Understanding the Effects of Temporal Energy-Data Aggregation on Clustering Quality

Holger Trittenbach, Jakob Bach and Klemens Böhm

Abstract: Energy data often is available at high temporal resolution, which challenges the scalability of data-analysis methods. A common way to cope with this is to aggregate data to, say, 15-minute-interval summaries. But it often is not known how much information is lost with this, i.e., how good analysis results on aggregated data actually are. In this article, we study the effects of aggregating energy data on clustering. We propose an experimental design to compare a wide range of clustering methods found in literature. We then introduce different ways to compare clustering results obtained with different aggregation schemes. Our evaluation shows that aggregation affects the clustering quality significantly. Finally, we propose guidelines to select an aggregation scheme.

ACM CCS: Information systems → Data mining → Clustering; Hardware → Smart grid.

Keywords: Data mining, clustering, energy data, aggregation, benchmark.

1 Introduction

A common way to scale analysis methods for energy data to high volumes is data reduction, such as downsampling or temporal aggregation. Data reduction induces a loss of information, which in turn may deteriorate the result quality of the analysis. Therefore, understanding the tradeoff between the volume of the reduced data and the information content is fundamental to design data processing pipelines.

In this article, we focus on clustering, a data-mining approach that is common with time series of energy data. Clustering has been used to obtain consumption profiles [1], to improve consumption forecasts [2–4], or to group the energy consumption of households for programs and policies of utilities [5,6]. Clustering of energy data must cope with various electrical quantities like voltage, current, and frequency. In industrial settings, high data volumes may come from hundreds of machines, collected in second intervals. This may require significant computational resources. Further, distances between time series become very similar when the length of the time series, i.e., the number of observations, increases. This is due to the curse of dimensionality [7] and often affects clustering-result quality.

To give way to scalability, users often reduce the data volume before clustering, by aggregating over time windows, mostly by averaging over 15 min intervals. We refer to these windows as *aggregation levels*. However, such an aggregation may distort the clustering result and the ensuing findings. Literature on energy data largely disregards these effects. We think that this is because (a) users may not be aware of the effects of aggregation, and (b) there is no systematic approach to compare different aggregation schemes, i.e., aggregation functions and levels.

Deriving a simple set of rules to select an aggregation scheme is unrealistic, given the diversity of energy data, clustering algorithms and the manifold applications. In a previous short article, we have sketched an experimental design to evaluate the tradeoff between clustering quality and energy-data aggregation [8]. We now propose and evaluate different ways to assess the effect of aggregation on clustering quality systematically. The outcome is evaluation methods and guidelines to arrive at informed decisions regarding the aggregation scheme for a specific application. This entails several challenges.

Data Characteristics: Electrical quantities may have different characteristics, like the typical shape of a time

series. This makes the choice of a suitable aggregation function for a specific quantity difficult. To illustrate, the active power of a machine depends on its type and its usage pattern. The net frequency in turn is determined and regulated by the grid operator. Hence, while "average" might be a suitable function to summarize power consumption in a time window, this may not be true for the net frequency. Here, volatility may be better.

Evaluation Methods: For unsupervised techniques like clustering, there is no ground truth to evaluate against. Instead, there exist several, often complementary measures of result quality. It is not clear how to compare quality values when aggregation schemes are different, as we will show later.

Design Space: There is a daunting variety of clustering algorithms and dissimilarity measures for time series [9, 10]. The best method usually is application-specific [11–13]. In addition, one must select an aggregation function and the size of the time window for aggregation. This results in a huge design space, i.e., the cross-product of data characteristics, clustering methods, dissimilarity measures, and aggregation functions. A full factorial experiment is prohibitive, and one needs to identify a subset which is conclusive.

Contributions. In this article, we present an experimental design to study the effects of aggregation on clustering results. It has certain dimensions, which we explain in the body of the paper, as follows:

(D1) *Data*: Time series of several electrical quantities in fixed and variable length.
(D2) *Clustering algorithms*: Representative-based, hierarchical and density-based algorithms.
(D3) *Dissimilarity measures*: Lock-step, elastic and complexity-based dissimilarities.
(D4) *Aggregation functions*: Summary statistics of location, shape and statistical dispersion.
(D5) *Aggregation levels*: Windows from 30 s up to 6 h.

Based on this design, we conduct experiments on high-resolution smart meter data from industrial production [14].

Next, we propose different ways to compare clustering results across aggregation levels. For instance, we cluster data with increasing level of aggregation, and we evaluate the resulting clusters both on the given aggregation level and against the original data. Our methods to compare results across different levels have led to insights not achievable with conventional evaluation methods for single aggregation levels. For instance, we have discovered that the size of the effect of aggregation depends on the data resolution, e.g., aggregating from 30 s to 1 min has a stronger impact than from 5 min to 10 min.

We conclude by proposing guidelines to select an aggregation scheme. For instance, some dissimilarities and aggregation functions perform poorly in almost every setting, and we suggest to exclude them from further experiments. In our use case, this has reduced the number of experiments necessary from 43092 to 378. Further, we have identified spurious improvement of clustering-validity indices, i.e., data aggregation may boost index values without real improvement in clustering quality. A consequence is that our guidelines strongly advise to validate clustering results against randomly generated sequences. Researchers and practitioners can use our guidelines as a reference to select aggregation schemes for their specific application.

Outline. Section 2 is a review of related work and fundamentals. Section 3 introduces our experimental design, and Section 4 proposes different ways to compare clustering results across aggregation levels. In Section 5, we present our experimental results and propose guidelines for selecting aggregation schemes. Section 6 concludes.

2 Related Work

In this section, we discuss related work on energy data aggregation, as well as on comparative studies on clustering algorithms and dissimilarities.

