Energy Informatics

RESEARCH Open Access

ESTSS—energy system time series suite: a declustered, application-independent, semi-artificial load profile benchmark set

Sebastian Günther[1], Jonathan Brandt[1], Astrid Bensmann[1*] and Richard Hanke-Rauschenbach[1]

*Correspondence:
astrid.bensmann@ifes.uni-hannover.de

[1] Institute of Electric Power Systems, Leibniz University Hannover, Appelstraße 9A, 30167 Hannover, Germany

Abstract

This paper introduces an univariate application-independent set of load profiles or time series derived from real-world energy system data. The generation involved a two-step process: manifolding the initial dataset through signal processors to increase diversity and heterogeneity, followed by a declustering process that removes data redundancy. The study employed common feature engineering and machine learning techniques: the time series are transformed into a normalized feature space, followed by a dimensionality reduction via hierarchical clustering, and optimization. The resulting dataset is uniformly distributed across multiple feature space dimensions while retaining typical time and frequency domain characteristics inherent in energy system time series. This data serves various purposes, including algorithm testing, uncovering functional relationships between time series features and system performance, and training machine learning models. Two case studies demonstrate the claims: one focused on the suitability of hybrid energy storage systems and the other on quantifying the onsite hydrogen supply cost in green hydrogen production sites. The declustering algorithm, although a bys study, shows promise for further scientific exploration. The data and source code are openly accessible, providing a robust platform for future comparative studies. This work also offers smaller subsets for computationally intensive research. Data and source code can be found at https://github.com/s-guenther/estss and https://zenodo.org/records/10213145.

Keywords: Time series, Time series analysis, Time series features, Feature engineering, Load profiles, Energy systems, Machine learning, Statistical analysis, Systems modeling

Introduction

This paper proposes a set of load profiles represented as time series data in the context of energy systems. The aim is to enable researchers to test and benchmark methodologies and algorithms. Additionally, the data set facilitates a deeper understanding of researchers' specific problems.

Given the imparative to transition from fossil-based to renewable energy systems, energy systems are currently a critical area of research, resulting in significant alterations in the energy system landscape (Proedrou 2021; Grandjean et al. 2012; Sandhaas

et al. 2022; Meinecke et al. 2020a; Intergovernmental Panel On Climate Change (Ipcc) 2023). Numerous methodologies and algorithms have been proposed to enhance renewable energy systems (Meinecke et al. 2020a, 2020b; Ammari et al. 2022; Olatomiwa et al. 2016; Yang et al. 2018). Examples include component sizing optimization in wind, photovoltaic, diesel, and battery systems; energy management design; self-consumption rate optimization; schedule optimization of flexible loads; or topology optimization in distributed microgrids (Ammari et al. 2022; Olatomiwa et al. 2016). These methodologies and algorithms target various objectives, such as increasing performance, lifetime, efficiency, or reliability and decreasing costs or system size (Olatomiwa et al. 2016; Yang et al. 2018; Anoune et al. 2018).

Methods and algorithms require data to determine and quantify their effectiveness (Proedrou 2021; Grandjean et al. 2012; Sandhaas et al. 2022; Meinecke et al. 2020b; Hulk et al. 2018). Comprehensive testing is essential and must involve both parameter sensitivity and input data sensitivity. Parameter sensitivity is often evaluated using methods like Monte Carlo simulations (Brandt et al. 2023), whereas input data sensitivity analysis for methods and algorithms necessitates a variety of input data types.

Studies on methods and algorithms in energy systems can be classified into three categories: (a) studies that use only one load profile, providing only qualitative or indicative results; (b) studies that utilize a few load profiles, which could be either real, artificial, or both, offering a better but still incomplete understanding; and (c) studies that conduct a methodicaly comprehensive input data analysis (Proedrou 2021; Meinecke et al. 2020b).

An issue is that even studies classified under category (c) can yield biased and non-representative results, thereby making a quantitative performance assessment elusive. Moreover, specialized data sets limit the reusability and comparability of results across different studies. Boundary conditions, such as when algorithms perform optimally or fail, are indeterminable. Data often lacks reproducibility due to confidentiality in industrial settings (Meinecke et al. 2020a; Sandhaas et al. 2022) or privacy concerns in residential setups (Proedrou 2021; Sandhaas et al. 2022; Meinecke et al. 2020a).

A viable solution is providing a representative open-source data set, an objective that has attracted increasing interest since 2005 (Proedrou 2021). Prior tools, methods, and data sets that provide or generate load profiles have been proposed. These mainly encompass residential, commercial, municipal load profiles (e.g. Pflugradt et al. (2022); Staudt et al. (2018); Tjaden et al. (2015); Wang and Hong (2020); Team (2022); Islam et al. (2020); Anvari et al. (2022); Marszal-Pomianowska et al. (2016); Fischer et al. (2015); Widén et al. (2009); Armstrong et al. (2009); McLoughlin et al. (2015); Granell et al. (2015); Grandjean et al. (2012); Wilson et al. (2021); Angizeh (2020); Meier et al. (1999); Park et al. (2019); Lindberg et al. (2019)), industrial load profiles (e.g. Braeuer (2020); Huber et al. (2019); Gotzens et al. (2020); Sandhaas et al. (2022); Binderbauer et al. (2022); Angizeh (2020); Meier et al. (1999)), electric vehicle load profiles (e.g. André (2004); Giorgi et al. (2021); Sorensen et al. (2022)), grid and microgrid load profiles (e.g. Behm et al. (2020); Meinecke et al. (2020a, 2020b)) and renewable energy load profiles (e.g. Pfenninger and Staffell (2016); Staffell and Pfenninger (2016)).

Methods to derive these data sets typically include augmentation (e.g. Tjaden et al. (2015); Meier et al. (1999)), behavior/agent-based techniques (e.g. Pflugradt et al. (2022); Hoogsteen et al. (2016); Widén et al. (2009); Armstrong et al. (2009)), probabilistic/

stochastic approaches (e.g. Anvari et al. (2022); Marszal-Pomianowska et al. (2016); Fischer et al. (2015); Sandhaas et al. (2022); Giorgi et al. (2021); Sorensen et al. (2022); André (2004)), clustering/segmentation (e.g. Islam et al. (2020); McLoughlin et al. (2015); Granell et al. (2015); Sandhaas et al. (2022); Binderbauer et al. (2022); Park et al. (2019); Kim et al. (2018)), regression (e.g. Gotzens et al. (2020); Lindberg et al. (2019)), Artificial Neural Networks (e.g. Wang and Hong (2020); Behm et al. (2020)), and physical simulations (e.g. Wilson et al. (2021); Pfenninger and Staffell (2016); Staffell and Pfenninger (2016)).

Notably, the scope of the papers listed is not exhaustive. For a more in-depth review, metastudies that investigate available data are available, and the reader is referred to them (Proedrou 2021; Grandjean et al. 2012; Sandhaas et al. 2022; Meinecke et al. 2020a).

In the context of this study, the *hctsa* package is particularly noteworthy (Fulcher et al. 2013; Fulcher and Jones 2017; Fulcher et al. 2023). It assembles an interdisciplinary representative set of data across various scientific fields, curiously neglecting energy data. This set is derived from a superset of data by equally weighting various scientific domains (Jones et al. 2021; O'Hara-Wild 2023). The present study shares some methodological similarities with *hctsa* and will delve into these later.

Each of the existing data sets offers unique value but is also highly specific, posing challenges for universal applicability across a wide array of individual studies. These challenges stem from various factors such as varying size, temporal and spatial resolution, presence of data gaps, level of data aggregation, intended use case, and generation methodology (Proedrou 2021; Meinecke et al. 2020a). Moreover, many of these studies aim to generate high-quality data for specific applications, leading to solutions that yield insights unique to those applications. By nature, this application specificity constrains the feature space and the diversity of the data, which may be adequate for algorithms tailored to specific applications. The present work acknowledges the value of these specialized data sets and does not intend to render them obsolete. Instead, it aims to address a different gap, recognizing that this makes the value of this work also highly specific.

The objective of this paper is to propose a carefully curated set of application-independent load profiles for use as input data to be able to quantify the impact of input data on various methodologies and algorithms. As a byproduct, the paper also introduces a declustering methodology to create this set. The dataset is derived from real-world energy system data and modeled to manifest high diversity and low-discrepancy. In other words, the integral features of the time series in this set span a broad range while retaining typical time and frequency domain characteristics inherent in energy system load profiles.

As a result, the present study delivers generalized univariate data, facilitating overarching, application-independent insights. Depending on the specific research question, the data set can be utilized as a standalone resource or as supplementary input data to draw additional conclusions. In particular, the data set aims to enable researchers to address questions such as identifying the requisite time series features for high performance of methodologies and establishing correlations between time series features and performance. Two case studies substantiate the effectiveness and value of this dataset: one focused on hybrid energy storage systems and another on hydrogen production

sites. The paper also provides a general guide on the dataset's usage. Associated with this study is a git-project that offers the source code and additional material. These resources can be accessed at https://github.com/s-guenther/estss. The datasets associated with the study are also available at https://zenodo.org/records/10213145.

The paper is structured as follows: "Methodology" section outlines the methodology to derive a low-discrepancy dataset, detailing the desired properties and the algorithm employed; "Resulting Sets of Time Series" section presents the resulting datasets; "Application and Usage" section demonstrates the claims through case studies and provides general usage information; and "Summary" section offers a summary and outlines future work.

Methodology

First, the desired properties of the data set are delineated, and an overview of the methodology is provided. This overview sets the stage for the forthcoming subsections, which delve into time series features, the initial dataset, and the manifold and decluster steps of the proposed methodology.

Data set specification and method overview

In this study, the terms *time series*, *load profile*, and *signal* are used interchangeably and will henceforth be referred to as *time series*. This term is predominantly used in machine learning methods in contrast to *load profile* from energy engineering or *signal* from information technology and electronics, and the methods employed in this research primarily draw upon machine learning paradigms.

A time series can be characterized by various scalar *features* such as the *mean* or *root mean square*, among others. The subsequent subsection comprehensively introduces these features. Multiple features form an *n*-dimensional *feature vector*, mapping each time series to a point in an *n*-dimensional *feature space*. Consequently, a *set* of multiple time series forms a point cloud within this feature space.

A key advantage of representing a set of time series in feature space is that it imposes order on an otherwise unordered set of time series in the time domain. This order is beneficial for uncovering functional relationships or correlations between the outcomes of various methodologies and a given time series feature. To elaborate: a methodology may calculate a scalar result for a given time series, such as the *minimum required storage dimension*. Such a result can be assigned to each time series in a set, and without features for ordering, only basic descriptive statistical measures like *mean*, *minimum*, *maximum*, and *quantiles* can be derived. However, the imposed order in the feature space enables plotting these results against a feature, thereby revealing potential correlations. Notably, these identified correlations are independent of the specific set and are generalizable, in contrast to statistical measures that are only valid for the particular set under consideration.

