https://dx.doi.org/10.1016/j.enbuild.2021.111817

Distance measures in building informatics: An in-depth assessment through typical tasks in building energy management

Ao Li[1], Cheng Fan[1,2], Fu Xiao[1, 3*] and Zhijie Chen[1]

[1] Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China

[2] Department of Construction Management and Real Estate, Shenzhen University, Shenzhen, China

[3] Research Institute for Smart Energy, The Hong Kong Polytechnic University

Abstract

Distance measurement (also known as similarity measurement) is used to evaluate pairwise similarities between data samples. It has been widely used in diverse building informatics research and applications to classify or cluster massive building data with the aim of improving prediction accuracy, identifying operation patterns, benchmarking and diagnosing building performance, etc. Various distance measures have been adopted to measure the distance/similarity of building data. However, the intrinsic complexity and diversity of building operational data bring considerable difficulties to the selection of a suitable distance measure for a specific task. There is a strong and urgent need for a comprehensive review and systematic comparison of existing distance measures in building informatics.

This study provides a comprehensive review on various distance measures and their applications in building operational data analysis. A systematic comparison is undertaken based on two typical tasks relying on building informatics, i.e., building energy usage pattern recognition, and clustering-based weather data segmentation for the customized development of building energy

prediction models. Nine widely adopted distance measures have been reviewed and compared, including Euclidean distance, Chebyshev distance, Manhattan distance, Mahalanobis distance, Hausdorff distance, Pearson correlation distance, Dynamic Time Warping, Edit distance on Real Sequence, and Cosine distance. Novel internal and external clustering validation approaches based on the cross-test and prediction accuracy are proposed and adopted to compare the clustering performance. The results in case studies showed that weather data clustering using the Cosine distance and Pearson correlation distance helps to obtain better energy prediction results in terms of MAPE (13.22% and 12.91%, respectively) than the commonly-used Euclidean distance (13.99%). The results also revealed that better clustering performance does not necessarily lead to higher prediction accuracy. The research results and insights obtained are valuable to guide distance-based research in building informatics.

Keywords: Distance measure, clustering, pattern recognition, time-series analysis

1. Introduction

Building informatics is concerned with building information management throughout the whole building lifecycle, which requires interdisciplinary collaboration, including Building Information Modeling (BIM) [1], big data [2], Internet of Things (IoT) [3], and machine learning. Building-related data/information contain a wealth of knowledge about the interactions between building energy consumption and the factors that influence it. And building informatics play an increasing important role in improving building management and reducing energy consumption. Machine learning is one of the most rapidly growing data-driven technical fields, lying at the intersection of computer science and statistics, and serving the core of Artificial Intelligence (AI) and data science [4]. The application of data-intensive machine-learning methods in building informatics leads to more evidence-based decision-making [5].

Distance-based learning algorithms are predominant in the field of machine learning, such as the k-means algorithms for clustering analysis, the k-Nearest Neighbors for classification and regression. They are inspired by one of the most critical components of numerous human cognitive processes, i.e., the capacity to quantify similarities between different objects [6]. Distance-based algorithms and approaches have been widely used in all kinds of applications oriented for improving building energy performance, including building energy prediction (e.g., k-NN [7]), fault detection and diagnosis (e.g., [8,9]), energy usage pattern recognition (e.g., fuzzy clustering [10], k-shape clustering [11], k-means clustering [12]), occupant behavior identification (e.g., Hierarchical clustering [13], k-means clustering [1]), energy benchmarking (e.g., Bin method [14,15]), and building typology analysis for obtaining reference buildings [16,17]. For instance, distance-based approaches have been widely used in FDD applications, where the difference between actual and predicted values are used as indicators for fault detection and diagnosis [18]. Li et al. [9] proposed a distance-based data-driven strategy for chiller FDD. The fault was detected and diagnosed if data measurements fall into one of the predefined fault clusters and within the predefined Manhattan distance range. The dynamic behavior of buildings and their energy systems generally exhibits extremely different operating characteristics under varying environmental conditions. Hence, distance-based data segmentation has been widely used for the customized development of machine learning models [19,20], RC model development [21], and energy saving analysis [22]). Building energy prediction is essential for control optimization and energy management towards energy saving and carbon emission mitigation [23]. Paudel et al. [19] proposed two AI modeling approaches (i.e., “all data” approach and “relevant data” approach) for predicting building heating energy consumption. The “all data” approach used all available training data, while the “relevant data” approach only used a small representative dataset with

similar climate conditions (judged by dynamic time warping) to prediction day. The numerical results showed that the "relevant data" modeling approach had higher prediction accuracy ($R^2 = 0.98; RMSE = 3.4$) than the "all data" modeling approach ($R^2 = 0.93; RMSE = 7.1$). Additionally, the distance-based Bin method is also frequently used in building energy consumption benchmarking [14,15]. More specifically, the historical loads are grouped together into a bin if their associated variables (such as hour of week, temperature, and humidity) are similar and fall into the same interval categories. The mean value of the bin is then used to predict/estimate building loads with similar associated variables. However, most of the existing research focuses on development and comparison of different distance-based algorithms/methods, while ignoring the research on distance measure between data samples.