The literature on clustering energy data mainly relies on load (kW) or consumption (kW h) measurements from households [1, 3, 5, 13, 15–25]. Some references consider commercial and industrial facilities [19, 26, 27], large buildings [28–31], medium voltage consumers [12, 32] and transformers [33]. The references differ by the sampling rate and the representation of the time-series measurements. The most common sampling rate is at least 15 min, and only a few approaches use more fine-granular data [1, 15, 33]. For the time-series representation, one can distinguish between raw data and feature-based approaches. Most approaches for raw data work with fewer than 100 measurements per time series, with a few exceptions [1, 23, 24]. Feature-based approaches transform the time series into a feature vector before clustering, e.g., mean, periodicity, or seasonal scores. The number of features varies, but can be as high as 96 [25].

Our work focuses on *temporal aggregation*. The idea is to split the time series into non-overlapping intervals. An aggregation function is then applied to each interval. The result still is a time series, but of reduced length. This is different from *instance-based aggregation*, i.e., aggregation across different consumers, often used for load forecasting [19, 23]. Instance-based aggregation is far more common in literature on clustering energy data. There only are few references for temporal aggregation [25, 29].

Literature on energy-data clustering mostly relies on a single clustering algorithm. In many cases, this is a variant of k-means. There only are few comparative studies for clustering that use energy-consumption data [12,28]. A recent one compares algorithms on residential household data [13]. It shows that the clustering results are not consistent across different validity indices, and that the choice of a suitable clustering algorithm depends on the application. However, all of these studies focus on a specific aspect of the experimental design, such as the clustering algorithm or the dissimilarity measure. They also do not cover data aggregation.

3 Experimental Design

We first discuss the data set and our pre-processing (D1) in Section 3.1 and Section 3.2. Then we present the dimensions (D2)-(D5) of our experimental design in Section 3.3 to Section 3.5. We present measures to assess the effects of aggregation in Section 4. In the following, *experiment setting* refers to a specific combination of (D1)-(D5).

3.1 Data Set (D1)

Our data is an extension of the HIPE data set [14], which contains smart meter data from a production site for power electronics. It includes 10 machines, e.g., a heat furnace, a soldering machine and a pick-and-place machine which have been monitored over eight months. Each machine is equipped with a smart meter that measures more than a hundred attributes with a sampling rate between 2 s and 28 s. We select a subset of six electrical quantities as representatives for different time-series characteristics. Active power, amperage and power factor depend on the machine usage and follow an on-off behavior. Frequency and voltage both depend on the electrical grid and fluctuate almost independently of the machine usage since the energy consumption of a machine is comparatively low. Positive energy is the total energy consumption. It is quantized to 1 kW h steps, which results in a staircase pattern.

3.2 Pre-Processing (D1)

Within a sensor, the time between subsequent measurements varies. So we use the mean over 30 seconds as the *base-level* data. From each time series, i.e., the measurements of one electrical quantity over the entire observation period, we extract non-overlapping intervals. We refer to these intervals as *sequences*. We use two methods to extract the sequences: fixed and variable length.

3.2.1 Fixed Length

We extract sequences of one day with fixed start and end time at midnight. This is the method frequently used in related work. We only consider sequences with less than 1 % missing values and exclude days where all measurements of a machine are constant.

3.2.2 Variable Length

The reason to extract variable-length sequences is that periods where the machine is idle may not be of interest. In general, several ways to extract variable-length sequences are conceivable. In this study, we proceed as follows. We first search for values which make up more than 5 % of the data and consider them as candidates for periods of inactivity. For example, this could be 0 when the machine is switched off or a small value for a stand-by state. We limit the minimum length of a sequence to 30 min and the maximum length to one day. Some electrical quantities like frequency do not have an on-/off-behavior. In this case, we use the active power measurements to select the start and end times.

For both extraction strategies, we fill missing values with linear interpolation. We remove sequences which contain outliers. These are measurement errors, i.e., negative spikes for single measurements, machine shutdowns, i.e., measurements drop to zero, and days where machine activity is less than 30 min. We select machines with a clear on-/off-behavior for active power, amperage, and power factor. We also apply min-max normalization to active power, amperage and positive energy, because the range of these attributes is machine-dependent. Table 1 summarizes the data set.

3.3 Clustering Algorithms (D2)

Clustering algorithms fall into several categories, e.g., representative-based, hierarchical, and density-based [34]. We select standard approaches from different categories which support arbitrary dissimilarity measures. The specific parameter settings are not essential for the further understanding and are listed in Appendix Table 4.

3.3.1 Representative-Based Methods

Methods from this category assign objects to clusters based on their dissimilarity to representative objects. A popular member is *Partitioning Around Medoids (PAM)* [35]. PAM takes the number of representative objects k as an input. To determine k, we run PAM with $k \in [2, 10]$ and choose the result with the maximum Silhouette Coefficient. In addition, we use *Affinity Propagation (AP)* [36], parameterized as proposed in [36] and [37].

3.3.2 Density-Based Methods

Density-based approaches focus on the local neighborhood of individual objects. Clusters are regions of high density, i.e., with many objects within a certain dissimilarity threshold. We select *DBSCAN* [38], a prominent algorithm from this category. It takes two parameters: the minimum number of objects in a cluster, which we

Table 1: Sequences of fixed (F) and variable (V) length.

Quantity	Sequences		Machines		Norm.
	F	V	F	V	
Active power	489	516	4	4	min-max
Amperage	479	540	6	6	min-max
Frequency	725	856	3	8	—
Positive Energy	479	819	7	7	min-max
Power Factor	702	789	8	8	—
Voltage	706	612	3	6	—

set to 1, so that all objects are part of a cluster, and the neighborhood radius ϵ, which we set to the mean one-nearest-neighbor dissimilarity.

3.3.3 Hierarchical Methods

Hierarchical methods are popular for shape-based time-series clustering [39]. They create a hierarchy of clusters either bottom-up (agglomerative) or top-down (divisive). The merging or splitting of clusters during hierarchy construction depends on a dissimilarity measure, the linkage criterion. This criterion can influence the clustering result significantly [12,13,32,34]. So we use different linkage criteria: single linkage, complete linkage, average linkage and Ward's criterion [34, 40]. We determine the number of clusters $k \in [2, 10]$ with the Silhouette Coefficient.