The objective is to generate a set of time series that is approximately uniformly distributed within a defined boundary in the feature space, essentially seeking a *low-discrepancy* set of points (Kuipers and Niederreiter 2006; Drmota and Tichy 1997). Figure 1 exemplifies this objective. Figure 1a presents an arbitrarily distributed set in a 2D-plane, where an accumulation or cluster is noticeable in the upper-left corner,

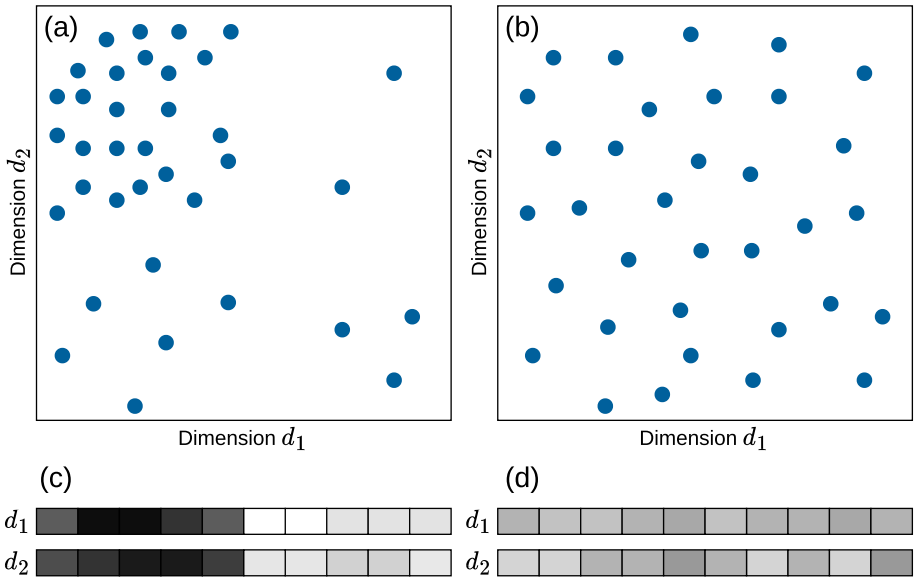


Fig. 1 Distribution of time series in feature space. **a** arbitrary distribution with evident clustering in 2D-plane; **c** same distribution visualized as *n*-dimensional histogram. **b** desired uniform distribution with low-discrepancy in 2D-plane and **d** its corresponding representation as *n*-dimensional histogram

and a sparse region is apparent in the right half. Figure 1c displays the same distribution but as an n-dimensional histogram. In this representation, each row corresponds to the distribution of a single dimension, depicted as a color-coded histogram. The cluster or accumulation is evident in both dimensions, characterized by densely populated bins on the left and sparsely populated bins on the right (darker shades indicate more densely populated bins). On the other hand, Fig. 1b and Fig. 1d illustrate a set that is uniformly distributed. Specifically, the points are evenly spread across the 2D-plane in Fig. 1b, while the histograms in Fig. 1d contain approximately equal numbers of points within each bin. These properties of uniform distribution extend to the n-dimensional set, as well.

The advantages of employing this uniform set over an arbitrary set are multifold. An arbitrary set tends to occupy a smaller volume within the feature space, limiting the scope of feature-bounds and potentially obscuring underlying relationships. Moreover, such sets are often highly clustered, increasing computational time and leaving sparse or empty regions in some feature-ranges, concealing functional relationships and widening confidence intervals.

The desired set of time series shall have additional characteristics.

- It should comprise time series with both varying and constant signs, each forming a subset of equal size.
- Both subsets, as well as their union, shall be uniformly distributed within the feature space.
- Furthermore, the integral of each time series should peak in the beginning at time $t = 0$ and reach its minimum at the last time step $t = t_{\text{end}}$, to simplify problem formulations for methodologies using this data, by establishing known boundary con-

ditions without sacrificing generalizability. For instance, this mimics an entire discharge cycle in energy storage applications.

- Each time series is discrete and shall consist of 1000 unitless points, serving as a trade-off between the informational content of a time series and the computational load for methodologies that utilize the data.
- Further, the time series are normalized to a maximum absolute value of one.
- Different sizes of time series sets shall be made available, each adhering to the above conditions. These sets will be defined to have sizes of 4096, 1024, 256, and 64 points, enabling users to balance computational complexity against resolution.
- The smaller sets shall be subsets of the larger ones.

Figure 2 schematically outlines the algorithm for obtaining the desired set of time series. The first step, detailed in "Initial Time Series" section, involves assembling an initial set of time series collected from various domains. This set is highly clustered within the feature space with several empty regions, as depicted in the lower portion in Fig. 2a. "Time Series Features" section introduces the features that form this feature space. In the second step, elaborated in "Manifold Step" section and represented in Fig. 2b, the set is manifolded through various transformations. These transformations involve the recombination of time series through methods such as superposition and concatenation. Additional transformations are achieved by applying various signal processors such as compressors and limiters. The resulting manifolded set retains high levels of clustering. However, a small share of time series now populates the entire range of the feature space.

Finally, the set is declustered and pared down to the desired number of time series. This process, described in "Decluster Algorithm" section and represented in Fig. 2c, involves selecting time series uniformly across the feature space, thus eliminating any existing clusters. The decluster step results in a set with the desired properties.

Transforming time series data into feature space is a prevalent step in data pipelines in machine learning (Fulcher et al. 2013; Hastie et al. 2009). While transformation into feature space is commonly used in literature to cluster data into distinct groups

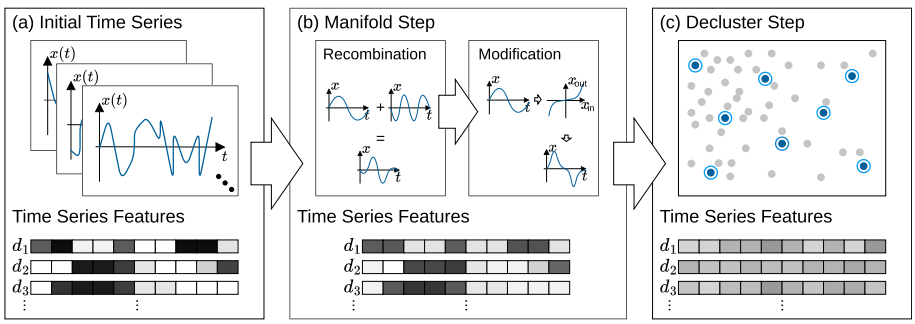


Fig. 2 Outline of the algorithm for generating a heterogeneous and low-discrepancy set of time series through three main steps: **a** Assembly of an initial set of time series from various domains (detailed in "Initial Time Series" section) that is highly clustered in feature space (introduced in "Time Series Features" section); **b** Manifold this set through recombination and modification (detailed in "Manifold Step" section). Clusters are retained; and **c** Decluster this set by finding a low-discrepancy subset from the previous superset (detailed in "Decluster Algorithm" section)

(Ismail Fawaz et al. 2019), the current study employs it to decluster data. This is achieved by populating sparse volumes and alleviating dense volumes within the feature space.

Some efforts towards declustering or anticlustering can be found in Mishra et al. (2017); Späth (1986); Valev (1998), but the state-of-the-art methodology stems from the *anticlust* package (Papenberg and Klau 2021; Brusco et al. 2020; Papenberg 2023). The method constructs subsets from the main set, where each subset is internally dissimilar but maximally similar to other subsets. It is based on a modified or inverted version of the *k-means* algorithm. In both *anticlust* and our approach, the goal is to maximize the dissimilarity within each subset. The key distinction lies in the composition of the final set. In *anticlust*, the union of subsets creates the main set, thereby preserving the overall density distribution albeit with fewer total elements.

In contrast, our method selects a single subset and discards the rest, which equalizes the density across the feature space. Consequently, *anticlust* can be considered an advanced form of stratified sampling, while our declustering approach transforms the shape of one distribution into another – in this work, a uniform distribution, although the methodology is extendable to other shapes. Furthermore, *anticlust* focuses on maximizing dissimilarity solely based on the four statistical moments, while the present study aims to maximize dissimilarity across a broader range of features.

Time series features

In the context of time series analysis, a feature maps a time series to a scalar value. Numerous features have been defined in the literature, and for this study, we compile a list of features derived from various sources. The *hctsa* package provides the most comprehensive collection, defining over 7700 features (Fulcher et al. 2013). However, many of these features stem from the same method but with different parameters, such as the *autocorrelation* with 100 different lags. The *catch22* package reduces this dimensionality to 22 features without a significant sacrifice in classification accuracy (Lubba et al. 2019).

Several other toolboxes for feature-calculation exist, including *kats* (Jiang et al. 2022), *tsfresh* (Christ et al. 2018), *tsfel* (Barandas et al. 2020), *theft* (Henderson and Fulcher 2022), *feasts* Ravi et al. 1994, and *tsfeatures* (Hyndman et al. 2023). In this study, we only consider those available in *Python*, namely *kats*, *tsfresh*, *tsfel*, and *catch22* and complement some features manually taken from Wang et al. (2015). After removing duplicates, redundancies, and features that require arbitrary parameters, the total number of unique features is reduced to 119 from an initial 255 base features without variants. This list includes simple features like *mean*, *standard deviation*, and *root mean square*, as well as more complex ones such as the *lag of the first minimum of the autocorrelation function* or the *goodness of exponential fit to embedding distance distribution*. A summary of the features used and a detailed discussion of the selected and dismissed features can be found in the additional material.

These features serve two distinct purposes within the methodology. A dimensionally-reduced set is used to build the feature space, in which the time series should be uniformly distributed. Meanwhile, the complete set is employed in the analysis to correlate results and identify functional relationships.

Initial time series

The starting point for the methodology is an *initial data set*. This data set comprises time series from various domains, including photovoltaic systems, wind farms, residential, commercial and municipal buildings, microgrids and grids, industrial loads, machine tools, production plants, real-world and artificial driving profiles, as well as train and tram load profiles for both onboard and substation systems. This compilation aligns with the literature presented in the introduction section. The lengths of these time series range from 420 to approximately 2×10^6, with temporal resolutions varying from milliseconds to hourly aggregation. The values of the measured quantities span from 1×10^{-2} to 400×10^6. However, both the temporal resolution and absolute value of the measured quantities are irrelevant due to subsequent normalization.