To measure the distance or similarity between data samples, it is necessary to introduce/define a distance measure. In the perspective of building informatics, there are many sorts of data, such as building physical parameters (e.g., floor area, U-value, window-to-wall ratio), building operational data (e.g., electrical energy consumption, operational data of HVAC system), climate data (e.g., outdoor temperature, solar radiation), user-related data (e.g., occupant activities, on-off appliances), and time variables (e.g., season, date). Building operational data can therefore be numerical or categorical, single- or multi-dimensional, homogeneous or inhomogeneous, sequential or cross-sectional in different scenarios. The complexity and diversity of building operational data bring considerable difficulties to the selection or definition of distance measures. Therefore, the distance measurement under varying/different scenarios needs to be scientifically defined/selected to accurately describe the pairwise similarities in building operational data. As an example, clustering analysis can be used to partition data into groups based on their internal similarities. Such methods have been frequently used to classify energy consumers [24], predict

future energy demand [19], occupant behavior identification [13] and detect distinguished, habitually undesirable behaviors [25]. The accurately quantification of pairwise data similarities serves as the fundamental basis for clustering algorithms, and a poorly-defined distance measure typically lead to meaningless and invalid clustering [10].

Although a multitude of (competitive and robust) distance measures (e.g., Euclidean distance, Mahalanobis distance, Dynamic Time Warping) have been proposed, choosing an appropriate distance is definitely not an easy task. Research efforts have been made to investigate the effects of various distance measures [22,26] on applications oriented for enhancing building energy efficiency. Iglesias and Kastner [22] investigated the influence of four similarity measures (i.e., Euclidean distance, Mahalanobis distance, Dynamic Time Warping distance, and distance based on Pearson's correlation) in the application of time series clustering in discovering typical building energy patterns. Results showed that Euclidean and DTW distances outperformed the other two methods. Euclidean distance was the measurement that obtained the best, balanced general solution. And DTW can be considered an improved alternative in applications that benefit from a better representation of the highly similar kernel (or parts of the samples). Al-Wakeel et al. [27] conducted k-means clustering analysis on power consumption estimation based on four kinds of distance measures (i.e., Canberra, Manhattan, Euclidean, and Pearson correlation distances). The simulation results revealed that the use of Canberra distance yields more accurate load estimates (around 7% MAPE) than other distance measures. However, the sorts of distance measures that have been examined in previous research are limited, and some of the research findings are in conflict with others [22,26,27].

It is vital but challenging to objectively and comprehensively analyze and compare various distance measures under different tasks. Yilmaz et al. [26] presented an analysis of clustering

approaches for grouping the electricity demand profiles for households in Switzerland. The clustering analysis was applied to average household electricity profiles and daily electricity profiles of 656 multi-family flats in Swizerland. Four methods of distance measurement (i.e., Euclidean, Manhattan, Canberra and Chebyshev distance) were compared. Results showed that, different methods achieved similar clustering performance in terms of Silhouette score (Euclidean: 0.186; Manhattan: 0.186; Canberra: 0.188; Chebyshev: 0.185). Leprince and Zeiler [28] proposed a clustering-based load pattern identification method and assessed the benefit of the attained information on enhancing accuracies of building energy prediction. The proposed method was tested using energy consumption data of 70 residential buildings located in Netherlands. A cross-clustering validation was illustrated on varying distance measures (i.e., Euclidean and Mahalanobis distances), algorithms (i.e., Fuzzy C-Means and Agglomerative Hierarchical clustering) and number of clusters (ranging from 2 to 30). Three internal clustering validity indexes were selected in validation analysis, namely Silhouette score, Calinski-Harabasz index, and Davies-Boudlin Index (DBI). The results showed that, Euclidean distance achieved better results compared to Mahalanobis distance with higher Calinski-Harabasz indexed and lower DBIs. Ding et al. [29] conducted an extensive set of time series experiments re-implementing 8 different representation methods and 9 similarity measures and their variants, and testing their effectiveness on 38 time-series data sets from a wide variety of application domains. One major drawback of previous research is that the validation/verification methods of distance-based algorithms are insufficiently systematic and comprehensive, such as adopting the default Euclidean distance measure in internal clustering validation indexes, and that the internal and external validation methods are not used together to evaluate the results.

Previous researches reveal a number of gaps and shortcomings, i.e., (1) the types of distance measures compared in individual work are limited (no more than four), (2) the application scenarios are unspecific or undefined, (3) some of the findings are inconsistent [22,26,27], and (4) the validation/verification methods are insufficiently systematic and comprehensive, which limits the generalization capability of the results [26,27,28]. To tackle the above-mentioned limitations, there is a strong need for comprehensive review and systematic comparison of existing distance measures in building data analytics. This study attempts to investigate and assess the nine widely-used distance measures by conducting comparison study on typical tasks in building energy management. The main contributions of this paper are summarized as follow:

- This study conducted a comprehensive literature review concerning applications of various distance measures in the building informatics field.
- A systematic comparison study is conducted on two application scenarios, i.e., Case I: building energy usage pattern recognition, and Case II: clustering-based weather data segmentation for prediction model development.
- Both internal and external clustering validation approaches are designed and adopted to evaluate the clustering results. The traditional internal clustering validation approach is modified based on the "cross-test" concept. A novel external clustering validation approach based on model prediction accuracy is designed for Case II.
- In-depth discussions are provided on the pros and cons of the distance measures, and how to select a suitable distance measure for a specific task based on the review and comparison study.