3.4 Dissimilarity Measures (D3)

There is a great variety of dissimilarity measures, and a good choice depends on the application [41]. So we select representative measures of different types, i.e., lock-step, elastic and complexity-based dissimilarities. When necessary, we make slight adaptations so that the measures are symmetric and have value zero if objects are indistinguishable. The adaptations and the parametrization are not essential for the further understanding and are listed in Appendix Table 5.

3.4.1 Lock-Step

These measures compare sequences element-wise. An element-wise comparison entails a linear time complexity, a sensitivity to noise and inflexibility to shifts on the time axis. Common representatives are the L_p norms, in particular L_1 (*Manhattan*), L_2 (*Euclidean*), and L_{max} (*Chebyshev*).

3.4.2 Elastic

Elastic measures allow for offsets between values. Given this, these measures also are applicable to sequences of different length. A popular representative is *Dynamic Time Warping (DTW)* [42]. DTW searches for a minimum distance assignment between elements of both series, the warping path, by allowing for stretching and compression of the time axis. To reduce computation time, one can control the warping procedure by global constraints [11, 43]. We apply the well-known Sakoe-Chiba band [44].

Other dynamic approaches transfer the concept of the string edit (Levenshtein) distance to numeric sequences. Instead of a binary match criterion like the one for string distances, the *Edit Distance With Real Penalty (ERP)* [45] calculates the absolute difference between time-series values. Operators allowed in the matching are edit, removal and addition of values.

We also select the *Shaped-Based Distance (SBD)* [46], a semi-elastic measure. It uses the maximum normalized cross-correlation coefficient between sequences across all possible time-axis offsets.

3.4.3 Complexity-Based

These methods compare complexities of time-series patterns. A measure based on information-theoretic principles is the *Compression-based Dissimilarity Measure (CDM)* [47, 48]. It compares the size of two sequences compressed individually to the size of compressing the concatenation of both sequences. Similarly to [47], we first convert the sequences to a discretized SAX representation [49] before compressing them. In addition, we use the *Permutation Distribution Dissimilarity (PDD)* [50]. It compares the frequency distribution of subsequence patterns, extracted with a sliding window from each sequence.

3.4.4 Correction Factors

In addition to the standard dissimilarities, we also use modified versions of L_2 and DTW. The first modification is *CORT* [51], which adjusts the dissimilarity according to the correlation between the sequences. It decreases the dissimilarity in case of positive correlation and increases it for negative correlation. The second modification is *Complexity-Invariant Distance (CID)* [52]. It compensates that dissimilarities between simple patterns tend to be lower than ones between complex patterns.

3.5 Aggregation (D4, D5)

We aggregate the time series along the time axis over non-overlapping windows of equal length. This is similar

to Piecewise Aggregate Approximation (PAA) [53, 54], but with varying *aggregation functions*. Our aggregation functions cover summary statistics for data location (*mean*, *median*, *minimum*, *maximum*). We further use measures of dispersion (*standard deviation*) and shape (*skewness*, *kurtosis*), although we expect these measures may perform worse than the location statistics. By increasing the window length, we reduce the number of aggregated measurements that are in a sequence. For instance, a full day sequence consists of 2880 measurements. An aggregation to 10 min intervals reduces this sequence to 144 values. We aggregate to intervals commonly used in the literature: 1 min, 5 min, 10 min, 15 min, 30 min, 1 h, 2 h and 6 h. We refer to longer aggregation intervals as higher aggregation levels.

4 Evaluating the Effects of Aggregation

A suitable evaluation of clustering depends on the application, be it to discover groups of similar consumers, to identify recurring voltage patterns, or to find nodes in an electrical grid with similar behavior.

A common way to evaluate clustering results is validity indices. *Internal* validity indices evaluate properties of the clusters, such as the shape, compactness or distinctness. They can be useful to determine parameters of clustering algorithms, e.g., the number of clusters. *External* validity indices compare clustering results to a pre-defined ground truth. They tend to be used with synthetic data where a ground truth is available.

Other methods have been proposed to evaluate further properties of the clustering result. Examples are the robustness of clustering on different samples [18], variability of consumption over time [5], and properties to distinguish between clusters [25]. There also are informal evaluation approaches, like textual description or visual inspection of cluster representatives [5,16,18,25,27,33]. One can also evaluate clustering indirectly, by using the clustering result for subsequent data mining. An example is the forecast of energy consumption [3, 19, 21, 23, 24, 31]. The hypothesis is that clustering of consumers into homogeneous groups might yield better forecasts. Clustering quality is the improvement of this accuracy.

In this article, we use an extensive set of evaluation metrics to cover a broad range of possible applications. To assess the effects of data aggregation, we propose several ways to compare validity indices between aggregation levels.

4.1 Internal Validity

Different internal validity indices focus on different aspects of clustering quality. Most of them quantify cohesion and separation of clusters [55]. Cohesion measures the dissimilarity within clusters, which should be low. Separation measures the dissimilarity between clusters, which should be high.

Table 2: Internal validity measures.

Name	Ref.	Range
Inverted Normalized Connectivity	[60]	$[0, 1]$
Inverted Generalized Davies-Bouldin	[57]	$(0, \infty)$
Generalized Dunn Index	[59]	$[0, \infty)$
Silhouette Coefficient	[56]	$[-1, 1]$

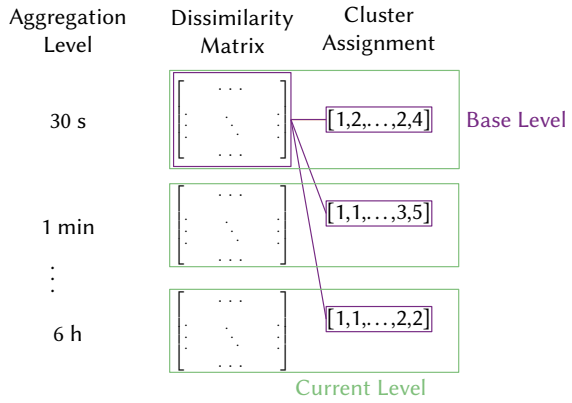

Figure 1: Schematic for comparing *internal* validity indices. The dissimilarity matrix to calculate the index either is from the current aggregation level or the base level [8].