The utilized time series data set includes confidential information. To maintain confidentiality while ensuring accessibility, a randomized, sectionized, and normalized derivative of this data set is created. This derivative is constructed by extracting randomly located sections of varying lengths from the raw data, resampling them to 1000 points via a piecewise cubic Hermite interpolating polynomial (Fritsch and Butland 1984), and normalizing them to a mean of zero and a maximum absolute value of one. The range of section lengths is manually determined for each time series to capture characteristic shapes within the complete time series while avoiding over- or under-sampling. The resulting time series set is made publicly available in the additional material to ensure transparency and reproducibility of subsequent steps and results.

After the selection and normalization processes, the initial data set contains $2^{11} = 2048$ time series. A random selection of 32 time series from this set is illustrated in Fig. 3 to provide a visual overview of the data set's diversity. The depicted time series exhibit a wide range of time- and frequency-domain characteristics.

Manifold step

A series of modifications *manifold* the initial data set: recombination via concatenation and superposition, application of signal processors, and fixing boundary conditions. The objective is to increase the diversity and variance within the feature space. The adopted manifold approach has been empirically derived to meet this objective.

Time series recombination

The recombination step involves two substeps: concatenation and superposition. The first substep, concatenation, employs the initial set of time series. A random selection of random count time series is made, from which a new subsection is extracted with a varying length and location. These subsections are then resampled to a different length and concatenated into a single time series. The resulting time series is resampled and normalized again. The parameters used for this substep have been empirically determined and are documented in the additional material.

The second substep, superposition, involves the random selection of random count time series from both the selection and concatenation pools. A randomly chosen scalar then scales each time series. Superposition is achieved by summing up the scaled time series at each time step. The resulting time series is again resampled and normalized.

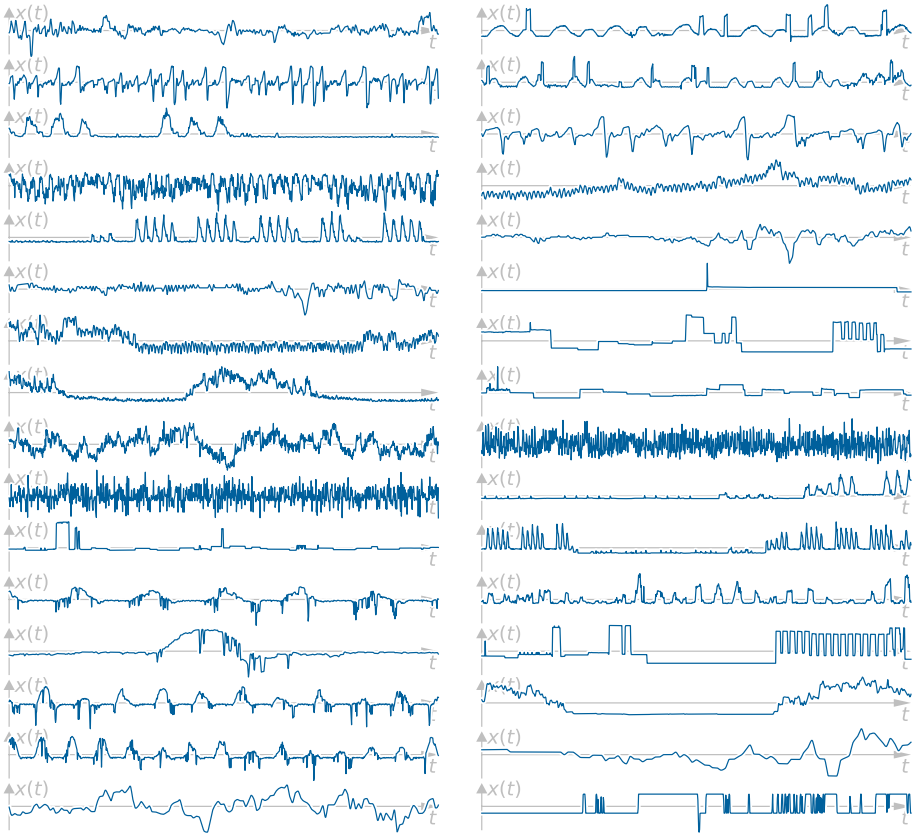


Fig. 3 A random selection of 32 time series of the initial set. A huge variety in time- and frequency-domain characteristics is apparent

Parameters for this substep have also been empirically determined and are documented in the additional material.

Both substeps, concatenation and superposition, are implemented to augment the diversity of the entire set. In energy systems, these abstract operations can be interpreted in functional terms. Concatenation can model a state change within a system, while superposition can represent various operational equipment contributing to a common electrical bus.

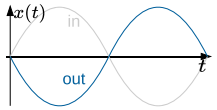
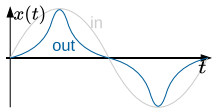
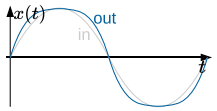
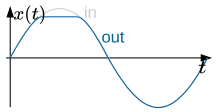
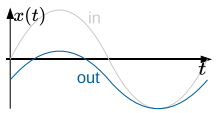
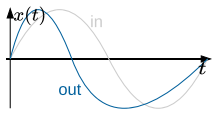
The data set manifolds to $2^{13} = 8192$ time series after concatenation and $2^{15} = 32768$ points after superposition.

Time series modifications

The recombined data set containing 32,768 time series is further manifolded by applying a chain of base signal processing operations to each time series. These operations include invert, expand, compress, curtail, shift, and time distortion. Table 1 presents an overview of these operations.

These base operations are concatenated in a chain with random sequence, random parameters, and random count to modify a given time series. Each base operation may be selected twice based on a predefined probability, and the list of realized operations in

Table 1 Time series modifications with corresponding equations and exemplary behaviour on a sine-wave

Invert	$y(t) = -x(t)$	
Expand	$y(t) = \begin{cases} x^a(t) & \text{for} \quad x(t) \geq 0 \\ -x^a(t) & \text{for} \quad x(t) < 0 \end{cases}$	
Compress	*see expand*	
Curtail	$y_1(t) = \begin{cases} x(t) & \text{for} \quad x(t) \leq a \\ a & \text{for} \quad x(t) > a \end{cases}$ $y_2(t) = \begin{cases} -a & \text{for} \quad x(t) \leq -a \\ x(t) & \text{for} \quad x(t) \geq -a \end{cases}$ $y_3(t) = \begin{cases} x(t) - c/2 & \text{for} \quad x(t) \geq a/2 \\ x(t) + c/2 & \text{for} \quad x(t) \leq a/2 \\ 0 & \text{otherwise} \end{cases}$	
Shift + Normalize	$y(t) = \frac{x(t)+a}{\max \lvert x(t)+a \rvert}$	
Time Distort	$y(t) = x(\tau) \quad \text{with} \quad \tau = f(t)$	

the chain is shuffled. Subsequently, the time series are normalized to a maximum absolute value of one after each operation. Exemplarily, the chains could appear as

- `compress: 0.70 |> expand: 2.17 |> invert |> curtail bottom: 0.14 |> shift: 0.65 |> curtail top: 0.24` or
- `curtail mid: 0.09 |> shift: 0.23 |> compress: 0.62 |> expand: 1.58`

where the number following the colon indicates the randomly selected parameter and the `|>` symbol denotes a pipe, which passes the output of the left-hand side function into the right-hand side one. The mean length of the chains based on selected probabilities is 4.9 with a standard deviation of 1.7. The parameters and probabilities for these operations are empirically selected to ensure a high variance in the time domain representation of the input time series, while preventing the time series from degenerating to a constant value for the majority of the time. The details are documented in the additional material.

The inclusion of these operations is motivated by several considerations. Curtailment is often encountered in energy systems due to equipment limitations or direct current (DC) bus restrictions. The shift operation reintroduces a DC component or offset removed during earlier normalization. Inversion is included, as the sign of the power flow often depends on its definition. Expand, compress, and time distortion, although not directly corresponding to operations in energy systems, are mainly included to augment the diversity of the time series pool. They may also result from energy management system rules and control inputs.

For each time series in the recombined set, 16 variations are created through randomly generated chains of operations, leading to a total of $2^{19} = 524288$ time series.

Fix boundary conditions

As stipulated in the data set specification in "Data Set Specification and Method Overview" section, each time series should have its maximum integral value at $t = 0$ and minimum at $t = t_{\text{end}}$. Consequently, the integral of any intermediate positive values must not exceed the integral of the prior time series. Furthermore, a time series must not start or end with positive values.

Figure 4 elucidates this issue and the principle for its resolution. Figure 4a presents the original, arbitrary time series in the time domain, depicted in grey, along with its corresponding integral shown in Fig. 4b. The maximum and its time stamp of this integral are identified. A negative offset is subsequently added to the original time series, spanning from the start to the identified time stamp, as indicated by the hatched area in Fig. 4a. The area of this offset equals the maximum of the integral. This offset is added as a sinusoidal quarter-wave rather than as a constant value, thus ensuring a smooth alteration of the time series shape instead of introducing a harsh step. If necessary, these steps are repeated.

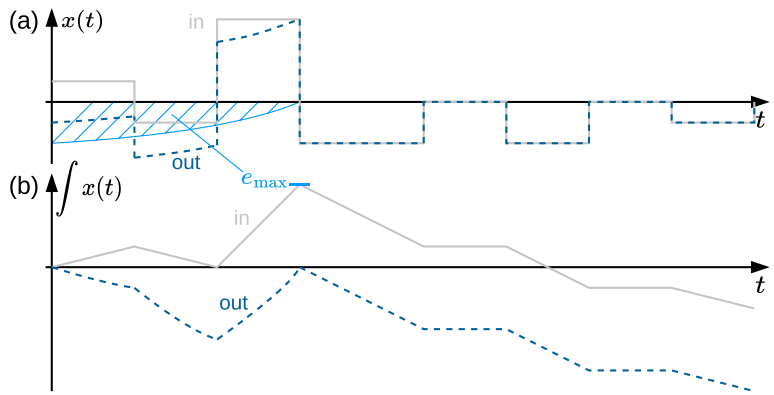


Fig. 4 Principle on how to fix the boundary conditions of a time series, so that the integral does not exceed zero. **a** shows the time series and **b** the corresponding integral. Determine the maximum integral of the grey input time series and superpose a sine-quarter-wave with the same integral to the beginning of the original time series to gain the dark-blue and dashed output time series

Decluster step

This section outlines the *decluster* step, beginning with some preprocessing measures, followed by the problem definition for the decluster algorithm. Finally, the algorithm designed to solve the previously defined objective is discussed.