The remaining part of the paper is organized as follows. Chapter 2 presents an overview of various distance measures and their applications in building energy field. Chapter 3 presents the comparison schemes and data used for assessment. The research results are presented in Chapter 4. Chapter 5 provides discussions on the pros and cons of the distance measures and how to select suitable distance measure based on the review and comparison study. Chapter 6 concludes the paper and discusses possible future research extensions.

2. Overview of distance measures

This section introduces the theoretical background on representative distance measures. lists the equations of several widely-used distance measures, and Figure 1 depicts their diagram. And a comprehensive literature review on their applications in building energy field will be provided. Table 2 summarizes distance measures used in relevant studies, along with the (type of) data samples (red: time series data) and research applications. The majority (19 out of 23) of the reviewed studies measured the distance of time series data, e.g., energy consumption subsequences and outdoor climate subsequences. And the applications of frequent occurrence include pattern recognition, data segmentation (for energy prediction), and FDD.

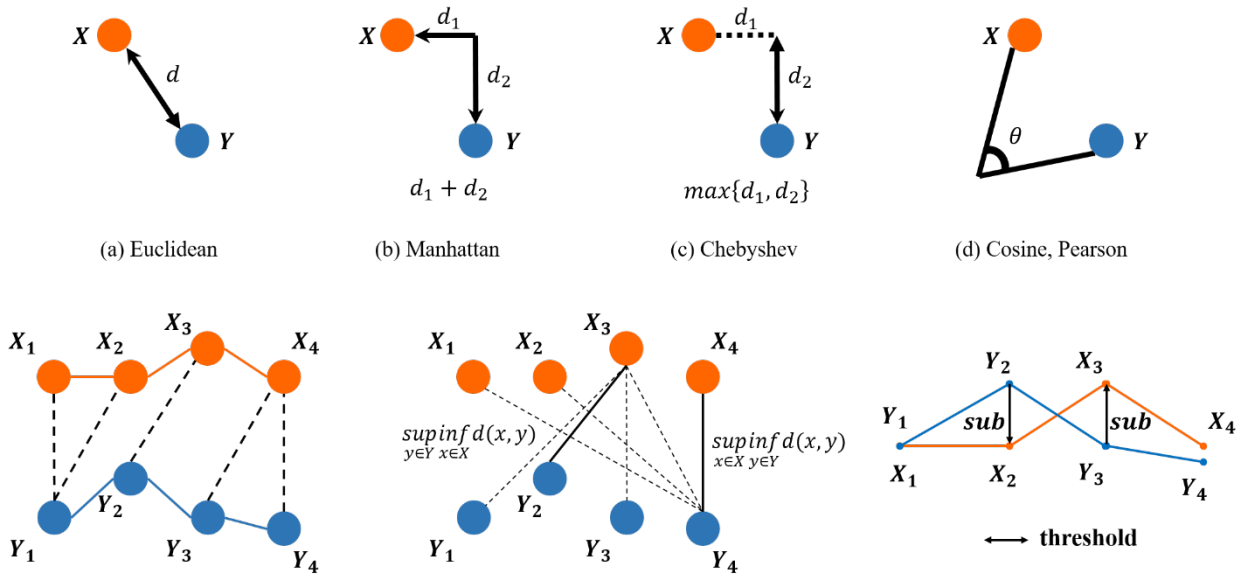


Figure 1. Diagram of several distance measures

Table 1. Summary of the nine distance measures discussed in this research

Distance measure	**Equation**		**Characteristics**
Euclidean distance	$d(X,Y) = \sqrt{\sum_{i=1}^{T}(X_i - Y_i)^2}$	(1)	Classical; Most widely-used; Geometric correspondence
Mahalanobis distance	$d(X,Y) = \sqrt{(X-Y)^T S^{-1}(X-Y)}$	(2)	Consider data correlation; Robust against data projection or rescaling
Hausdorff distance	$d(X,Y) = max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\}$	(3)	No data alignment restrictions; High computation time
Manhattan distance	$d(X,Y) = \sum_{i=1}^{T}\lvert X_i - Y_i\rvert$	(4)	Applicable for both categorical and numerical sequences; Geometric correspondence
Chebyshev distance	$d(X,Y) = \lim_{k\to\infty}\left(\sum_{i=1}^{T}\lvert X_i - Y_i\rvert^k\right)^{\frac{1}{k}} = \max_i(\lvert X_i - Y_i\rvert)$	(5)	Dominated by maximal component-wise difference
Pearson distance	$d(X,Y) = 1 - \frac{\sum_{i=1}^{T}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{T}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{T}(Y_i - \bar{Y})^2}}$	(6)	Scale-independent; Reflect linear correlation; Geometric correspondence (a function of the angle)
Cosine distance	$d(X,Y) = 1 - \frac{\sum_{i=1}^{T} X_i \times Y_i}{\sqrt{\sum_{i=1}^{T}(X_i)^2} \times \sqrt{\sum_{i=1}^{T}(Y_i)^2}}$	(7)	Scale-independent; Geometric correspondence (vector similarity)
EDR	$EDR[i,j] = \begin{cases} m & \text{if } i = 0 \\ n & \text{if } j = 0 \\ min\begin{cases} EDR[i-1][j-1] + subcost[i][j] \\ EDR[i][j-1] + 1 \\ EDR[i-1][j] + 1 \end{cases} & \text{otherwise} \end{cases}$	(8)	Applicable to categorical and numerical sequences; Robust against data imperfections
DTW	$DTW(X,Y) = \frac{Minimum\ accumalated\ warping\ cost}{Length\ of\ optimal\ warping\ path}$		Effective and popular for time series; High computation time