We use three respective indices which have been used before when clustering energy data: the *Silhouette Coefficient* [56], the *Davies-Bouldin Index* [57] and the *Dunn Index* [58]. We use a generalized version of the Dunn Index [59] for stability against outliers. We generalize Davies-Bouldin so that one can use it for clustering algorithms without representative objects. In addition, we use *Connectivity* as a neighborhood-based index [60]. While many internal validity indices prefer spherical clusters [61], Connectivity works well for clusters of arbitrary shape. We invert some of these indices so that higher index values stand for better clusterings. See Table 2 as an overview, and Appendix 1.1 for details.

So internal validity evaluates a cluster assignment based on the dissimilarity matrix of the objects. In the conventional case, i.e., without aggregation, the same dissimilarity matrix is used to calculate the assignment and the validity index. With aggregation, however, one may assign clusters with the dissimilarity on one aggregation level, but calculate the validity index with the dissimilarities from a different one. By doing so, one can assess the clustering obtained from aggregated data by evaluating it on the original data. We therefore propose two alternatives to evaluate internal validity for different aggregation levels, see Figure 1.

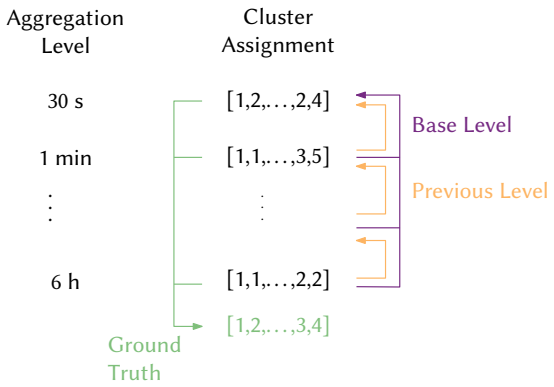

Figure 2: Schematic for comparing *external* validity indices. The cluster assignments are compared against the base level, the previous aggregation level or a ground truth [8].

4.1.1 Base Level

The idea of the base-level comparison is to evaluate how well aggregation preserves the structure of the unaggregated data. So the cluster assignments come from the clustering of aggregated data. The index calculation however uses the dissimilarities of the unaggregated data.

4.1.2 Current Level

With this variant, both the cluster assignment and the index calculation use the aggregated data. This quantifies the clustering quality on a given aggregation level.

4.2 External Validity

External validity indices quantify the similarity of an actual clustering result and a target cluster assignment. We select representative indices from common categories [62].

The first category is pair counting with the *Fowlkes-Mallows* [63], *Phi* [64] and *Rand Index* [65]. These indices rely on the confusion matrix between the actual clustering result and the target assignment. The Rand Index relates to accuracy, Fowlkes-Mallows to the geometric mean of precision and recall, and the Phi Index to the Pearson correlation of the actual and the target assignment. The second category of external validity indices is the set overlap, from which we select the *van Dongen measure* [66]. Indices from this category map each cluster from the actual assignment to a cluster in the target assignment and then calculate the maximum overlap. The third category is information-theoretic measures, from which we select the *Normalized Mutual Information* [67] between two clusterings. We normalize all indices according to [68], so that the index value is 1 if clusterings are indistinguishable. Table 3 is an overview of the external validity indices.

Like for internal indices, we propose variants to compare different aggregation levels. More specifically, we suggest to compare an actual clustering assignment to three different target assignments, see Figure 2.

Table 3: Normalized external validity measures.

Name	Ref	Category	Range
Fowlkes-Mallows	[63]	pair-counting	$(-1, 1]$
Phi Index	[64]	pair-counting	$(-1, 1]$
Rand Index	[65]	pair-counting	$(-1, 1]$
Normalized Mutual Information	[67]	information-theoretic	$[0, 1]$
Inverted van Dongen	[66]	set overlap	$[0, 1]$

4.2.1 Ground Truth

Cluster assignments are compared to domain-specific information. For example, we could expect data from the same machine, from the same sensor type, or collected at the same day of the week to be in the same cluster. If available, one may also use production-specific information such as the product type or the production step.

4.2.2 Base Level

In this variant, the base-level clustering is the target assignment. Clustering results from other aggregation levels are compared to it. This quantifies the information loss for a specific aggregation level. A similar method has been applied in [25], but with fewer experiment settings.

4.2.3 Previous Level

This variant is a relative comparison of clustering for adjacent aggregation levels. The idea is to quantify the change in the clustering when increasing the aggregation level by one step.

5 Results

Our experimental design includes 7 clustering algorithms, 13 dissimilarity measures, 7 aggregation techniques and 9 aggregation levels to cluster subsequences for 6 electrical quantities of variable and fixed length. In total, we evaluate 43092 settings. Our implementation is publicly available.[1]

5.1 Design Space

The large design space results in daunting runtimes for all experiment settings. The summary and interpretation of such a large result set would be challenging as well. So we first discuss effects of aggregation functions and dissimilarity measures, to see where differences are marginal. This might help to reduce the design space for further experiments.

[1] `https://www.ipd.kit.edu/clustagg`

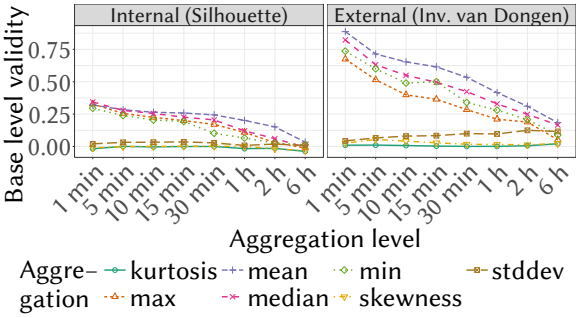

Figure 3: Base Level comparison of internal and external validity index. The figure shows the median validity over all experiment settings grouped by aggregation function.

