On the Tradeoff between Energy Data Aggregation and Clustering Quality

Holger Trittenbach, Jakob Bach, Klemens Böhm
Karlsruhe Institute of Technology
{holger.trittenbach,jakob.bach,klemens.boehm}@kit.edu

ABSTRACT

Energy data from industrial facilities is collected with high frequency. The resulting data volumes pose a scalability challenge for subsequent analyses. While data aggregation can be used to address it, the quality of analyses on aggregated data often is unknown. In our work, we propose an experimental design to evaluate the effects of aggregation on clustering energy data.

CCS CONCEPTS

• **Information systems → Data analytics**; **Clustering**; • **Computing methodologies → Cluster analysis**;

KEYWORDS

Time-Series Clustering, Smart Meter Data, Data Reduction

ACM Reference Format:
Holger Trittenbach, Jakob Bach, Klemens Böhm. 2018. On the Tradeoff between Energy Data Aggregation and Clustering Quality. In *e-Energy '18: The Ninth International Conference on Future Energy Systems, June 12–15, 2018, Karlsruhe, Germany*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3208903.3212038

1 INTRODUCTION

In the past, research on the management and analysis of energy data has focused on consumption data from households. Typical sampling intervals for such data are between 15 minutes and one hour. However, a significant share of today's energy consumption takes place in industrial settings. There, smart meters collect various quantities like voltages, currents, and harmonic distortions. They may do so from hundreds of machines, in second intervals. Hence, the extent of resources necessary for subsequent data mining, including processing power and storage, is much higher.

Data reduction, e.g., downsampling or aggregation, is a common way to deal with this challenge. Data reduction typically induces a loss of information. This in turn may bring down the result quality of data mining. So there is a tradeoff between data volume and information content. Understanding it is important when designing data-processing pipelines for energy data. However, existing work on, say, the clustering of energy-consumption curves uses one fixed sampling or aggregation level. Only a few studies have explicitly looked at different levels of temporal aggregation [5, 14]. In our work, we propose different ways to assess the effects of aggregation on clustering quality. We present an experimental design to this end and results on a real-world dataset from a production site.

2 EXPERIMENTAL DESIGN

2.1 Design Space

Several aspects have to be considered when evaluating the effects of aggregation on the clustering of time series:
(D1) Data set: Different physical quantities exhibit different behavior. For example, power and amperage depend on machine activity and show an on-off behavior. Frequency and voltage depend on the electrical grid and behave independently of the machine usage.
(D2) Clustering algorithm: Clustering algorithms fall into several categories, e.g., representative-based, hierarchical, and density-based [1]. The best choice usually is application-specific [3, 6, 13].
(D3) Dissimilarity measure: Clustering relies on dissimilarity measures. There is a great variety, and the choice is unclear [8].
(D4) Aggregation function: A simple form of aggregation is computing summary statistics for windows of equal length, i.e., Piecewise Aggregate Approximation (PAA) [7, 9]. For example, one can compute the mean over windows of 5 min, 10 min, 30 min etc.
(D5) Aggregation level: When computing aggregates over windows, the length of the window determines the amount of aggregation and therefore the length of the resulting time series.

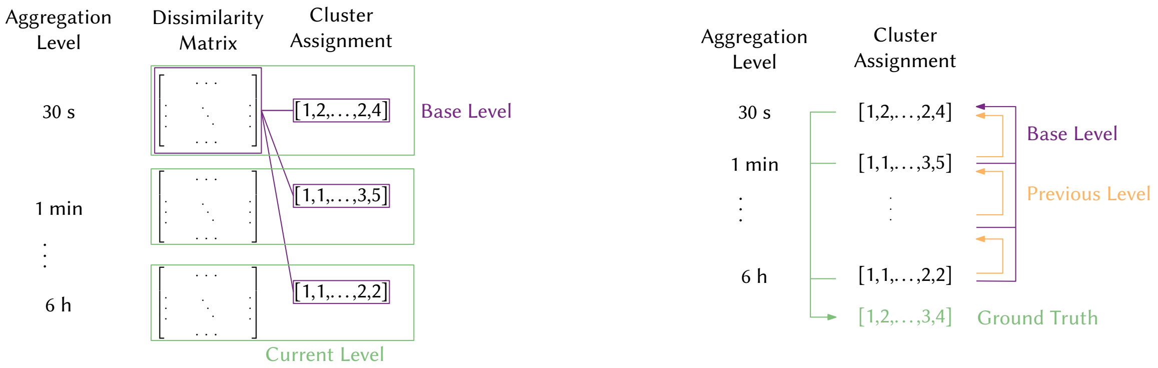
2.2 Clustering Evaluation

Clustering is an unsupervised technique, and there typically is no ground truth. A suitable evaluation depends on the goal, be it to discover groups of similar consumers, to identify recurring voltage patterns, or to find nodes in a grid with similar behavior. To cover a broad range of possible applications, we use different metrics to evaluate the experimental results. We focus on the tradeoff between data volume, i.e., the level of aggregation, and the information content, i.e., the quality of clustering.

(E1) Clustering Structure: One way of assessing the clustering structure is by looking at the distribution of cluster sizes. To evaluate the effects of aggregation, we are interested in the variation of the clustering structure with changing aggregation level. We use entropy as a measure of imbalance of the relative cluster sizes [12].

(E2) Internal Validity: Internal validity indices like the Silhouette Coefficient [11] evaluate the quality of clusterings. Most indices quantify cohesion within and separation between clusters [2]. As shown in Figure 1a, we propose two variants to evaluate internal validity when aggregating. The dissimilarity matrix contains the pairwise dissimilarities between time series. A cluster assignment is an integer vector, listing the number of the cluster of each time

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
e-Energy '18, June 12–15, 2018, Karlsruhe, Germany
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5767-8/18/06…$15.00
https://doi.org/10.1145/3208903.3212038



(a) Internal validity.