Table 2. Summary of the applications of distance measures in the building energy field

Distance	Data	Application	Ref
DTW	Outdoor temperature	Relevant data selection for energy consumption prediction	[19]
	BACS data	Application-agnostic	[30]
	Building energy consumption	Detect building energy usage patterns Improve accuracy of forecasting model	[11]
	Outdoor temperature	Relevant data selection for building energy demand prediction	[20]
	Energy signature profiles	Identify heating system and building type Building retrofit analysis	[31]
	Building energy system operation data	Building energy usage pattern recognition	[32]
	Transient power waveform	Non-intrusive load transient identification	[33]
Chebyshev	Historical (cooling load) prediction residuals	Quantify the uncertainties in building cooling load prediction	[34]
Hausdorff	Building electricity consumption	Determine natural segmentation of customers Identify temporal consumption patterns	[35]
Mahalanobis	Energy signature	Outlier detection	[31]
	Building load profile	Pattern identification	[28]
	Climate profile	Building climate zoning	[36]
	Building energy consumption	Pattern identification	[37]
	Real value and prediction error of energy demand	Ensemble 4 different (building energy demand) prediction models	[38]
Manhattan	Occupant's Preference and energy consumption profile	Light intensity setup Personalized control visual comfort	[39]
	Outdoor temperature	Data separation for building energy saving analysis	[22]
	Chiller operational data	FDD of building chiller faults	[9]
Edit distance	Belgian time-of-use data	Discover occupancy pattern	[13]
Cosine	All variables in the system	Initialize the input weights for Extreme learning machine for energy consumption robust prediction	[40]
	Solar power variations on consecutive days	Solar power prediction	[41]
Pearson	Energy consumption data	Detect customers with anomalous drops in their consumed energy	[42]
Euclidean, DTW, Mahalanobis, Pearson	Building energy consumption	Building energy pattern recognition	[10]

Euclidean, Chebyshev, Pearson	Building energy daily profile	Group daily electricity usage profiles Relevant data selection for forecasting model	[43]
Euclidean, Chebyshev, Pearson	Building energy daily profile	Identify informative typical daily electricity usage profiles	[44]

Euclidean distance, Manhattan distance and Chebyshev distance

The definitions of these three distances are all based on the Minkowski distance. Two data samples X and Y of length T are defined as following:

$$X = [X_1, X_2, \dots, X_T]$$

$$Y = [Y_1, Y_2, \dots, Y_T]$$

The Minkowski distance of order p (where p is an integer) between two samples X and Y is defined as following equation:

$$d(X,Y) = \left(\sum_{i=1}^{n} |X_i - Y_i|^p \right)^{\frac{1}{p}} \quad (9)$$

Minkowski distance measures the difference between two samples in the format of multi-dimensional vectors from an average perspective, and the order p represents the significance of individual component-wise difference (i.e., $|X_i - Y_i|$). With the increase of p, large component-wise difference will contribute more to the distance. When p equals 1 and 2, the Minkowski distance corresponds to the Manhattan distance (Eq. (4) in Table 1) and Euclidean distance (Eq. (1)) , respectively. When p approaches infinity, the Chebyshev distance can be obtained (Eq. (5)).

The Euclidean distance is the distance in the Euclidean space, which is the fundamental space of classical geometry. It is a distance measure that can be best interpreted as the length of a line segment connecting two points. Euclidean distance is definitely the most applied similarity

measure and usually appropriate for applications that do not present directly or necessarily correlation among distinct features [10]. Under normal circumstances, if not specifically specified, the default distance measure is Euclidean distance. Despite its widely utilization, Euclidean distance is not scale in-variant which means that distances computed might be skewed depending on the units of the features. And the distribution of each component (expected value, variance) may be different. Normally, one needs to normalize or standardize the data before using this distance measure [45].

The Manhattan distance, also called city block distance or taxicab distance, is a form of geometry in which the usual distance function or metric of Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. The reason for the naming of Manhattan distance is from the shortest driving path between cities planned as square building blocks (such as Manhattan). Kar et al. [39] proposed a recommender system-based approach for personalized visual comfort control in buildings. Initially, the individual user-preferences and energy consumption profiles were extracted from historical data. Then, the collaborative user-preferences are learnt/calculated based on the Manhattan distance of the target occupant from every other occupant. The proposed recommender system will generate the final recommended light intensity based on both individual and collaborative user-preferences, which is sent to the actuator for setting up the light intensity.

The Chebyshev distance is the indicator of maximal component-wise difference between two multi-dimensional vectors. Understanding this kind of correspondence/trend is particularly useful in many cases, e.g., loss function design (or model evaluation) (in machine learning model development) [46]. Haben et al. [46] proposed a new forecast error measure based on Minkowski distance for domestic household electrical energy prediction. In this study, 4-norm (i.e., $p = 4$)

was adopted, rather than the more common 2-norm (i.e., Mean squared error, the implementation of Euclidean distance), to penalize large errors (i.e., missed peaks) much more than small errors.