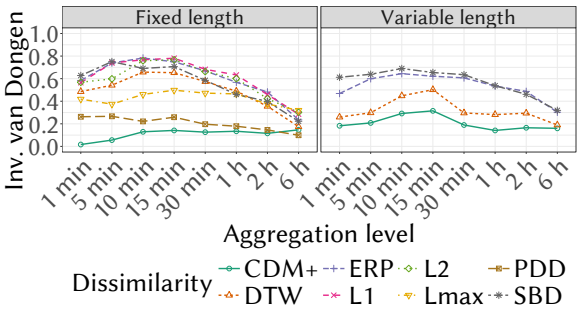

Figure 4: External validity (Inverted van Dongen) comparison against the Previous Level. The figure shows the median over all experiment settings grouped by dissimilarity.

5.1.1 Aggregation Function

Figure 3 graphs the median base-level Silhouette Coefficient and the Inverted van Dongen measure for different aggregation functions. For both external and internal validity, the results differ between location statistics and measures of shape and dispersion. For low aggregation levels in particular, location statistics lead to a significantly better validity. It is intuitive that a piecewise location summary is more characteristic for energy data than a piecewise summary of shape or dispersion. With increasing aggregation level, the validity decreases for location statistics. A reason might be that, for short intervals, location statistics are a good approximation of the raw data. With increasing window size, the error of this approximation increases and affects the clustering quality negatively.

5.1.2 Dissimilarity

In general, there is no dissimilarity measure that is best for all experiment settings. For example, Figure 4 displays the previous-level external validity for different measures. For fixed-length sequences, elastic dissimilarities and some of the lock-step dissimilarities yield similar results. However, the runtime complexity for elastic measures is quadratic or log-linear in the sequence length. For lock-step measures, it is linear. So lock-step measures yield roughly the same results and have a computational advantage.

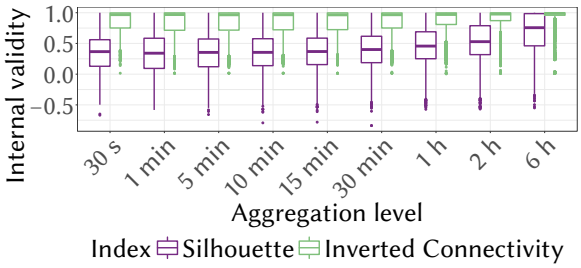

Figure 5: Current Level comparison of internal validity over all experiment settings.

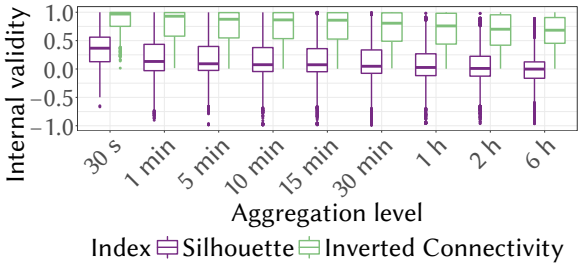

Figure 6: Base Level comparison of internal validity over all settings.

The corrections CORT and CID for the L_2 and DTW dissimilarities do not generally improve clustering validity. The complexity-based dissimilarities PDD and CDM+ have a low overall internal and external validity. We also observe that L_{max} results in lower external and also lower internal validity than L_1 and L_2. One explanation is that noise has a higher impact on L_{max} than on L_2 and L_1.

5.2 Evaluation Methods

We now discuss the usefulness of evaluation methods to assess the effects of aggregation.

5.2.1 Validity Index Selection

We first strive to identify the most relevant validity indices for a comprehensive view on the effects of aggregation. The rationale behind our selection is that indices with little or no correlation give complementary information, and highly correlated indices are redundant. For the internal indices, the correlation is moderate to strong. We select Silhouette Coefficient and Inverted Connectivity, which have the lowest correlation. For external validity indices, we find a strong correlation between all indices. In the following, we select the Inverted van Dongen measure. This is in line with recommendations in [68].

5.2.2 Method of Comparison

We now discuss insights from using the evaluation methods presented in Section 4. The following observations apply to all experiment settings. However, the extent of the effects may vary.

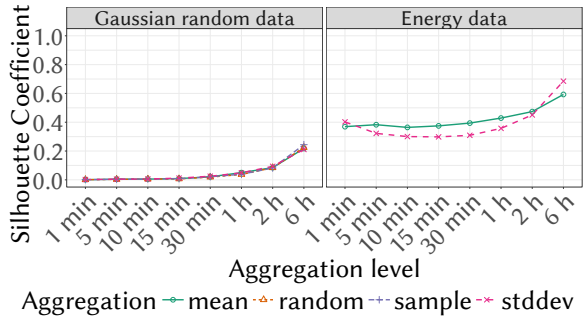

Figure 7: Comparison of energy data clustering and clustering of random sequences based on internal Current Level validity (Silhouette Coefficient). The figures shows the median over all experiment settings.

Internal Validity: Figure 5 graphs the current-level internal validity over all experiment settings with increasing aggregation level. The index values increase for levels larger than 15 min, with the best index value at 6 h aggregation. However, one can observe the same effect for randomly generated data as well. To illustrate this, we sample 600 random sequences independently from a Gaussian distribution, of the same length as the sequences on the base level. We then apply our setup to these random sequences. We also use two further strategies besides the seven standard aggregates. The first one is *sample* which randomly samples one value per aggregation interval from the base-level data. The second one is *random* which generates sequences randomly, i.e., independently of the base-level data. Figure 7 shows the current-level Silhouette Coefficient both based on random data and for our dataset. As one might expect, the absolute validity of clustering random sequences is lower than for the real-world data. However, we observe an increase in current-level validity in both cases. In particular, validity increases for the *random* strategy. This rules out smoothing effects introduced by aggregation as a possible explanation for the increasing validity.

If one considers a sequence as a vector, each observation corresponds to one dimension in the data space. Thus, the difference between aggregation levels is the number of dimensions. This indicates that the reason for increasing clustering quality is the curse of dimensionality.