Preprocessing

Preparation steps include *normalization* in both the time domain and the feature space and *dimensionality reduction* of the feature space. The time series are normalized to a maximum absolute value of one with a non-zero mean. The non-zero mean poses a challenge since some features are sensitive to the *mean* and the *standard deviation* (Fulcher et al. 2013). A common approach to mitigate this issue involves *z-score* normalization of the time series prior to feature calculation, given by the formula (Fulcher et al. 2013)

$$\tilde{y}(t) = \frac{x(t) - \bar{x}(t)}{\sigma_{x(t)}} \tag{1}$$

where $\bar{x}(t)$ denotes the *mean* and $\sigma_{x(t)}$ the *standard deviation* of the time series $x(t)$. This approach is also employed in previous works (Fulcher et al. 2013; Lubba et al. 2019), and the features calculated using z-scored data are documented in the additional material.

The feature space is also normalized as each feature vector across the time series set can possess highly differing ranges and skewed distributions, potentially with outliers. Differing ranges and outliers are a concern because distance measures in the feature space will overly emphasize features with large ranges (Murphy 2013; Hastie et al. 2009). A common method to address this is to normalize the feature vector, too (Murphy 2013; Hastie et al. 2009). In this study, an *outlier-robust sigmoidal transformation* is employed, given by Fulcher et al. (2013)

$$\hat{f} = \left(1 + \exp\left(-\frac{f - \text{median(f)}}{1.35 \cdot \text{iqr(f)}}\right)\right)^{-1} \tag{2}$$

where f denotes the vector of a single feature of all time series of the data set, median denotes the *median* and iqr denotes the *inter-quartile range*. This nonlinear transformation maps an unbounded feature space to the interval [0, 1], approximately maintaining linearity within the *inter-quartile range* and compressing outliers to the interval bounds. This method is similar to the logistic sigmoid but is robust due to its dependence on the *median* and *inter-quartile range* instead of *mean* and *standard deviation* (Fulcher et al. 2013).

The choice of normalization directly impacts the ensuing algorithm's objective of uniformly distributing points within a nonlinear space. When mapped back to linear space, the nonlinearity leads to a thinning out of points in the boundary regions. However, this is considered acceptable because most of the feature space remains densely populated, and the alternative, dismissing outliers, would result in information loss.

Following normalization, dimensionality reduction becomes essential. In this study, a 119-element feature vector represents each time series, creating a 119-dimensional feature space. The ambition to establish a uniformly distributed set in such a high-dimensional space encounters challenges due to the curse of dimensionality (Houle et al.

2010). With increasing dimensions, the feature space becomes sparser, necessitating an exponentially larger number of points for adequate coverage. Additionally, traditional distance measures become less meaningful, computational complexity proliferates, and redundant or irrelevant attributes introduce noise (Duda et al. 2001; Witten et al. 2017).

To address these issues, this study adopts dimensionality reduction techniques that retain only the most relevant dimensions, eliminate highly correlating dimensions, and consequently reduce noise and absolute dimensions within the feature space (Duda et al. 2001; Witten et al. 2017). To accomplish dimensionality reduction, a *hierarchical clustered correlation matrix* is utilized (Müllner 2011; Bar-Joseph et al. 2001), a method also employed in prior works (Fulcher et al. 2013; Lubba et al. 2019). This approach computes the Pearson correlation between all features and arranges them in a matrix such that highly correlated features are adjacent.

Figure 5 presents this clustered matrix, framing 14 clusters with orange squares. Within each cluster, features are sorted by their correlation strength to other features in the same cluster. The representative feature for each cluster is indicated with an orange square on the diagonal and is generally the feature with the highest intra-cluster correlation. Exceptions are made when a different feature demonstrates a significantly lower correlation with outer-cluster features or is easier to interpret while maintaining

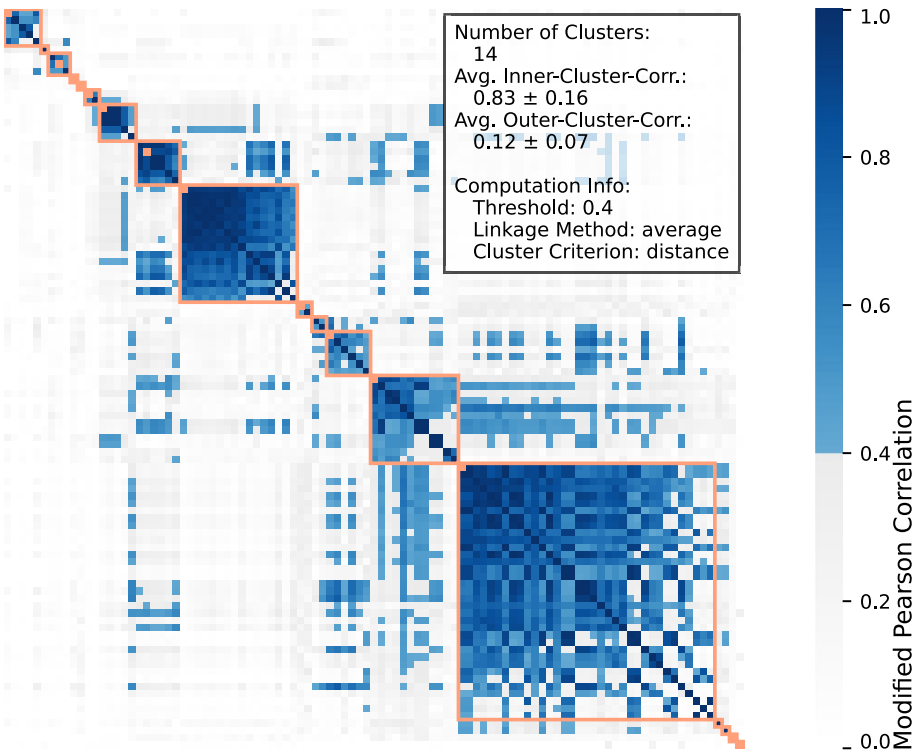


Fig. 5 Hierarchical clustered correlation matrix employed for dimensionality reduction, (cp. Fulcher et al. (2013); Lubba et al. (2019)). Fourteen clusters, outlined with orange squares, emerge from the original 119-dimensional feature space. Color encodes modified pearson correlation between features. Representative features for clusters are marked on the diagonal. These are (from top left to bottom right): temporal_centroid, loc_of_last_max, dfa, rs_range, mode5, share_above_mean, iqr, mean, rcp_num, acf_first_zero, median_of_abs_diff, freq_mean, mean_2nd_diff, trev. See the additional material for more information

comparable performance. This study further refined the correlation matrix to include non-linear correlations. Specifically, matrices for $1/x$, $\sqrt{x}$, and x^2 correlation are also computed. The final combined correlation matrix takes the maximum correlation value from these individual matrices. The 14 selected cluster representatives (cp. Fig. 5) determined in this phase form the prerequisite for the subsequent step.

Problem definition

A low-discrepancy subset shall be extracted from a superset in n-dimensional space, which in this case comprises 14 feature-dimensions. The optimization problem can be stated as follows:

Given a search space set $\mathbb{M}$ consisting of m points – specifically, the manifold 2^{19}-element time series set in this context – in n-dimensional space, the objective is to find a low-discrepancy subset $\mathbb{K}$ containing k points, where $k < m$, that minimizes the functional

$$\sum_i \sum_j \frac{1}{r_{ij}^2} \quad , \tag{3}$$

where r_{ij} is the distance metric (euclidean, manhattan, etc.) between two points i and j in the subset $\mathbb{K}$. This problem is akin to a variant of the k-minimum-spanning-tree problem and has been proven to be NP-hard (Ravi et al. 1994; Chlebík and Chlebíková 2008). The distance calculation is computationally expensive as well, operating in polynomial time $\mathcal{O}(nk^2)$.

To address this, an alternative metric named *heterogeneity* h_{nb} is introduced:

$$h_{nb} = \frac{1}{n} \sum_n \sum_b \left(H_{nb} - \frac{1}{b} \right)^2 \quad , \tag{4}$$

where H_{nb} is a normalized histogram array with b bins over a set with n dimensions. This metric minimizes to zero when each bin in each dimension contains an equal fraction of $\frac{1}{b}$ points and disproportionately penalizes bins that deviate significantly from the average.

The alternative heterogeneity optimization target only approximates the original measure and may introduce additional correlations between previously uncorrelated dimensions. These drawbacks are acceptable, as the distance measure in the exact optimization problem becomes increasingly vague in higher dimensions, with unpredictable consequences. The alternative metric primarily enhances the uniformity or low-discrepancy in each dimension.

It should be noted that the minimization problem will not yield a perfectly uniform distribution; instead, it will find the most uniform distribution possible within the volume of the given superset. Due to the NP-hard nature of the problem, the algorithm outlined in the following section will only approximate the optimal solution.

Decluster algorithm

The algorithm implemented in this study follows the paradigm of classical genetic optimization algorithms, albeit without crossover and relying solely on mutation. First, an initial population consisting of one candidate with k elements or time series is generated

from the search space set $\mathbb{M}$. The fitness of this candidate is calculated, and l new candidates are generated through mutation. Subsequently, the fitness is calculated for each new candidate, and the one with the best fitness is chosen. This process is iterated until the fitness ceases to improve.

The fitness function employed here evaluates the heterogeneity as defined in Equation (4). Additionally, custom initialization and mutation steps have been implemented, which are detailed below.

The initial population is generated using a quasi-random Halton sequence (Owen 2017) in the n-dimensional feature space, with each point mapped to its nearest neighbor in the search space set $\mathbb{M}$. More details and parameters can be found in the additional material.

The custom mutation step is designed to efficiently improve fitness or heterogeneity (as in Equation (4)) by directly attempting to equalize the n-dimensional histogram as follows: A selection probability is assigned to each bin of each dimension of the n -dimensional histogram, proportional to the sparsity of that bin. Upon selecting a bin, a point from the search space set $\mathbb{M}$ is chosen and inserted to fall within the bin's edges. Afterward, another point is removed from the candidate set in the same manner but based on a probability that is assigned to each bin in proportion to the density of the bin. It should be noted that these insertion and deletion steps affect not only the chosen bin but also corresponding bins in other dimensions, potentially leading to a degradation in fitness.