Mahalanobis distance

There are primarily two issues with the aforementioned distances (i.e., Euclidean distance, Manhattan distance, Chebyshev distance): 1) The dimension of each component is treated as the same; 2) The components' distributions (e.g., expectation, variance) may vary, but are not considered [47,48].

The Mahalanobis distance, as shown in Eq. (2), is defined as a distance measurement between two vectors X and Y of the same distribution with the covariance matrix S. Mahalanobis distance can be considered as an evolution of the Euclidean distance (if S is the identity matrix, then Mahalanobis distance is equal to the Euclidean distance) that takes into account data correlation. The covariance matrix S is adopted for weighting different features. Mahalanobis distance usually performs successfully with large data sets with reduced features, otherwise undesirable redundancies (brought by the covariance matrix) tend to distort the results [49]. When working with random information, the Mahalanobis distance exhibits stability against data projections or rescalings without degeneration [50]. This is particularly useful when outliers exist. Westermann et al. [31] adopted Mahalanobis distance in outlier filtering process of building energy signature data. The covariance-based method used a multivariate Gaussian distribution and classifies the points most distant from the center as outliers.

Pearson distance

Pearson correlation coefficient (PCC), also known as the product moment correlation coefficient, is widely used to reflect the linear correlation between two sets of data [51]. The PCC between two vectors X and Y is defined as Eq. (9). It is the covariance of two variables, divided by the

product of their standard deviations. Therefore, PCC is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

$$PCC(X,Y) = \frac{Covariance(X,Y)}{SD(X) \times SD(Y)} = \frac{\sum_{i=1}^{T}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{T}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{T}(Y_i - \bar{Y})^2}} \quad (10)$$

where SD denotes the standard deviation. The closer the absolute value of PCC is to 1, the stronger the correlation between the two vectors. PCC can be seen as a function of the angle between the two variable vectors [52]. PCC is perhaps the most broadly applied index in all of statistics [52], although not free of distortions or problems [10]. Pearson distance is developed/defined based on Pearson's correlation coefficient, as shown in Eq. (6) in Table 1.

Cosine similarity

The cosine similarity is widely used for analyzing vector similarity. Unlike the Euclidean distance, the cosine similarity pays more attention to the direction and the angle between two vectors. The cosine similarity between two vectors X and Y is defined as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^{T} X_i \times Y_i}{\sqrt{\sum_{i=1}^{T}(X_i)^2} \times \sqrt{\sum_{i=1}^{T}(Y_i)^2}} \quad (11)$$

where θ is the angle between X and Y. A small angle means the tested two vectors are expected to have high similarity. The value of Eq. (10) is limited between -1 and 1. Based on Eq. (10), the cosine distance can be defined as Eq. (7) in , ranging from [0,2].

Xu et al. [40] also proposed a modified cosine similarity measure for initializing the input weights of building energy consumption prediction model (i.e., extreme learning machine), defined in follows:

$$cos'(\theta) = \frac{\sum_{i=1}^{T}(X_i - \bar{X}) \times (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{T}(X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{T}(Y_i - \bar{Y})^2}} \quad (12)$$

where $\bar{X}$ and $\bar{Y}$ is the mean value of X and Y, respectively. Instead of Euclidean distance, which is highly sensitive to magnitudes, the modified cosine similarity coefficient is adopted to initialize the weights connecting the input neurons and the hidden neurons of extreme learning machine, for improving its generalization ability.

Dynamic time warping (DTW)

Dynamic time warping (DWT) is one of the most popular and field-tested similarity measures is called the "time warping" distance measure. For the distance measurement of two sequences, Minkowski distance requires strict alignment (calculating all the component-wise difference between $[X_i, Y_i]$, and somehow averaging them), while DTW relaxes this alignment restriction. The principles of the DTW algorithm can be summarized as follows: a local cost n-by-m matrix C which contains all pairs of corresponding distances between two sequences X (of length n) and Y (of length m), with the element $C_{ij} = d(X_i, Y_j)$ representing the distance between the two points X_i and Y_j,, where the mapping d is called local cost function (usually the Euclidean distance). Assuming that under the constraints of the boundary condition, monotonicity condition and step size condition, as shown in formula 13, any warping path (or mapping) between X and Y from $(1,1)$ to (i,j) is $p = (p_1, p_2, \dots, p_k, \dots, p_l)$, where the kth element $p_k = (i_k, j_k) \in \{1,2, \dots, i\} \times \{1,2, \dots, j\}$. The accumulated cost associated with the warping path p can then be defined as $D = \sum_{k=1}^{l} d(i_k, j_k)$ with $k \in \{1,2, \dots, l\}$ and $l \in [\max(i,j), i+j-1]$, where l represents the path length.

$$\begin{cases} p_1 = (1,1), p_l = (i,j) \\ i_1 \le i_2 \le \cdots \le i_l, j_1 \le j_2 \le \cdots \le j_l \\ p_{k+1} - p_k \in \{(1,1),(0,1),(1,0)\} \end{cases} \quad (13)$$

Then, the optimal warping path (or optimal matching path) between X and Y satisfying the above constraints, which minimizes the warping cost D, is searched using dynamic programming algorithm, and the corresponding minimum accumulated distance D is calculated. Accordingly, the DTW distance between X and Y can be defined as $DTW(X,Y) = D/l$. According to the above principle, DTW algorithm aligns the data points of the two time-series while finding the optimal matching path, such that the "distance" between them is minimized. The Euclidean distance between two sequences (of same length) can be seen as a special case of DTW, where the kth element of warping path P is constrained such the $p_k = (i,j)_k, i = j = k$.