The base-level comparison of internal validity in Figure 6 shows an opposite trend. It decreases with increasing aggregation level. As before, we also compare the results to randomly generated data. With random data, the median base-level Silhouette Coefficient is close to zero on all aggregation levels. We conclude that the decrease of base-level validity is not just a random effect.

In summary, a high internal validity might support the interpretability of the results, because clusters are more distinct. Increasing the aggregation level improves current-level internal validity. But this is also true for random data. So a base-level comparison is necessary to identify potential information loss. This finding shows that our evaluation methods are indeed useful to assess the effects of aggregation, as claimed in the introduction.

External Validity: Our results on external validity support this finding. With the base-level comparison, external validity decreases with increasing aggregation level. This means that similar sequences on the base level are assigned to different clusters after aggregation. The effect is strongest in the initial aggregation steps, e.g., from 30 s to 1 min. There only is little difference between 15 min and 30 min aggregation. This is one of the subtle insights announced in the introduction. The previous-level comparison supports the finding.

Nevertheless, relying only on external validity can also be misleading. For example, DBSCAN yields the highest external base validity across all aggregation levels. But it also has a low internal base-level validity. In such a case, i.e., the base-level clustering already is of poor quality, further comparison across aggregation levels is not meaningful.

5.3 Guidelines for Selecting Aggregation Schemes

Our experiment results suggest that there is no simple set of rules to select an aggregation function and level. The selection of a suitable aggregation scheme rather depends on the application as well as on practical constraints, like an upper limit on runtime. We therefore have extracted guidelines to help selecting an aggregation scheme. Researchers and practitioners can use them as a reference for their application.

Experimental Design: We propose to reduce the experimental design to ease analysis of aggregation effects with industrial smart-meter data. For fixed-length sequences, elastic dissimilarities and some of the lock-step dissimilarities yield similar clustering quality. In this case, one may reduce the experiments to lock-step measures, because runtimes for elastic measures are much higher. Further, the corrections CORT and CID for the L_2 and DTW dissimilarities do not generally improve clustering validity. The complexity-based dissimilarities PDD and CDM+ have a low overall internal and external validity. We also observe that L_{max} results in lower external and also lower internal validity than L_1 and L_2. Consequently, we have decided to remove all of these dissimilarities and corrections from the experimental setup.

For aggregation functions, location-based statistics generally outperform measures of shape and dispersion for base-level comparisons. Among them, *mean* has been most robust against aggregation, and we deem it the preferred choice.

On the other hand, the clustering algorithms and the linkage criteria do affect result quality significantly. More specifically, a good choice depends on the electrical quantity, and on whether sequences are of fixed or variable length. So we suggest to compare these dimensions carefully.

Evaluation Method: One has to consider both external and internal validity indices, since they evaluate two different aspects of clustering quality, see Section 4. However, our results suggest that it is sufficient to limit the analysis to one internal and one external validity index, see Section 5.2. Next, comparing aggregation functions by external base level is only meaningful if the base-level internal validity is high. In any case, one should compare against a ground truth, if available. For internal validity, one should rely on base-level comparisons.

Random Effects: We strongly advise to validate against randomly generated sequences to check whether quality improvements through aggregation are spurious.

Base-Level Selection: Selection of the base aggregation level should be with respect to a ground truth, if available. Namely, all conclusions rest on the assumption that clustering of the base level is the best possible clustering from an application perspective. The lowest aggregation level might not always be suited as the base level, for instance when the curse of dimensionality affects the data, or when the data contains a lot of noise.

6 Conclusions

The large volume of energy data challenges the scalability of data-analysis methods. In this article, the focus is on clustering. A common way to deal with that challenge is data aggregation. However, it often is not known how aggregation affects the quality of analysis results.

We have proposed an experimental design and different evaluation methods to compare clustering results across aggregation levels. Our experiments show that the aggregation function and level can have a significant effect on clustering results. Based on our experimental results, we have extracted guidelines to help researchers and practitioners when selecting an aggregation scheme. They can be useful to validate our findings with other specific applications, such as data from residential areas, or with activities of different length, such as machine-startups.

The guidelines are already helpful with our use case (cf. [14]). For instance, we now strive for ground-truth data to improve the base-level selection. Next, for a possible extension of our smart meter installations, we now deem it sufficient to estimate the effect of aggregation on a subset of the initial experimental space. In this case, we may focus on fixed-length sequences, mean aggregation and the $L1$ metric. This is a reduction from 43092 to 378 instances, which will reduce the experimental burden significantly.

Acknowledgement

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.*

Literature

[1] Omar Al-Jarrah et al. Multi-layered clustering for power consumption profiling in smart grids. *IEEE Access*, 2017.

[2] Sambaran Bandyopadhyay et al. Individual and aggregate electrical load forecasting: One for all and all for one. In *e-Energy*, 2015.

[3] Mohamed Chaouch. Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves. *IEEE Smart Grid*, 2014.

[4] Wen Shen et al. An ensemble model for day-ahead electricity demand time series forecasting. In *e-Energy*, 2013.

[5] Jungsuk Kwac, June Flora, and Ram Rajagopal. Household energy consumption segmentation using hourly data. *IEEE Smart Grid*, 2014.

[6] Ranjan Pal et al. Challenge: On online time series clustering for demand response: Optic – a theory to break the 'curse of dimensionality'. In *e-Energy*, 2015.

[7] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *IWANN*, 2005.

[8] Holger Trittenbach, Jakob Bach, and Klemens Böhm. On the tradeoff between energy data aggregation and clustering quality. In *e-Energy*, 2018.

[9] T Warren Liao. Clustering of time series data – a survey. *Pattern Recognition*, 2005.

[10] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering– a decade review. *Inform Syst*, 2015.

[11] Xiaoyue Wang et al. Experimental comparison of representation methods and distance measures for time series data. *Data Min Knowl Disc*, 2013.

[12] Gianfranco Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 2012.