Resulting sets of time series

After thoroughly outlining the methodology, we now focus on the data sets resulting from the decluster algorithm applied to the manifolded set, highlighting its performance in feature space heterogeneity. This performance is evident in Fig. 6, depicting the n-dimensional histograms visualizing the distribution in feature space for the

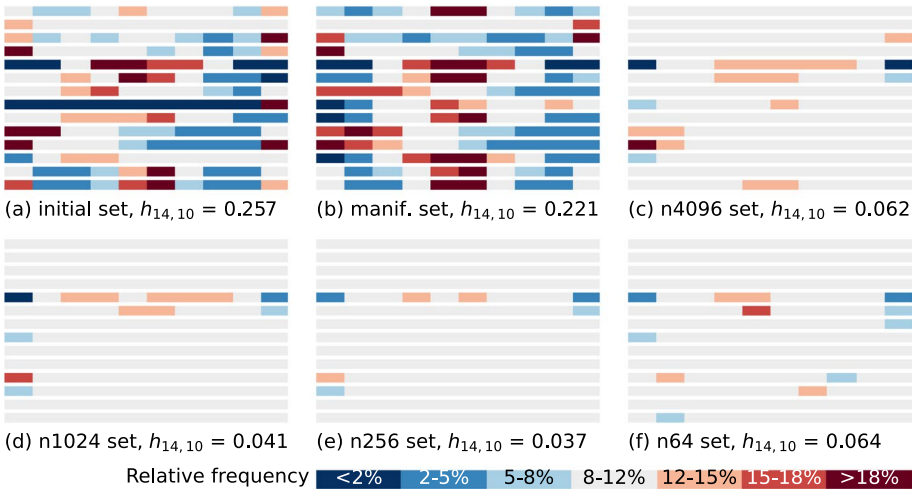


Fig. 6 The n-dimensional histograms of **a** initial, **b** manifolded, and **c**–**f** declustered sets. Blue/red shades indicate bins that are too sparse/dense. Feature names (rows) of the individual dimensions from top to bottom are documented in Fig. 5. Columns represent the value range of each feature. In **a**, **b**: many clusters and skewed distributions; **c**–**f**: improved uniformity, enhancing with smaller sets

various sets. The color encoding for the histograms in Fig. 6 distinguishes bins as follows: Grey shades mark bins that contain 8% to 12% of all points, which corresponds to the expected number of points in a uniform distribution within a 10-bin histogram and with a tolerance of ±2%. Blue shades signify a downward deviation from this optimal uniform distribution. Darker shades of blue indicate higher levels of deviation. Red shades indicate similarly an upward deviation from the optimal uniform distribution. The rows in Fig. 6a to Fig. 6f represent the distribution in the cluster-representative feature-dimensions that are documented in Fig. 5.

In Fig. 6a, the *n*-dimensional histogram of the initial set displays dimensions that are fairly uniformly distributed, especially in dimensions 1 to 3, where each bin contains 5% to 15% of all points. However, most dimensions are highly clustered and skewed, notably dimensions 5 and 8. Similar general characteristics are observed for the manifolded set represented in Fig. 6b. The critical difference lies in the increased number of points available for selection during the decluster step to generate the sets shown in Fig. 6c to Fig. 6f.

Figure 6c displays the largest declustered set comprising 4096 points, furtherly abbreviated as `n4096set`. Most dimensions exhibit a uniform distribution with only minor skewness, specifically in dimensions 5 and 11. In a weakened form, these slight deviations echo the irregularities found in the manifolded set. The histograms for smaller subsets comprising 1024 and 256 points, furtherly abbreviated as `n1024set` and `n256set` and shown in Fig. 6d and Fig. 6e, respectively, converge closely to an ideal uniform distribution, exhibiting only minor deviations. In the smallest subset, denoted as `n64set` and displayed in Fig. 6f, the heterogeneity is relatively low albeit not as optimal as in the superset `n256set`. Two reasons account for this: first, an aliasing effect caused by representing 64 points in a 10-bin histogram, as 64 and 10 are not integer-divisible, and second, the impact of a single point in a 14-dimensional space becomes significant due to the limited number of points. Despite these challenges, the `n64set` remains fairly uniform and representative.

Additionally, the feature space can be represented by an *n*-dimensional histogram of all 119 dimensions, not just the 14 dimensions defined by the cluster representatives. Although the heterogeneity is less optimal, the decluster effect is still discernible in the omitted dimensions. For instance, the initial set has a 119-dimensional heterogeneity of 0.35 with a standard deviation of 0.16 in the individual dimensions, while `n256set` achieves a 110-dimensional heterogeneity of 0.17 with a standard deviation of 0.09.

Figure 7 presents the subset comprising the 32 signed time series from the smallest declustered `n64set`, displayed in the time domain to provide a visual representation of the results. These time series are sorted from top to bottom and from left to right according to their *mean*.

At first glance, the time series in Fig. 7 appear to be quite similar to those in the initial data set shown in Fig. 3. However, a comparison between Fig. 6a and Fig. 6f reveals that this subset is much more uniform in feature space, despite having fewer elements. Importantly, it also covers a larger volume of the feature space. The observed similarity in the time domain can also be corroborated in the frequency domain, although these details have been omitted for conciseness and can be found in the additional material.

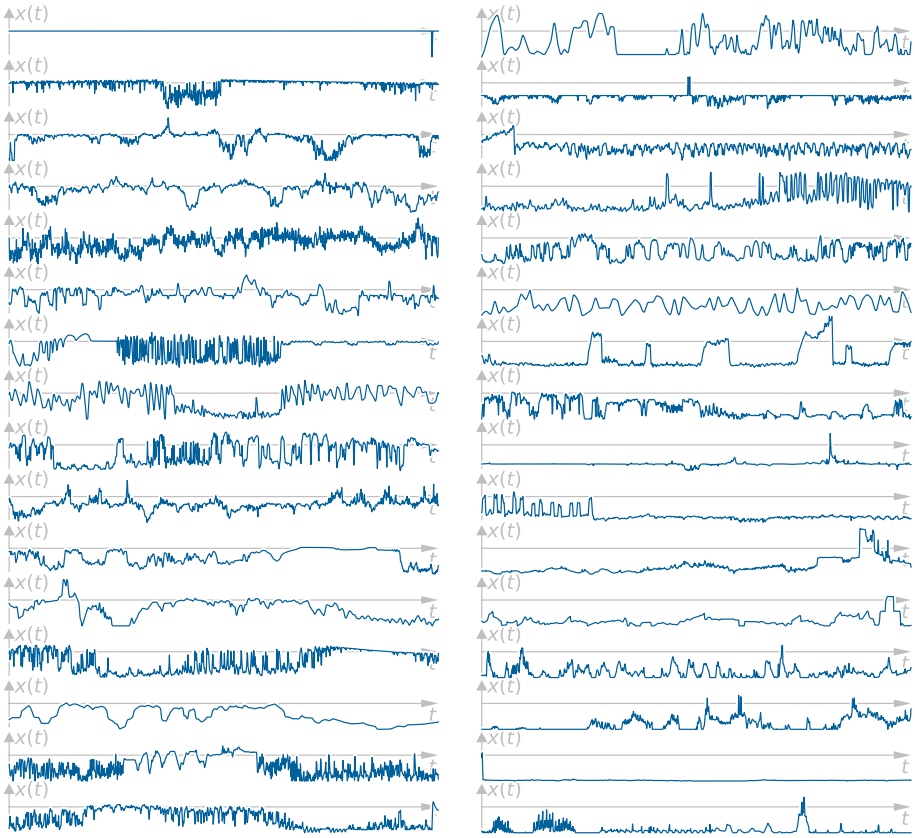


Fig. 7 Time series of the subset comprising the 32 signed time series of the smallest declustered `n64set`, showing good compliance in time and frequency domain with the initial set (cp. Fig. 3) but exposing much better distribution in feature space (cp. Fig. 6)

Application and usage

The effectiveness of the declustered data set is demonstrated through two examples: a hybrid energy storage system (HESS) and a hydrogen production system. These examples demonstrate that abstract data can yield application-specific insights. Additionally, general guidelines for working with the data sets are derived from these examples.

Hybrid energy storage systems

"A HESS combines two different energy storage technologies into a single storage system to increase the performance of the overall storage system, decrease costs and dimensions and increase the overall systems lifetime, efficiency, and response time". Günther et al. (2022). Previous work by the authors (Günther et al. 2018) shows that the feasibility of implementing a HESS depends on the application or the input time series. Some applications offer few opportunities for effective HESS implementation and are thus unsuitable, while others present multiple options and high suitability. The work introduces a metric called *hybridisation potential P*, ranging from 0, indicating no suitability for HESS, to 1, indicating high suitability for HESS. For a more detailed discussion, the reader is referred to Günther et al. (2018).

This example aims to identify the factors or time series feature that dictate the hybridisation potential P. This identification is achieved by mapping the hybridisation potential P as an additional feature onto the declustered set. We then seek functional relationships between this and existing features. Calculating the correlation with all other features reveals high-correlating pairs (with a correlation greater than 0.8) such as *mean*, *rms*, or *ecdf20*, as well as low-correlating pairs (with a correlation less than 0.1) like *embedding distance*, *acf1*, or *seasonality strength* (see the additional material for a description of these features). Visualizing these high-correlating pairs through scatterplots identifies the most meaningful relation, as shown in Fig. 8.

From Fig. 8, we derive a functional relationship

$$P \leq 1 - \overline{x}/\hat{x} \quad \forall\, x(t) \in \texttt{n4096set} \quad , \tag{5}$$

establishing an upper limit. Here, $\hat{x}$ denotes the absolute maximum of the time series. Although most time series are near this line, deviations down to zero are possible. Exact quantiles can be read directly from the figure.

For practitioners, this translates to the following actionable insights: if the normalized mean of the application is high, considering a HESS is unwarranted and the analysis can be aborted at this early stage. Conversely, if it is low, investigating a HESS further is worthwhile, with a high probability of identifying a suitable HESS configuration for successful implementation.

Exercise caution when making quantitative assertions: For instance, the assertion that 50% of the time series deviate by only 10% from the established upper limit in Equation (5) holds true only for the chosen data set that covers a large volume in feature space. An application-specific data set covering only a fraction of the feature space may exhibit different behavior. Nonetheless, the upper limit itself appears to be a relatively safe assumption, irrespective of the data set in question. This assertion is by no means a proof from a mathematical standpoint but an acceptable axiom from an engineering perspective.

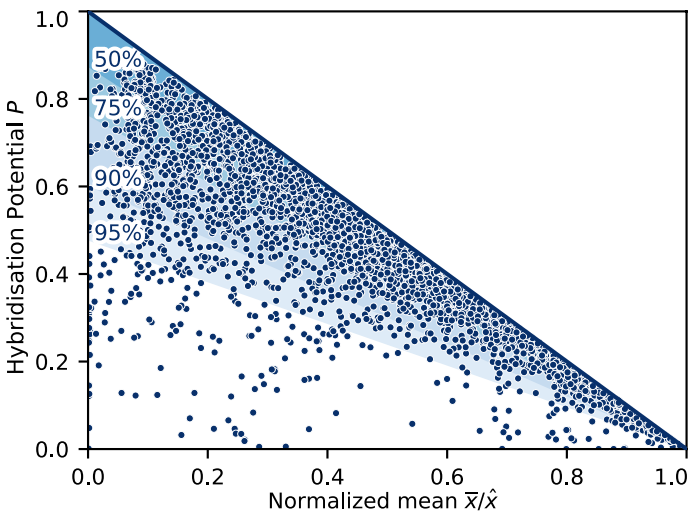


Fig. 8 Mapping the target feature hybridisation potential P to the time series feature *normalized mean* $\overline{x}/\hat{x}$ in a scatterplot for each time series in the `n4096set`. A strong correlation is evident with $P \sim 1 - \overline{x}/\hat{x}$

Various approaches for advancing this analysis exist but are outside the scope of this work. One possible path is to identify a second influencing variable that explains deviations from the threshold. This could be accomplished by manually inspecting further strong correlating features or outliers within the time domain to derive the underlying reasons and influences for deviations from the threshold. The search for further influencing variables could also include classic machine learning methods, such as linear and nonlinear classifiers, including linear regression, elastic net, Bayesian regressors, or generalized linear models models (Hastie et al. 2009). Feature selection and elimination techniques can also be employed (Hastie et al. 2009).