Dau et al. [53] compared Euclidean distance and DTW for classification cases on the University of California Riverside (UCR) time series archive, which contained 128 datasets of different types (e.g., image, ECG, motion, audio data). The classification results strongly supported that well-constrained DTW is better than Euclidean distance for most datasets (in UCR archive). In building energy field, DTW is frequently used as distance measurement in clustering analysis [30,31]. Westermann et al. [31] conducted clustering analysis on building energy signatures extracted by SVR for building retrofit analysis. In clustering analysis, the similarity of two energy signatures is calculated/measured using both Euclidean distance and DTW. The results showed that DTW performed better than Euclidean distance with either k-means or Hierarchical clustering algorithm. The authors stated that the DTW-based clustering can preserve the shape of the energy signature profiles (i.e., the entries of the profile vector remain), and at the same time it was capable to find energy signatures with similar shape even if they are offset. Bode et al. [30] proposed two clustering schemes (i.e., raw-data-based and feature-based) and tested them on time series data

extracted from the E.ON Energy Research Center. The raw-data based clustering scheme adopted DTW for similarity measurement. In the test case, the DTW technique showed higher clustering accuracy than statistical features for the long term time frame. Some researchers employed DTW in data segmentation based on weather clustering for separate prediction model development. For example, Paudel et al. [20] also adopted DTW for training data selection. This research developed a SVM model for building energy demand prediction. For each day prediction condition, via DTW-based method, the training data of relevant days are selected based on climatic conditions and functioning profile of building. The advantage of this training design is that it leads to higher accuracy and better computational efficiency in comparison to those based on the whole training data. DTW is also capable of dealing with variable-length time series. Liu et al. [33] adopted DTW to measure the similarity between the variable-length raw transient power waveform (TPW) sample and template time-series, for load transient identification.

As DTW is computationally expensive, different methods are proposed to speed-up the DTW matching process [54,55]. On the other hand, Keogh and Pazzani [56] pointed out the potential problems of DTW that it can lead to unintuitive alignments, where a single point on one time series maps onto a large subsection of another time series ([31] in our field also pointed out this problem). Additionally, DTW may fail to find obvious and natural alignments in two time series caused by a single feature (i.e., peak, valley, inflection point, plateau, etc.). One of the causes is due to the great disparity in the lengths of the comparing series. Therefore, besides improving the performance of DTW, methods are also proposed to improve the accuracy of DTW [56,57]. Interested readers are encouraged to refer to a comprehensive review [58].

Edit distance on Real sequence (EDR)

For categorical sequences (e.g., strings), one common distance measure is Edit distance. The Edit distance of two sequences X and Y is defined as the minimum number of edits needed to transform X into Y, allowing insertions, deletions and substitutions. Each edit operation may have a different cost. Typically, using "Levenshtein distance" (a cost of one is applied for each edit operation), the Edit distance from sequence $X = [X_1, \dots, X_m]$ to $Y = [Y_1, \dots, Y_n]$ is given by d_{mn}, defined by the recurrence [59]:

$$d_{i0} = \sum_{k=1}^{i} w_{\text{del}}(a_k), \text{ for } 1 \le i \le m \tag{14}$$

$$d_{0j} = \sum_{k=1}^{j} w_{\text{ins}}(b_k), \text{ for } 1 \le j \le n \tag{15}$$

$$d_{ij} = \begin{cases} d_{i-1,j-1} & \text{for } a_i = b_j \\ min\begin{cases} d_{i-1,j} + w_{\text{del}}(a_i) \\ d_{i,j-1} + w_{\text{ins}}(b_j) \\ d_{i-1,j-1} + w_{\text{sub}}(a_i, b_j) \end{cases} & \text{for } a_i \ne b_j \text{ for } 1 \le i \le m, 1 \le j \le n \end{cases} \tag{16}$$

Based on Edit distance, Chen et al. [60] introduced a novel distance function, i.e., Edit Distance on Real sequence (EDR) which is robust against data imperfections (e.g., noise, shifts and scaling of data that commonly occur due to sensor failures, disturbance signals and different sampling rates). The EDR from sequence $X = [X_1, \dots, X_m]$ to $Y = [Y_1, \dots, Y_n]$ can also be calculated in recurrence:

$$EDR[i,j] = \begin{cases} m & \text{if } i = 0 \\ n & \text{if } j = 0 \\ min\begin{cases} EDR[i-1][j-1] + subcost[i][j] \\ EDR[i][j-1] + 1 \\ EDR[i-1][j] + 1 \end{cases} & \text{otherwise} \end{cases} \tag{8}$$

where $subcost[i][j]$ is a function depending on whether X_i equals Y_j (1 if equal, 0 if not). EDR can also be employed for numerical sequences. In this case, a threshold ε should be assigned for judging that two points/values are equal (if the absolute residual/difference lies within the range). Aerts et al. [13] proposed a clustering-based method for identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison. In this study, each element of the dataset represented one individual's daily occupancy sequence. Each sequence in turn was constructed from 144 characters, one for each 10-min time step, with a value that corresponded to one of three possible occupancy states (i.e., at home and awake, sleeping, or absent). To be able to treat the difference between all states equally, each element was handled as a string instead of an integer (categorical sequence instead of numerical sequence). Therefore, Edit distance was adopted in hierarchical clustering, and seven typical occupancy patterns were identified.

