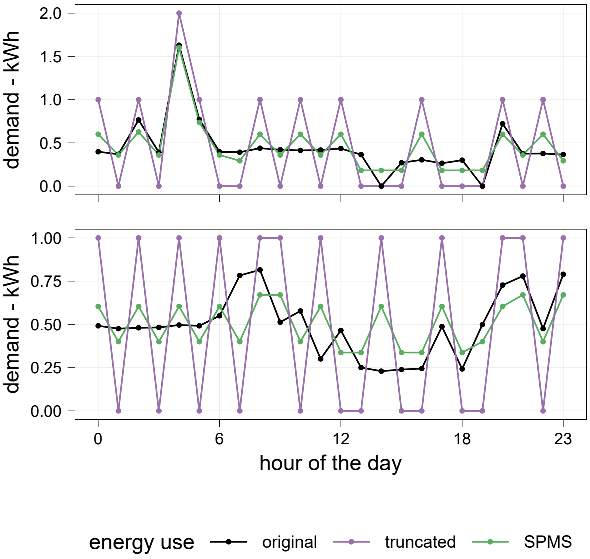

Figure 8: Daily energy use curves for two exemplary days, based on the original high resolution data, truncated data and results obtained from SPMS

Based on the previously observed influence of the truncated data on the results of SPMS, the change in NRMSE and PCC for each day compared to the truncated data (in relation to the high resolution data) was analysed for SPMS with optimal settings. As shown in Figure 9, there is a clear correlation between the results of the truncated data and the results obtained from SPMS, confirming the relationship seen previously. Although SPMS did not yield favourable results for all days, overall the positive effect exceeds the negative in quantity and magnitude. This confirms SPMS' overall usability but also highlights that SPMS' accuracy correlates to the truncated data's accuracy.

Conclusion & Discussion

Recent research has shown the principal potential of smart heat meter (SHM) data in the district heating (DH) sector. Nevertheless, it has also become evident that the low transmitted measurement resolution from commercial SHMs is a hurdle for many applications. Therefore, for the first time, the error introduced by commercial SHMs when rounding down the transmitted data has been analysed. Additionally, a method, SPMS, has been proposed to improve the accuracy and hence the usability of such truncated data.

The analysis of the error introduced by truncation showed that for cumulative energy data, the error in principle follows a uniform distribution with limits 0 kWh and 1.0 kWh. However, frequent periods with an actual demand of 0 kWh can shift the error away from a uniform distribution. For the energy use (non-cumulative values) in the winter months, the error is a combination of an approximately normal distribution when the actual energy use is not zero and zero errors when the actual energy use is zero. No theoretical distribution could be found for the summer months. This knowledge can support the estimation of confidence in the results obtained from such SHM data.

The proposed method to improve the accuracy of such truncated data, SPMS, shows promising results. With optimal settings, a moving average with linear weighting, a centre-aligned window of length 5, and a maximum pointwise deviation of ± 0.4 kWh, SPMS can reduce the NRMSE (-0.20) and increase the PCC (+0.03) of the truncated data relative to the high-resolution data. This improves the usability of such data for subsequent analysis. However, the results also clearly show that the performance is correlated with the truncated data, i.e. if the truncated data has a low PCC and/or a high NRMSE, the results of SPMS are limited by this. Therefore, while SPMS can mitigate the problems caused by low transmitted resolution, it cannot completely overcome them. This highlights the need to change the transmitting resolution of commercial SHMs. Therefore, utility companies need to start recognising the potential value of SHM data beyond billing and act to adapt their systems and transmission infrastructure to increase the usability of the data.

Data and code availability

The data used in this study cannot be shared due to GDPR restrictions. All code used in this work, is available at: `https://github.com/markus-schaffer/shm-truncated-analyses`

Acknowledgments

This work was funded by the Independent Research Fund Denmark under FOREFRONT project (0217-00340B). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 893945 (E-DYCE). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 958345 (PRELUDE). The authors would like to acknowledge the Himmerland housing association and Varmekontrol A/S for their support in data collection in Denmark. The authors declare no conflict of interest.