A question remains about the necessity of employing a specialized data set for the analysis instead of using a randomly selected one. Figure 9 displays the same scatterplot as in Fig. 8 but contrasts the declustered `n64set` (Fig. 9a), a random selection of 64 time series from the initial set (Fig. 9b), and a random selection of 64,000 time series from the manifold set (Fig. 9c).

The relationship in Equation (5) depicted in Fig. 8 and derived from the `n4096set` is also reasonably represented by the `n64set` in Fig. 9a. The initial set in Fig. 9b still reveals the correlation, albeit with increased uncertainty due to empty or unknown regions. Due to the curtailed boundaries, deriving the upper limit and functional relationship is difficult. It should be noted that the visibility of the correlation in Fig. 9b is coincidental, stemming from the strong correlation between hybridisation potential P and normalized mean $\overline{x}/\hat{x}$, and the reasonably-distributed nature of the *mean*-dimension in the initial set, as opposed to many other dimensions, cp. Fig. 6. Figure 9c shows that the large data set yields the same relationship as the `n4096set`, albeit at a higher computational cost, which may not be feasible for many analyses. Furthermore, this amount of data is likely unavailable in most studies.

Another operational guideline can be inferred: for one's study, employ the largest declustered data set that is computationally feasible. Initial explorations might begin with the smallest set, with the results being refined using larger data sets in subsequent stages.

Hydrogen production system

To demonstrate the data set's applicability to another technical application, we use the optimization model presented by Brandt et al. 2023. This model minimizes the

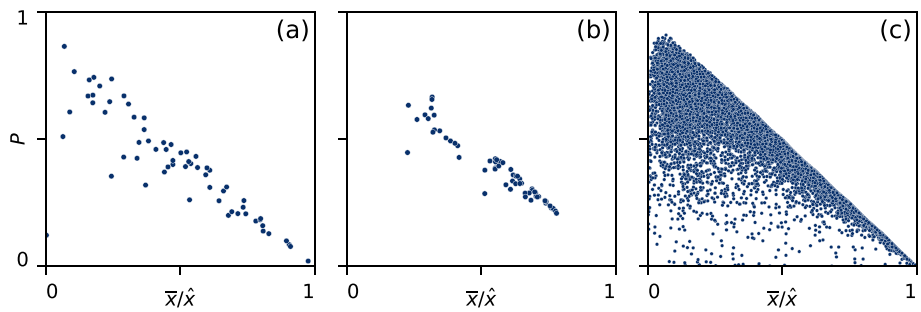


Fig. 9 Scatterplots that show hybridisation potential vs. normalized mean for different sets: **a** the declustered `n64set`; **b** a random 64-point subset of the initial set; **c** a random 64,000-point subset of the manifold set

onsite hydrogen supply cost (OHSC) C to meet a given hydrogen demand and optimizes the design and operation of a hydrogen production system for this purpose. The system includes various power sources such as utility-scale photovoltaic, wind turbines and a connection to the electricity grid, and an electrolysis system, compressor, and pressure tank for hydrogen storage. In this analysis, we reassess the impact of demand characteristics. To achieve this, we adapt the original model by selecting 2048 time series, characterized by exclusively negative values, from the declustered `n4096set`. This selection mirrors the exclusive demand nature of the problem. These profiles span a year, have hourly resolution, and are normalized to an average demand of $10 t_{H_2}$/day.

Generally, it is possible to rescale the time series along the x- and t- axes to meet the specific requirements. Upscaling in time domain is generally unproblematic; however, exercise caution when downscaling: Reducing the scale could result in information loss, potentially altering time series characteristics and features. Downscaling by a factor of two is generally acceptable for the presented sets without introducing significant errors; higher levels of downscaling are discouraged.

In Brandt et al. 2023, a variable importance analysis (VIA) was used to determine the most impactful uncertain input parameters on the OHSC. These included, among others, fluctuating electricity prices, renewable energy availability, and varying demand profiles. The uncertainty of demand profiles in the original study was limited to constant, daily, and weekly variations. Figure 10a reveals that the influence of their uncertainty on the OHSC was minimal compared to other parameters. The first-order Sobol' index in the figure indicates the individual impact of each uncertain parameter, while the total Sobol' index considers combinatorial effects with other parameters.

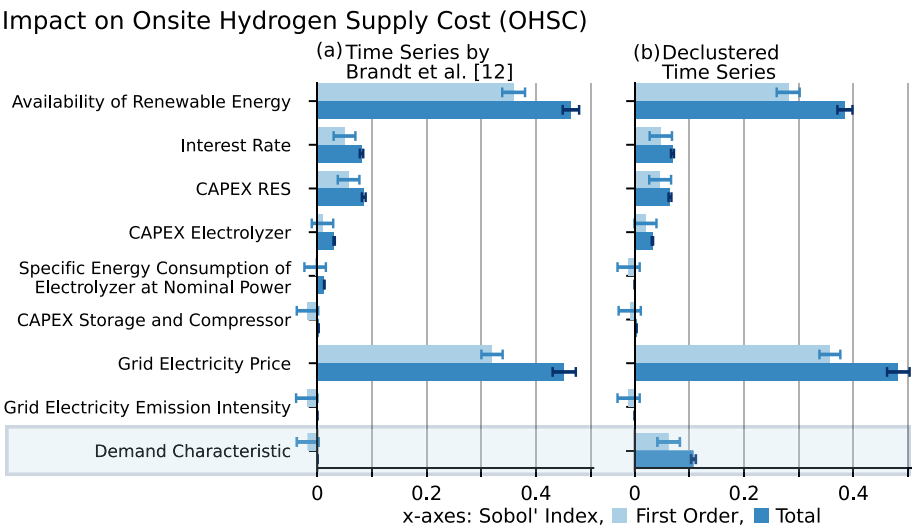


Fig. 10 Results of the VIA on the OHSC. **a** Original VIA results featuring constant, daily and weekly changing time series (Brandt et al. 2023). **b** Updated results utilizing the declustered set of time series. The first-order Sobol' index measures individual parameter impact on model output, while the total index additionally considers combinatorial effects with other uncertain input parameters. The error bars represent bootstrap confidence intervals with a confidence level of 95%. A comparison of **a**, **b** shows a significant increase in the impact of demand profile uncertainty on the OHSC due to the increased diversification of demand characteristics

When the VIA is rerun using the newly generated set of more declustered demand profiles, Fig. 10b shows a notable increase in the impact of the uncertainty in demand profiles on the resulting OHSC. The broader range of demand characteristics leads to a greater dependency of supply cost on these profiles. Thus, demand profiles that differ from constant or periodic patterns can significantly vary the resulting OHSC.

To investigate which features of the demand profiles contribute to the increased variability in OHSC, we conducted optimizations for all 2048 profiles of the declustered set. All other uncertain input parameters were held constant, based on the values defined in the parameter studies by Brandt et al. 2023.

Figure 11 reveals that the normalized mean $\overline{x}/\hat{x}$ strongly influences the variability in OHSC of the respective demand profile. Specifically, the dispersion increases exponentially with a decreasing normalized mean $\overline{x}/\hat{x}$ value, indicating that the uncertainty in OHSC also rises with a decreasing normalized mean $\overline{x}/\hat{x}$ value. Practically speaking, if the normalized mean $\overline{x}/\hat{x}$ value of a specific demand profile is higher than 0.6, the resulting OHSC will not exceed 8.5€/kg. However, for normalized mean $\overline{x}/\hat{x}$ values between 0.2 and 0.4, the cost can rise to 13€/kg and even reach 37€/kg for lower normalized mean $\overline{x}/\hat{x}$ values.

Additionally, a clear trend emerges: the higher the necessity for grid electricity integration, the higher the resulting supply cost (cp. color-coding of Fig. 11). This trend implies that if the demand profile does not correlate well with available renewable energy, the deficit must be balanced by comparably expensive grid electricity. Another practical observation from the plot indicates that, irrespective of how low the normalized mean $\overline{x}/\hat{x}$ of the demand profile is, achieving a supply cost lower than 6.3€/kg is not realizable within the boundaries of the model and its chosen parameters. Note that the mean was intentionally selected as a feature for simplicity. While other correlations and insights are present in the data, they fall outside the scope of the current discussion.

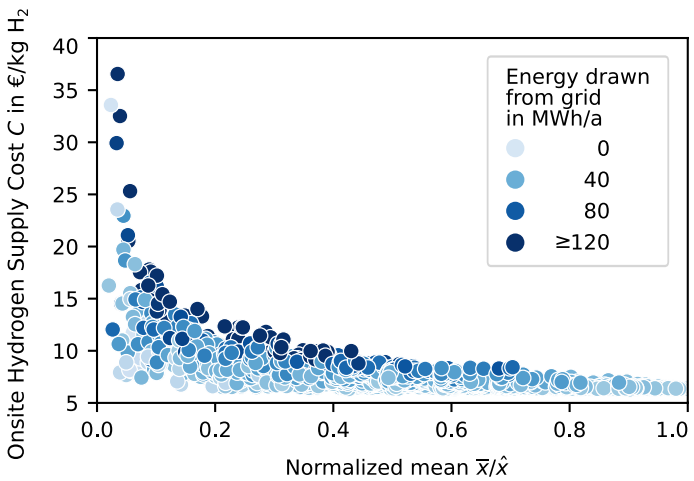


Fig. 11 Optimization results for 2048 declustered time series showing OHSC vs. normalized mean $\overline{x}/\hat{x}$ as a scatterplot. The color of each point encodes the *energy drawn from the grid* besides that from optimized renewable energy sources, to meet the given hydrogen demand. An exponential increase in OHSC variability with decreasing normalized mean $\overline{x}/\hat{x}$ of the demand profiles is evident

Summary

The objective of this paper was to create a set of application-independent load profiles or time series, specifically within the context of energy systems and based on real-world data. The creation involved two principal stages: (a) manifolding the initial data to increase the diversity and heterogeneity of the data using signal processors like expanders, compressors, and limiters; and (b) subsequently declustering this data by eliminating redundant points in densely clustered regions through a genetic optimization algorithm aimed at minimizing set discrepancy. The study employed standard feature engineering and machine learning techniques: (a) transforming the time series into a normalized feature space, (b) executing dimensionality reduction via hierarchical clustering, and (c) performing optimization in this reduced feature space.