(b) External validity.

Figure 1: Schematics for comparison of validity indices across different aggregation levels (here: from 30 s to 6 h).

series. *Base level* internal validity indicates how well aggregation preserves the structure of the unaggregated data. So we use cluster assignments obtained with the aggregated data, but for index computations the dissimilarities of the unaggregated data. *Current level* internal validity uses both the cluster assignments and the dissimilarities of the aggregated data. This quantifies how well clustering represents aggregated data on a given level, independently from the unaggregated data.

(E3) External Validity: External validity indices like the Rand Index [10] quantify the similarity of a clustering result and a target cluster assignment. As displayed in Figure 1b, we propose three variants of external validity to compare aggregation levels. *Ground truth* validity uses domain-specific information. For example, one could expect data from the same sensor type, or day of the week to be in the same cluster. *Base level* external validity takes the clustering of unaggregated data as the target assignment. This allows analyzing the overall loss of information. *Previous level* external validity compares clusterings for adjacent aggregation levels.

(E4) Forecasting: Clustering can be a pre-processing step for other data mining methods, allowing to quantify clustering quality indirectly. A method often used with energy data is forecasting. To evaluate the effects of aggregation, we compare the forecasting error when combining forecasting with clustering on different aggregation levels.

3 PRELIMINARY RESULTS

We illustrate the effects of aggregation with data from a production site for power electronics. Ten machines are equipped with smart meters which measure more than a hundred attributes.

In our example, we choose active power as physical quantity (D1). We compare summary statistics like mean and standard deviation (D4) over aggregation windows from 1 min to 6 h (D5). The base dataset has a sampling rate of 30 s, each time series representing a full day. We select 7 clustering algorithms (D2) and 13 dissimilarity measures (D3) commonly used in the literature. Figure 2 shows base level internal and external validity, taking the median over the different clustering algorithms and dissimilarity measures.

We measure internal validity with the Silhouette coefficient [11], higher values indicating higher clustering quality. It decreases with growing aggregation for location statistics like the mean. Shape and dispersion measures like the standard deviation always yield a value close to zero, equivalent to the quality of a random clustering. The Adjusted Rand Index [4, 10] is our measure of external validity. A value of one indicates that cluster assignments on the unaggregated and aggregated data are equal. Again, we can see a decreasing trend with aggregation for location aggregates like the mean. Shape and dispersion measures result in cluster assignments different from the one for the base dataset at all levels.

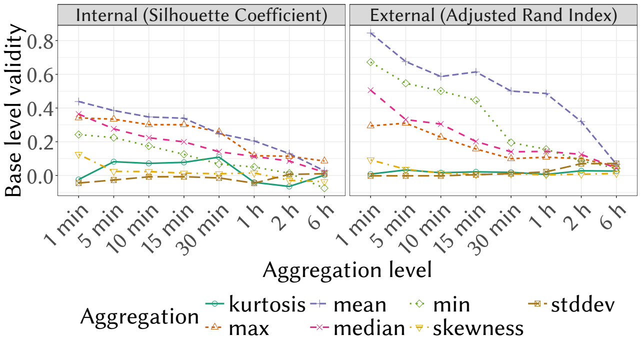


Figure 2: Base level validity by aggregation function.

4 CONCLUSION AND OUTLOOK

We have presented an experimental design to evaluate the effects of aggregation on clustering energy data. In future research, we will conduct an in-depth analysis of various experimental settings and evaluation metrics. In particular, we will study several physical quantities like amperage, voltage etc. and propose decision guidelines for domain experts.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG) as part of the Research Training Group GRK 2153: *Energy Status Data – Informatics Methods for its Collection, Analysis and Exploitation.*

REFERENCES

[1] Charu C Aggarwal. 2015. *Data Mining: The Textbook.* Springer.

[2] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M Pérez, and Inigo Perona. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition* 46, 1 (2013), 243–256.

[3] Gianfranco Chicco. 2012. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* 42, 1 (2012), 68–80.

[4] Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *J Classification* 2, 1 (1985), 193–218.

[5] Rishee K Jain, Kevin M Smith, Patricia J Culligan, and John E Taylor. 2014. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy* 123 (2014), 168–178.

[6] Ling Jin, Doris Lee, Alex Sim, Sam Borgeson, Kesheng Wu, C Anna Spurlock, and Annika Todd. 2017. *Comparison of Clustering Techniques for Residential Energy Behavior using Smart Meter Data.* Technical Report. LBNL.

[7] Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. 2001. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowl Inf Syst* 3, 3 (2001), 263–286.

[8] Eamonn Keogh and Shruti Kasetty. 2003. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Min Knowl Disc* 7, 4 (2003), 349–371.

[9] Eamonn J Keogh and Michael J Pazzani. 2000. Scaling up Dynamic Time Warping for Datamining Applications. In *KDD*. 285–289.

[10] William M Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. *J Am Stat Assoc* 66, 336 (1971), 846–850.

[11] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20 (1987), 53–65.

[12] Silke Wagner and Dorothea Wagner. 2007. *Comparing Clusterings – An Overview.* Technical Report. Faculty of Informatics, Universität Karlsruhe (TH).

[13] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. 2013. Experimental comparison of representation methods and distance measures for time series data. *Data Min Knowl Disc* 26, 2 (2013), 275–309.

[14] Tri Kurniawan Wijaya, Tanuja Ganu, Dipanjan Chakraborty, Karl Aberer, and Deva P Seetharam. 2014. Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data. In *ICDM*. 226–234.