Hausdorff distance

The Hausdorff distance measures how far two subsets are from each other; two sets are close if every point of either set is close to some point of the other set. The Hausdorff distance between two subsets X and Y is calculated by computing the shortest distance between each feature X_1 in set X with respect to features in set Y, and then maintain the largest value. In other words, Hausdorff distance is the greatest of all distances from a point X_i in one set to the closest point Y_i in the other set. Formally, the Hausdorff distance between X and Y is defined as following Eq. (17):

$$d(X,Y) = \left\{ \sup_{x \in X} d(x,Y), \sup_{y \in Y} d(X,y) \right\} = max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y) \right\} \quad (17)$$

where sup represents the supremum, inf represents the infimum, and where $d(a, B) = \inf_{b \in B} d(a, b)$ quantifies the distance from a point a to the subset B. It should be mentioned that Hausdorff distance has rather long computation time.

Chelmis et al. [35] adopted clustering algorithms for determining natural segmentation of smart grid electricity customers (i.e., buildings) and identification of temporal consumption patterns. To avoid clustering individual daily consumption observations for a given building, which can in turn results in a building participating in numerous clusters, the Hausdorff distance is employed in the clustering algorithm to determine one point per building instead. And the results showed that the Hausdorff-based clustering algorithm was capable of identifying good clusters of similar buildings by operating on sets of observations and their respective distances rather than considering individual points.

3. Description of comparison schemes and data used for assessment

This research aims to assess different distance measures by conducting a systematic comparison study of two typical application tasks based on building informatics: 1) Building energy usage pattern recognition (Case I); 2) Clustering-based weather data segmentation for the customized development of building energy prediction models (Case II). Figure 2 depicts the overall research outline. Data preprocessing is first carried out to enhance the data quality by filling in missing values, removing outliers, and preparing required data attributes for further analysis. The building energy consumption data and outdoor climate data are transformed into daily subsequences with an interval of one hour. Afterwards, clustering analysis (i.e., k-means clustering) is adopted based on different distance measures. Overall, nine widely-used methods of distance measurement are tested in this study, including Euclidean distance, Chebyshev distance, Mahalanobis distance, Hausdorff distance, Manhattan distance, Pearson correlation distance, Dynamic Time Warping,

Edit distance on Real Sequence, and Cosine similarity. The detailed clustering validation methods for two cases will be introduced in the following sections.