The outcome of this research is a set of load profiles or time series that is approximately uniformly distributed across each dimension of the feature space by simultaneously retaining typical time and frequency domain characteristics commonly found in energy system time series. These generated time series serve multiple purposes: they can be used standalone for method and algorithm exploration, enable the identification of correlations and functional relationships to time series features, facilitate the training of machine learning models with an unbiased dataset, and act as supplementary data to yield additional, more profound insights into specific problems.

Two examples substantiate the utility of these time series. The first involved a method to classify the suitability of a hybrid energy storage system for an application, revealing a strong correlation to the mean of a time series and a corresponding upper limit. The second example provided an uncertainty quantification optimization model used to minimize the onsite hydrogen supply cost to meet a predefined demand time series. Utilising the created set of time series refined previous results by attributing greater weight to the influence of the time series on onsite hydrogen supply cost and showed that the variance is highly dependent on the mean of the time series.

Both the manifold and the decluster algorithms presented in this paper have broader applicability; they can be customized and applied to various other problems. The decluster algorithm, while serving as a means to an end in this research, has inherent scientific merit worthy of independent investigation. We anticipate significant improvements in the speed and quality of this algorithm. The declustered dataset produced in this work could be enhanced through (a) hyperparameter optimizations, (b) incorporation of a more extensive and diverse input dataset, and (c) refining the utilized feature set. Moreover, the methodology for creating declustered datasets is generally applicable across domains beyond energy systems.

This work also provides smaller subsets of the dataset to facilitate computationally intensive studies, thereby offering a trade-off between accuracy and computational speed. Researchers are relieved from curating individual databases as the dataset and generating source code are openly accessible, promoting potential comparability with future and past studies.

Abbreviations

catch22	Canonical time-series characteristics
DC	Direct current
feasts	Feature extraction and statistics for time series

HESS	Hybrid energy storage system
hctsa	Highly comparative time-series analysis
`n64set`	Declustered set comprising 64 points
`n256set`	Declustered set comprising 256 points
`n1024set`	Declustered set comprising 1024 points
`n4096set`	Declustered set comprising 4096 points
OHSC	Onsite hydrogren supply cost
theft	Tools for handling extraction of features from time series
tsfeatures	Time series feature extraction
tsfresh	Time series feature extraction based on scalable hypothesis tests
VIA	Variable importance analysis

Acknowledgements
Not applicable.

Author contributions
SG developed, realized, and implemented the presented concept and method and designed and provided the first application example. JB designed and provided the second application example. The manuscript writing and visualization is in line with the specified contributions. The presented work was supported and supervised by AB and RHR. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was supported and funded by the German Federal Ministry of Education and Research (BMBF) under the grant 03SF0560A. The publication of this article was funded by the Open Access Fond of Leibniz University Hannover in context of the project DEAL.

Data availibility
The datasets generated and analysed during the current study are available in the ESTSS repository, https://github.com/s-guenther/estss and https://zenodo.org/records/10213145.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 21 December 2023 Accepted: 6 January 2024
Published online: 22 January 2024

References

Ammari C, Belatrache D, Touhami B, Makhloufi S (2022) Sizing, optimization, control and energy management of hybrid renewable energy system - a review. Energy Built Environ 3(4):399–411. https://doi.org/10.1016/j.enbenv.2021.04.002

André M (2004) The ARTEMIS European driving cycles for measuring car pollutant emissions. Sci Total Environ 334–335:73–84. https://doi.org/10.1016/j.scitotenv.2004.04.070

Angizeh F (2020) Dataset on Hourly Load Profiles for a Set of 24 Facilities from Industrial, Commercial, and Residential End-use Sectors. Mendeley. https://doi.org/10.17632/RFNP2D3KJP.1 . https://data.mendeley.com/datasets/rfnp2d3kjp/1 Accessed 10 Apr 2023

Anoune K, Bouya M, Astito A, Abdellah AB (2018) Sizing methods and optimization techniques for PV-wind based hybrid renewable energy system: a review. Renew Sustain Energy Rev 93:652–673. https://doi.org/10.1016/j.rser.2018.05.032

Anvari M, Proedrou E, Schäfer B, Beck C, Kantz H, Timme M (2022) Data-driven load profiles and the dynamics of residential electricity consumption. Nat Commun 13(1):4593. https://doi.org/10.1038/s41467-022-31942-9

Armstrong MM, Swinton MC, Ribberink H, Beausoleil-Morrison I, Millette J (2009) Synthetically derived profiles for representing occupant-driven electric loads in Canadian housing. J Build Perform Simul 2(1):15–30. https://doi.org/10.1080/19401490802706653

Barandas M, Folgado D, Fernandes L, Santos S, Abreu M, Bota P, Liu H, Schultz T, Gamboa H (2020) TSFEL: time series feature extraction library. SoftwareX 11:100456. https://doi.org/10.1016/j.softx.2020.100456

Bar-Joseph Z, Gifford DK, Jaakkola TS (2001) Fast optimal leaf ordering for hierarchical clustering. Bioinformatics 17(suppl 1):22–29. https://doi.org/10.1093/bioinformatics/17.suppl_1.S22

Behm C, Nolting L, Praktiknjo A (2020) How to model European electricity load profiles using artificial neural networks. Appl Energy 277:115564. https://doi.org/10.1016/j.apenergy.2020.115564

Binderbauer PJ, Kienberger T, Staubmann T (2022) Synthetic load profile generation for production chains in energy intensive industrial subsectors via a bottom-up approach. J Clean Prod 331:130024. https://doi.org/10.1016/j.jclepro.2021.130024

Braeuer F (2020) Load profile data of 50 industrial plants in Germany for one year. Zenodo. https://doi.org/10.5281/ZENODO.3899018
Brandt J, Iversen T, Eckert C, Peterssen F, Bensmann B, Bensmann A, Beer M, Weyer H, Hanke-Rauschenbach R Cost and competitiveness of green hydrogen in europe: effects of the european union regulatory framework https://doi.org/10.21203/rs.3.rs-3164444/v1
Brusco MJ, Cradit JD, Steinley D (2020) Combining diversity and dispersion criteria for anticlustering: a bicriterion approach. Br J Math Stat Psychol 73(3):375–396. https://doi.org/10.1111/bmsp.12186
Chlebík M, Chlebíková J (2008) The Steiner tree problem on graphs: inapproximability results. Theor Comput Sci 406(3):207–214. https://doi.org/10.1016/j.tcs.2008.06.046
Christ M, Braun N, Neuffer J, Kempa-Liehr AW (2018) Time series feature extraction on basis of scalable hypothesis tests (tsfresh - APython package). Neurocomputing 307:72–77. https://doi.org/10.1016/j.neucom.2018.03.067
Drmota M, Tichy RF (1997) Sequences, discrepancies, and applications. Lecture notes in mathematics, vol. 1651. New York: Springer, Berlin
Duda RO, Hart PE, Stork DG (2001) Pattern classification. A Wiley-Interscience publication, 2nd edn. John Wiley & Sons Inc, New York Chichester Weinheim Brisbane Singapore Toronto
Fischer D, Härtl A, Wille-Haussmann B (2015) Model for electric load profiles with high time resolution for German households. Energy Build 92:170–179. https://doi.org/10.1016/j.enbuild.2015.01.058
Fritsch FN, Butland J (1984) A method for constructing local monotone piecewise cubic interpolants. SIAM J Sci Stat Comput 5(2):300–304. https://doi.org/10.1137/0905021
Fulcher BD, Jones NS (2017) HCTSA : a computational framework for automated time-series phenotyping using massive feature extraction. Cell Syst 5(5):527–5313. https://doi.org/10.1016/j.cels.2017.10.001
Fulcher BD, Little MA, Jones NS (2013) Highly comparative time-series analysis: the empirical structure of time series and their methods. J R Soc Interface 10(83):20130048. https://doi.org/10.1098/rsif.2013.0048
Fulcher B, Cliff O, Harris B, Philiphorst Sethi S, Lubba CH, Alam I, Lukas Vysyaraju KP, McCormac J (2023) VP007-Py, Xavier-FPMorris, Kaede Shiina: benfulcher/hctsa: v1.09. Zenodo. https://zenodo.org/record/8155940 Accessed 10 May 2023
Giorgi L, Obushevs A, Korba P (2021) Electric Vehicles Load Profile Generator Based on the Probability Density Functions. In: 2021 IEEE 62nd International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON), IEEE, Riga, Latvia. pp. 1–8. https://doi.org/10.1109/RTUCON53541.2021.9711591
Gotzens F, Gillessen B, Burges, S., Hennings, W., Müller-Kirchenbauer J, Seim S, Verwiebe P, Tobias S, Jetter F, Limmer T DemandRegio - Harmonisierung und Entwicklung von Verfahren zur regionalen und zeitlichen Auflösung von Energienachfragen : Abschlussbericht. IEK-STE, ITM, E &R, FfE München, BMWi (2020) https://doi.org/10.34805/ffe-119-20
Grandjean A, Adnot J, Binet G (2012) A review and an analysis of the residential electric load curve models. Renew Sustain Energy Rev 16(9):6539–6565. https://doi.org/10.1016/j.rser.2012.08.013
Granell R, Axon CJ, Wallom DCH (2015) Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. IEEE Trans Power Syst 30(6):3217–3224. https://doi.org/10.1109/TPWRS.2014.2377213
Günther S, Bensmann A, Hanke-Rauschenbach R (2018) Theoretical dimensioning and sizing limits of hybrid energy storage systems. Appl Energy 210:127–137. https://doi.org/10.1016/j.apenergy.2017.10.116
Günther S, Weber L, Bensmann AL, Hanke-Rauschenbach R (2022) Structured analysis and review of filter-based control strategies for hybrid energy storage systems. IEEE Access 10:126269–126284. https://doi.org/10.1109/ACCESS.2022.3226261
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer Series in Statistics. Springer, New York, NY
Henderson T, Fulcher BD (2022) Feature-based time-series analysis in R using the theft package. https://doi.org/10.48550/ARXIV.2208.06146
Hoogsteen G, Molderink A, Hurink JL, Smit GJM (2016) Generation of flexible domestic load profiles to evaluate Demand Side Management approaches. In: 2016 IEEE International Energy Conference (ENERGYCON), IEEE, Leuven, Belgium. pp. 1–6. https://doi.org/10.1109/ENERGYCON.2016.7513873
Houle ME, Kriegel H-P, Kröger P, Schubert E, Zimek A (2010) Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Vardi MY, Weikum G, Gertz M, Ludäscher B (eds.) Scientific and statistical database management vol. 6187, pp. 482–500. Springer, Berlin, Heidelberg. http://link.springer.com/10.1007/978-3-642-13818-8_34 Accessed 10 May 2023
Huber J, Klempp N, Becker J, Weinhardt C (2019) Electricity consumption of 28 German companies in 15-min resolution. Karlsruhe. https://doi.org/10.5445/IR/1000098027
Hülk L, Müller B, Glauer M, Förster E, Schachler B (2018) Transparency, reproducibility, and quality of energy system analyses-a process to improve scientific work. Energy Strat Rev 22:264–269. https://doi.org/10.1016/j.esr.2018.08.014
Hyndman R, Kang Y, Montero-Manso P, O'Hara-Wild M, Talagala T, Wang E, Yang Y (2023) Tsfeatures: time series feature extraction. https://pkg.robjhyndman.com/tsfeatures/, https://github.com/robjhyndman/tsfeatures. Accessed 10 May 2023
Intergovernmental Panel On Climate Change (Ipcc) (2023) Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 1st edn. Cambridge University Press. https://www.cambridge.org/core/product/identifier/9781009325844/type/book Accessed 10 May 2023
Islam SN, Rahman A, Robinson L (2020) Load Profile Segmentation using Residential Energy Consumption Data. In: 2020 International Conference on Smart Grids and Energy Systems (SGES), IEEE, Perth, Australia, pp. 600–605. https://doi.org/10.1109/SGES51519.2020.00112
Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A (2019) Deep learning for time series classification: a review. Data Min Knowl Discov 33(4):917–963. https://doi.org/10.1007/s10618-019-00619-1
Jiang X, Srivastava S, Chatterjee S, Yu Y, Handler J, Zhang P, Bopardikar R, Li D, Lin Y, Thakore U, Brundage M, Holt G, Komurlu C, Nagalla R, Wang Z, Sun H, Gao P, Cheung W, Gao J, Wang Q, Guerard M, Kazemi M, Chen Y, Zhou C, Lee S,