[13] Ling Jin et al. Comparison of clustering techniques for residential energy behavior using smart meter data. Technical report, LBNL, 2017.

[14] Simon Bischof et al. HIPE–An Energy-Status-Data set from industrial production. In *e-Energy*, 2018.

[15] Ian Dent et al. Finding the creatures of habit; clustering households based on their flexibility in using electricity, 2012.

[16] Vera Figueiredo et al. An electric energy consumer characterization framework based on data mining techniques. *IEEE Power Systems*, 2005.

[17] Alejandro Gómez-Boix, Leticia Arco, and Ann Nowé. Consumer segmentation through multi-instance clustering time-series energy data from smart meters. In *Soft Computing for Sustainability Science*. Springer, 2018.

[18] Stephen Haben, Colin Singleton, and Peter Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Smart Grid*, 2016.

[19] Peter Laurinec and Mária Lucká. Comparison of representations of time series for clustering smart meter data. In *WCECS*, 2016.

[20] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, 2015.

[21] Franklin L Quilumba et al. Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities. *IEEE Smart Grid*, 2015.

[22] Teemu Räsänen and Mikko Kolehmainen. Feature-based clustering for electricity use time series data. In *ICANNGA*, 2009.

[23] Abbas Shahzadeh, Abbas Khosravi, and Saeid Nahavan-

di. Improving load forecast accuracy by clustering consumers using smart meter data. In *IJCNN*, 2015.

[24] Yogesh Simmhan and Muhammad Usman Noor. Scalable prediction of energy consumption using incremental time series clustering. In *Big Data*, 2013.

[25] Tri Kurniawan Wijaya et al. Consumer segmentation and knowledge extraction from smart meter and survey data. In *ICDM*, 2014.

[26] Alexander Lavin and Diego Klabjan. Clustering time-series energy data from smart meters. *Energy Efficiency*, 2015.

[27] Luis Hernández et al. Classification and clustering of electricity demand patterns in industrial parks. *Energies*, 2012.

[28] Félix Iglesias and Wolfgang Kastner. Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, 2013.

[29] Rishee K Jain et al. 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*, 2014.

[30] A Vaghefi, Farbod Farzan, and Mohsen A Jafari. Modeling industrial loads in non-residential buildings. *Applied Energy*, 2015.

[31] Junjing Yang et al. k-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energ Buildings*, 2017.

[32] George J Tsekouras, Nikos D Hatziargyriou, and Evangelos N Dialynas. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Power Systems*, 2007.

[33] Bogdan Neagu et al. Patterns discovery of load curves characteristics using clustering based data mining. In *Cpe-Powereng*, 2017.

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

[35] Leonard Kaufman and Peter J Rousseeuw. *Clustering by Means of Medoids*. Elsevier, 1987.

[36] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 2007.

[37] Ulrich Bodenhofer, Andreas Kothmeier, and Sepp Hochreiter. Apcluster: An r package for affinity propagation clustering. *Bioinformatics*, 2011.

[38] Martin Ester et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[39] Dimitrios Kotsakos et al. Time-series data clustering. In *Data Clustering: Algorithms and Applications*. CRC Press, 2014.

[40] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann series in data mangement systems. Elsevier, 2012.

[41] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Min Knowl Disc*, 2003.

[42] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *ICDM*, 1994.

[43] Joan Serra and Josep Ll Arcos. An empirical evaluation of similarity measures for time series classification. *Knowl-Based Syst*, 2014.

[44] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE T Acoust Speech*, 1978.

[45] Lei Chen and Raymond Ng. On the marriage of lp-norms and edit distance. In *VLDB*, 2004.

[46] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In *PODS*, 2015.

[47] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. Towards parameter-free data mining. In *KDD*, 2004.

[48] Ming Li et al. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 2001.

[49] Jessica Lin et al. A symbolic representation of time series, with implications for streaming algorithms. In *Workshop on DMKD*, 2003.

[50] Andreas M Brandmaier. *Permutation Distribution Clustering and Structural Equation Model Trees*. PhD thesis, Universität des Saarlandes, 2011.

[51] Ahlame Douzal Chouakria and Panduranga Naidu Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *ADAC*, 2007.

[52] Gustavo Batista et al. Cid: an efficient complexity-invariant distance for time series. *Data Min Knowl Disc*, 2014.

[53] Eamonn J Keogh and Michael J Pazzani. Scaling up dynamic time warping for datamining applications. In *KDD*, 2000.

[54] Eamonn Keogh et al. Dimensionality reduction for fast similarity search in large time series databases. *Knowl Inf Syst*, 2001.

[55] Olatz Arbelaitz et al. An extensive comparative study of cluster validity indices. *Pattern Recognit*, 2013.

[56] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*, 1987.

[57] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE Pattern Analysis and Machine Intelligence*, 1979.

[58] Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J Cybernetics*, 1973.

[59] James C Bezdek and Nikhil R Pal. Some new indexes of cluster validity. *IEEE T Sys Man Cy B*, 1998.

[60] Julia Handl and Joshua D Knowles. Exploiting the trade-off – the benefits of multiple objectives in data clustering. In *EMO*, 2005.

[61] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. Understanding and enhancement of internal clustering validation measures. *IEEE Cybernetics*, 2013.

[62] Silke Wagner and Dorothea Wagner. Comparing clusterings – an overview. Technical report, Faculty of Informatics, Universität Karlsruhe (TH), 2007.

[63] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *J Am Stat Assoc*, 1983.

[64] Karl Pearson. Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively measurable. *Philos T R Soc Lond*, 1900.

[65] William M Rand. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*, 1971.

[66] Stijn Van Dongen. Performance criteria for graph clustering and markov cluster experiments. Technical report, CWI, 2000.

[67] Ana LN Fred and Anil K Jain. Robust data clustering. In *CVPR*, 2003.

[68] Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *KDD*, 2009.

[69] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*, 1963.