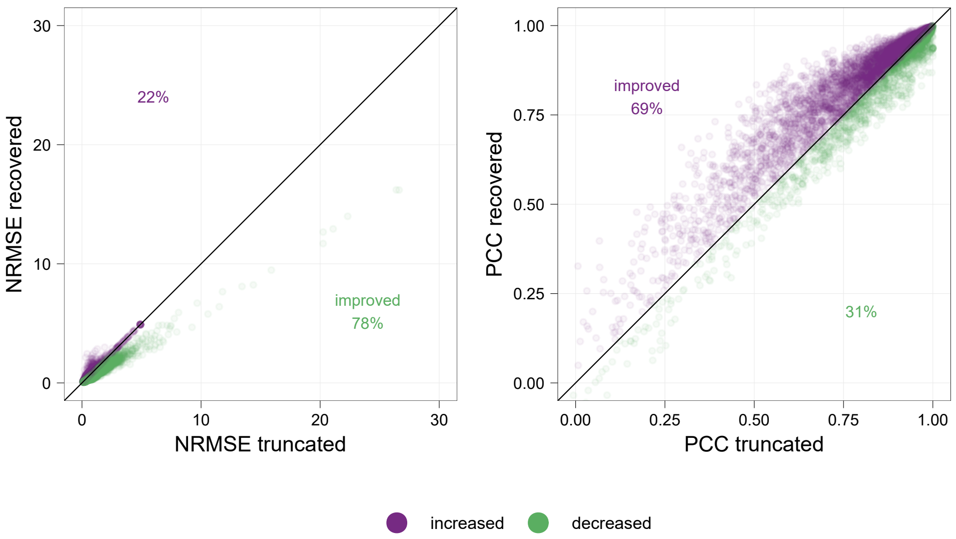

Figure 9: Comparison of the evaluation criteria of the data obtained from SPMS with ideal settings and the truncated data evaluated against the high-resolution data.

References

Calikus, E., Nowaczyk, S., Sant'Anna, A., Gadd, H., and Werner, S. (2019). A data-driven approach for discovering heat load patterns in district heating. Applied Energy, 252(January):113409.

do Carmo, C. M. R. and Christensen, T. H. (2016). Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy and Buildings, 125:171–180.

European Parliament (2018). Directive (EU) 2018/2002 amending Directive 2012/27/EU on energy efficiency. Official Journal of the European Union, (L 328/210).

Gianniou, P., Liu, X., Heller, A., Nielsen, P. S., and Rode, C. (2018a). Clustering-based analysis for residential district heating data. Energy Conversion and Management, 165(December 2017):840–850.

Gianniou, P., Reinhart, C., Hsu, D., Heller, A., and Rode, C. (2018b). Estimation of temperature setpoints and heat transfer coefficients among residential buildings in Denmark based on smart meter data. Building and Environment, 139:125–133.

Hauge Broholt, T., R. L. Christensen, L., and Petersen, S. (2022). Effect of measurement resolution on data-based models of thermodynamic behaviour of buildings. In REHVA 14th World Congress CLIMA.

Hedegaard, R. E., Kristensen, M. H., and Petersen, S. (2020). Experimental validation of a model-based method for separating the space heating and domestic hot water components from smart-meter consumption data. In Kurnitski, J. and Kalamees, T., editors, E3S Web of Conferences, volume 172, page 12001. EDP Sciences.

Kristensen, M. H., Hedegaard, R. E., and Petersen, S. (2018). Hierarchical calibration of archetypes for urban building energy modeling. Energy and Buildings, 175:219–234.

Kristensen, M. H., Hedegaard, R. E., and Petersen, S. (2020). Long-term forecasting of hourly district heating loads in urban areas using hierarchical archetype modeling. Energy, 201:117687.

Leiria, D., Johra, H., Belias, E., Quaggiotto, D., Zarrella, A., Marszal-Pomianowska, A., and Pomianowski, M. Z. (2022). Validation of a new method to estimate energy use for space heating and hot water production from low-resolution heat meter data. In BuildSim Nordic Conference 2022.

Leiria, D., Johra, H., Marszal-Pomianowska, A., Pomianowski, M. Z., and Kvols Heiselberg, P. (2021). Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case. Smart Energy, 3:100035.

Marszal-Pomianowska, A., Zhang, C., Pomianowski, M., Heiselberg, P., Gram-Hanssen, K., and Rhiger Hansen, A. (2019). Simple methodology to estimate the mean hourly and the daily profiles of domestic hot water demand from hourly total heating readings. Energy and Buildings, 184:53–64.

Schaffer, M., Tvedebrink, T., and Marszal-Pomianowska, A. (2022). Three years of hourly data from 3021 smart heat meters installed in Danish residential buildings. Scientific Data 2022 9:1, 9(1):1–13.