Laptev N, Levendovszky T, Taylor J, Qian H, Zhang J, Shoydokova A, Singh T, Zhu C, Baz Z, Bergmeir C, Yu D, Koylan A, Jiang K, Temiyasathit P, Yurtbay E (2022) Kats. https://github.com/facebookresearch/Kats
Jones N, Fulcher B, Sethi S, Lubba C CompEngine. 2021. www.comp-engine.org Accessed 10 May 2023
Kim N, Park S, Lee J, Choi J (2018) Load profile extraction by mean-shift clustering with sample Pearson correlation coefficient distance. Energies 11(9):2397. https://doi.org/10.3390/en11092397
Kuipers L, Niederreiter H (2006) Uniform Distribution of Sequences, Unabr. republ. of orig. publ. by wiley, new york, 1974 edn. Dover books on mathematics. Dover, Minneola
Lindberg KB, Bakker SJ, Sartori I (2019) Modelling electric and heat load profiles of non-residential buildings for use in long-term aggregate load forecasts. Util Policy 58:63–88. https://doi.org/10.1016/j.jup.2019.03.004
Lubba CH, Sethi SS, Knaute P, Schultz SR, Fulcher BD, Jones NS (2019) catch22: canonical Time-series characteristics: selected through highly comparative time-series analysis. Data Mining Knowl Discov 33(6):1821–1852. https://doi.org/10.1007/s10618-019-00647-x
Marszal-Pomianowska A, Heiselberg P, Kalyanova Larsen O (2016) Household electricity demand profiles–a high-resolution load model to facilitate modelling of energy flexible buildings. Energy 103:487–501. https://doi.org/10.1016/j.energy.2016.02.159
McLoughlin F, Duffy A, Conlon M (2015) A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 141:190–199. https://doi.org/10.1016/j.apenergy.2014.12.039
Meier H, Fünfgeld C, Adam T, Schieferdecker B (1999) Repräsentative VDEW-Lastprofile. Technical report, VDEW Frankfurt (Main)
Meinecke S, Thurner L, Braun M (2020) Review of steady-state electric power distribution system datasets. Energies 13(18):4826. https://doi.org/10.3390/en13184826
Meinecke S, Sarajlić D, Drauz SR, Klettke A, Lauven L-P, Rehtanz C, Moser A, Braun M (2020) SimBench – A benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. Energies 13(12):3290. https://doi.org/10.3390/en13123290 . Accessed 10 Apr 2023
Mishra M, Bhardwaj CA, Desikan K (2017) A maximal heterogeneity based clustering approach for obtaining samples
Müllner D (2011) Modern hierarchical, agglomerative clustering algorithms
Murphy KP (2013) Machine Learning: a Probabilistic Perspective, 4. print. (fixed many typos) edn. Adaptive computation and machine learning series. MIT Press, Cambridge, Mass
O'Hara-Wild M (2023) Feasts: feature extraction and statistics for time series
Olatomiwa L, Mekhilef S, Ismail MS, Moghavvemi M (2016) Energy management strategies in hybrid renewable energy systems: a review. Renew Sustain Energy Rev 62:821–835. https://doi.org/10.1016/j.rser.2016.05.040
Owen AB (2017) A randomized Halton algorithm in R. arXiv:stat.CO. https://doi.org/10.48550/ARXIV.1706.02808 . Accessed 10 May 2023
Papenberg M (January 2023) k-plus anticlustering: an improved k-means criterion for maximizing between-group similarity. preprint, PsyArXiv. https://osf.io/7jw6v. Accessed 10 May 2023
Papenberg M, Klau GW (2021) Using anticlustering to partition data sets into equivalent parts. Psychol Methods 26(2):161–174. https://doi.org/10.1037/met0000301
Park JY, Yang X, Miller C, Arjunan P, Nagy Z (2019) Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Appl Energy 236:1280–1295. https://doi.org/10.1016/j.apenergy.2018.12.025
Pfenninger S, Staffell I (2016) Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. Energy 114:1251–1265. https://doi.org/10.1016/j.energy.2016.08.060
Pflugradt N, Stenzel P, Kotzur L, Stolten D (2022) LoadProfileGenerator: an agent-based behaviorsimulation for generating residential load profiles. J Open Source Softw 7(71):3574. https://doi.org/10.21105/joss.03574
Proedrou E (2021) A comprehensive review of residential electricity load profile models. IEEE Access 9:12114–12133. https://doi.org/10.1109/ACCESS.2021.3050074
Ravi R, Sundaram R, Marathe MV, Ravi SS, Rosenkrantz DJ (1994) Spanning trees short or small. arXiv:math.CO. https://doi.org/10.48550/ARXIV.MATH/9409222 . Accessed 10 May 2023
Sandhaas A, Kim H, Hartmann N (2022) Methodology for generating synthetic load profiles for different industry types. Energies 15(10):3683. https://doi.org/10.3390/en15103683
Sorensen AL, Westad MC, Delgado BM, Lindberg KB (2022) Stochastic load profile generator for residential EV charging. E3S Web Conf 362:03005. https://doi.org/10.1051/e3sconf/202236203005
Späth H (1986) Anticlustering: maximizing the variance criterion. Control Cybern 15(2):213–218
Staffell I, Pfenninger S (2016) Using bias-corrected reanalysis to simulate current and future wind power output. Energy 114:1224–1239. https://doi.org/10.1016/j.energy.2016.08.068
Staudt P, Ludwig N, Huber J, Hagenmeyer V, Weinhardt C (2018) SCiBER: a new public data set of municipal building consumption. In: Proceedings of the Ninth International Conference on Future Energy Systems, ACM, Karlsruhe Germany. pp. 618–621. https://doi.org/10.1145/3208903.3210281
Team SDC (2022) Metadata record for: Dataset on electrical single-family house and heat pump load profiles in Germany. figshare. https://springernature.figshare.com/articles/dataset/Metadata_record_for_Dataset_on_electrical_single-family_house_and_heat_pump_load_profiles_in_Germany/17206271 Accessed 10 Apr 2023
Tjaden T, Bergner J, Weniger J, Quaschning V (2015) Representative electrical load profiles of residential buildings in Germany with a temporal resolution of one second. Unpublished. https://doi.org/10.13140/RG.2.1.3713.1606/1
Valev V (1998) Set partition principles revisited. In: Goos G, Hartmanis J, Van Leeuwen J, Amin A, Dori D, Pudil P, Freeman H (eds.) Advances in Pattern Recognition vol. 1451, Springer, Berlin, Heidelberg. pp. 875–881. http://link.springer.com/10.1007/BFb0033314 Accessed 10 Nov 2023
Wang Z, Hong T (2020) Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN). Energy Build 224:110299. https://doi.org/10.1016/j.enbuild.2020.110299
Wang X, Zheng Y, Zhao Z, Wang J (2015) Bearing fault diagnosis based on statistical locally linear embedding. Sensors 15(7):16225–16247. https://doi.org/10.3390/s150716225

Widén J, Lundh M, Vassileva I, Dahlquist E, Ellegård K, Wäckelgård E (2009) Constructing load profiles for household electricity and hot water from time-use data-modelling approach and validation. Energy Build 41(7):753–768. https://doi.org/10.1016/j.enbuild.2009.02.013

Wilson E, Parker A, Fontanini A, Present E, Reyna J, Adhikari R, Bianchi C, CaraDonna C, Dahlhausen M, Kim J, LeBar A, Liu L, Praprost M, White P, Zhang L, DeWitt P, Merket N, Speake A, Hong T, Li H, Mims Frick N, Wang Z, Blair A, Horsey H, Roberts D, Trenbath K, Adekanye O, Bonnema E, El Kontar R, Gonzalez J, Horowitz S, Jones D, Muehleisen R, Platthotam S, Reynolds M., Robertson J, Sayers K, Li Q (2021) End-Use Load Profiles for the U.S. Building Stock. DOE Open Energy Data Initiative (OEDI); National Renewable Energy Laboratory (NREL). https://doi.org/10.25984/1876417

Witten IH, Frank E, Hall MA, Pal CJ (2017) Data mining: practical machine learning tools and techniques, Fourth edition edn. Elsevier, Morgan Kaufmann, Amsterdam Boston Heidelberg London New York Oxford Paris San Diego San Francisco Singapore Sydney Tokyo

Yang Y, Bremner S, Menictas C, Kay M (2018) Battery energy storage system size determination in renewable energy systems: a review. Renew Sustain Energy Rev 91:109–125. https://doi.org/10.1016/j.rser.2018.03.047

Publisher's Note