1 Appendix

1.1 Adaptation of Indices

Notation Let $D = \{x_1, x_2, \ldots, x_m\}$ be a set of m time series. A cluster $C_i \subseteq D$ is a subset of all time series. A clustering C partitions the data set D into k clusters $C_1, \ldots, C_k$. The dissimilarity between two time series x and x' is $d(x, x')$.

Connectivity In the original version, high Connectivity indicates poor clustering quality [60]. We invert the index such that higher values indicate good clustering quality. As an intermediate step, we normalize Connectivity to $[0, 1]$ by dividing through the maximum Connectivity possible. Connectivity obtains its maximum if for all objects, the L nearest neighbors are assigned to a different cluster. The inverted and normalized Connectivity is:

$$i.Con(C) = 1 - \frac{Con(C)}{|D| \cdot \sum_{l=1}^{L} \frac{1}{l}}$$

Davies-Bouldin Index The original Davies-Bouldin Index [57] relies on dissimilarities between and to cluster centroids. To make the index applicable to non-representative-based algorithms, we use average-based instead of centroid-based dissimilarities. The average intra-cluster dissimilarity of a cluster C_i is:

$$\delta_{intra}^{avg}(C_i) = \frac{1}{|C_i| \cdot (|C_i| - 1)} \cdot \sum_{x, x' \in C_i,\ x \neq x'} d(x, x') \quad (1)$$

The average inter-cluster dissimilarity between two clusters C_i and C_j is:

$$\delta_{inter}^{avg}(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \cdot \sum_{x \in C_i, x' \in C_j} d(x, x') \quad (2)$$

We also invert the summands of the original Davies-Bouldin definition such that higher index values indicate good clustering quality. Our generalized and inverted version of the Davies-Bouldin Index is:

$$i.D\text{-}B(C) = \frac{1}{k} \cdot \sum_{C_i \in C} \min_{C_j \in C,\ i \neq j} \left\{ \frac{\delta_{inter}^{avg}(C_i, C_j)}{\delta_{intra}^{avg}(C_i) + \delta_{intra}^{avg}(C_j)} \right\}$$

Dunn Index The original Dunn Index [58] is defined as the ratio of the minimum dissimilarity between any two objects in different clusters, and the maximum dissimilarity between any two objects belonging to the same cluster. We use one of the generalized forms proposed in [59] to make the index more stable and less prone to outliers. With Equation 1 and Equation 2, the generalized version of the Dunn Index is:

$$Dunn(C) = \frac{\min\limits_{C_i \in C,\ C_j \in C,\ i \neq j} \delta_{inter}^{avg}(C_i, C_j)}{\max\limits_{C_i \in C} \delta_{intra}^{avg}(C_i)}$$

External Indices We apply the normalizations proposed in [68]. We also invert the normalized van Dongen measure by subtraction from 1 such that higher values indicate good clustering quality.

Holger Trittenbach is working towards the PhD degree at Karlsruhe Institute of Technology (KIT) at the Department of Informatics. His current research interest is data mining and machine learning in the field of outlier detection and active learning.

Address: Karlsruhe Institute of Technology (KIT), E-Mail: holger.trittenbach@kit.edu

Jakob Bach is working towards the PhD degree at Karlsruhe Institute of Technology (KIT) at the Department of Informatics. His current research interest is machine learning in the fields of feature selection and meta-learning.

Address: Karlsruhe Institute of Technology (KIT), E-Mail: jakob.bach@kit.edu

Prof. Dr. Klemens Böhm is full professor at Karlsruhe Institute of Technology (KIT), since 2004. Current research topics at his chair are knowledge discovery and data mining in big data, data privacy and workflow management.

Address: Karlsruhe Institute of Technology (KIT), E-Mail: klemens.boehm@kit.edu

Table 4: Overview of clustering algorithms.

Algorithm	Ref.	Category	Parameters
PAM	[35]	representative-based	$2 \leq k \leq 10$ with maximum Silhouette
AP	[36]	representative-based	$s(x,y) = -d(x,y)^*$, $s(x,x) = \underset{x,y}{median}(s(x,y))$, max iterations = 1000, $\lambda = 0.9$
DBSCAN	[38]	density-based	$minPts = 1$, $\epsilon = \underset{x}{mean}(d_{1NN}(x))^*$
Hier.avg	[40]	hierarchical	average linkage, $2 \leq k \leq 10$ with maximum Silhouette
Hier.comp	[40]	hierarchical	complete linkage, $2 \leq k \leq 10$ with maximum Silhouette
Hier.sin	[40]	hierarchical	single linkage, $2 \leq k \leq 10$ with maximum Silhouette
Hier.ward	[69]	hierarchical	Ward's criterion, $2 \leq k \leq 10$ with maximum Silhouette

* $s(\cdot,\cdot)$ = similarity, $d(\cdot,\cdot)$ = dissimilarity

Table 5: Overview of dissimilarity measures.

Diss.	Ref.	Category	Parameters	V*
CDM+[a]	[47]	complexity	SAX alphabet size = 8, compression = gzip	✓
DTW	[42]	elastic	—	✓[b]
DTW.CID	[42, 52]	elastic + complexity	—	✓[b]
DTW.CORT	[42, 51]	elastic + lock-step	tuning parameter $k = 2$	X
DTW.Band	[42]	elastic	Sakoe-Chiba window size = 10%	X[c]
ERP	[45]	elastic	gap value $g = 0$	✓[b]
L_1		lock-step	—	X
L_2		lock-step	—	X
L_2.CID	[52]	lock-step + complexity	—	X
L_2.CORT	[51]	lock-step	tuning parameter $k = 2$	X
L_{max}		lock-step	—	X
PDD	[50]	complexity	embedding dimension m by entropy heuristic	X[d]
SBD	[46]	elastic	—	✓

* Applicable to sequences of variable length (yes/no).
[a] We modify the formula of CDM slightly to obtain a dissimilarity in $[0, 1]$ instead of $(0.5, 1)$
[b] We additionally normalize the resulting dissimilarities to account for differences in length.
[c] Undefined if lengths of sequences differ too much and therefore not used.
[d] Undefined for sequences of length 1 and therefore not used.