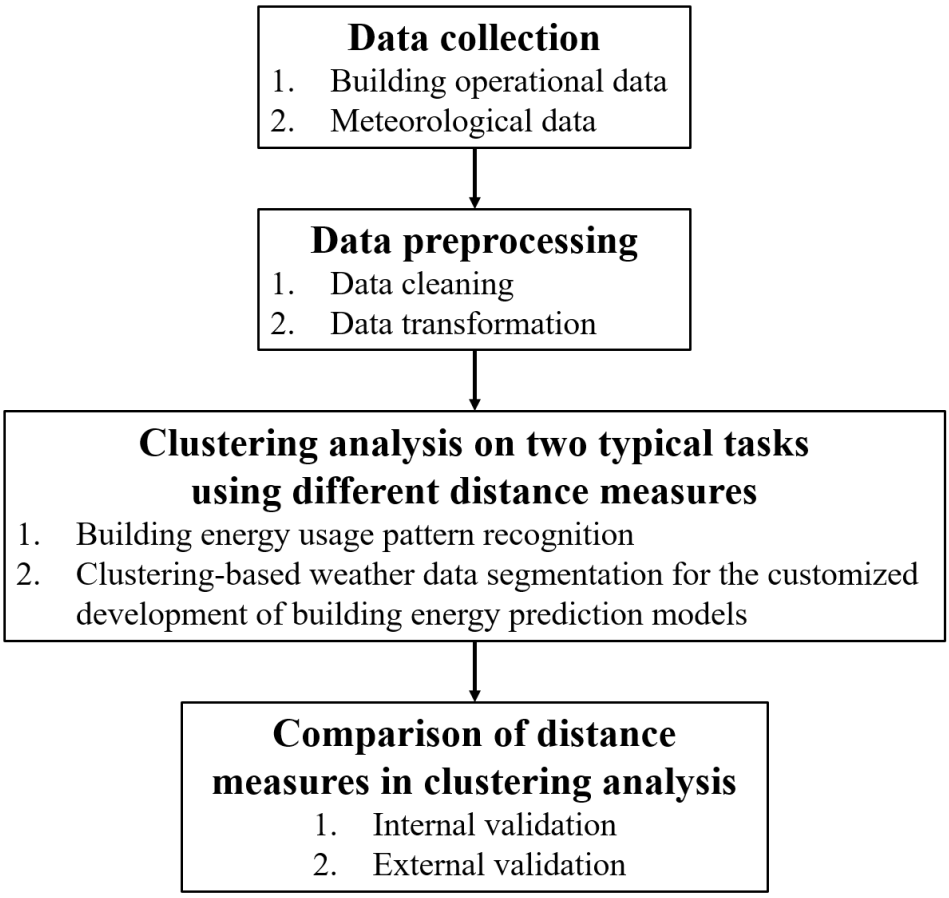


Figure 2. Outline for comparison study for assessing distance measures

3.1 Comparison scheme of distance measures in Case I

Figure 3 shows the intra-cluster similarity (between samples in the same cluster) and inter-cluster similarity (between samples in different clusters). The objective of clustering algorithms is to maximum the intra-cluster similarity and minimum the inter-cluster similarity.

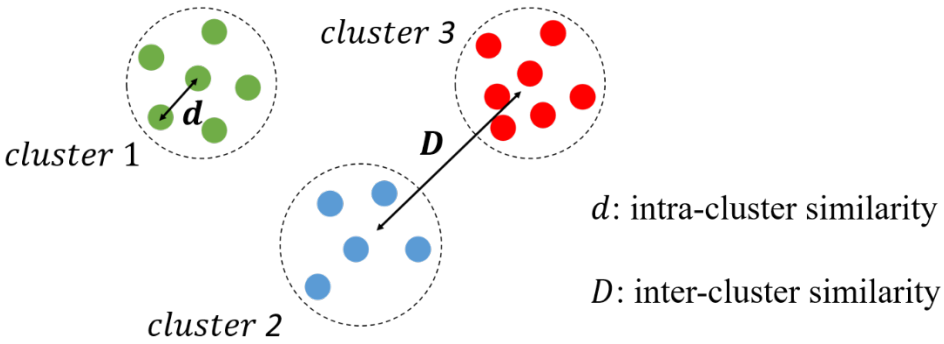


Figure 3. Intra-cluster and Inter-cluster similarity

As an unsupervised learning task, it is necessary to find a way to validate the goodness of partitions after clustering. Otherwise, it would be difficult to make use of different clustering results [46]. Clustering validation, which evaluates the goodness of clustering results, has long been acknowledged as one of the vital issues essential to the success of clustering applications [46]. The clustering validation methods (or methods of checking clustering solutions) can be broadly classified into two categories: internal clustering validation (or clustering validation methods), and external clustering validation (or clustering evaluation methods) [10]. And the performance metrics used in clustering validation are therefore named as internal indexes and external indexes. The primary distinction (between internal and external clustering validation methods) is whether or not external information is used for clustering validation. For internal clustering validation, the (clustering) results are evaluated through mathematical analysis and direct observation of solutions based on the inherent characteristics owned by the input data set. In a sense, it consists of idealistic analytic methods as they are concerned with the definition assigned to a cluster regardless of the reason for its deployment (i.e., the eventual application) [10]. As the goal of clustering is to make objects/samples within the same cluster similar and objects in different clusters distinct, internal validation measures are often based on two criteria [61,62]: *Compactness* and *Separation*. *Compactness* measures how closely related the objects in a cluster are, while *Separation* measures

how distinct or well-separated a cluster is from other clusters. On the other hand, the external clustering validation is a practical (or engineering) approach that focuses on application-based tests/assessments. The clustering solutions can be benchmarked and checked directly by the (clustering) application (or an environment that simulates the application). However, generalizations are more dangerous and riskier in this case, seeing that corruption and deformations may be introduced by the application, the boundary conditions and the specific data used for testing [10].

Two (widely-used) internal indexes (i.e., Dunn's index, and S_Dbw index) are selected in this study for clustering results validation. Dunn's index (DI) uses minimum pairwise distance between objects in different clusters as the inter-cluster between objects in different clusters as the inter-cluster separation and the maximum diameter among all clusters as the intra-cluster compactness [63]. S_Dbw index measures inter-cluster separation based on density and intra-cluster compactness based on variances of cluster objects [64]. Liu et al. [65] investigated 11 internal clustering validation indexes in five different aspects (i.e., monotonicity, noise, density, subclusters and skewed distributions). The S_Dbw index was the only measure performed well in all five aspects (therefore selected in this research).

For all internal clustering validation methods, a distance measure should also be assigned (in most case, the default is Euclidean distance). This will not cause a problem when comparing different clustering algorithms or determining the optimal cluster number. However, when evaluating the clustering results based on different distance measures, a paradox appears. Theoretically, when an internal index based on a particular distance measure is used, the clustering algorithm based on the same distance measure should produce the best results, which is entirely reasonable and easily-expected. Therefore, if only the default distance measure (typically Euclidean distance) is used in

internal validation indexes, then Euclidean distance and other distance measures similar to it should perform better. This issue was neglected by several previous studies [1,28], and will diminish the research's influence and contribution. To overcome this issue, Iglesias and Kastner [10] proposed a cross-test cluster validation procedure. That is, all distance measures tested in clustering analysis will be used in internal validation indexes/algorithms. In this research, this "cross-test" concept is adopted in Case I to modify the traditional internal clustering validation approach.

3.2 Comparison scheme of distance measures in Case II

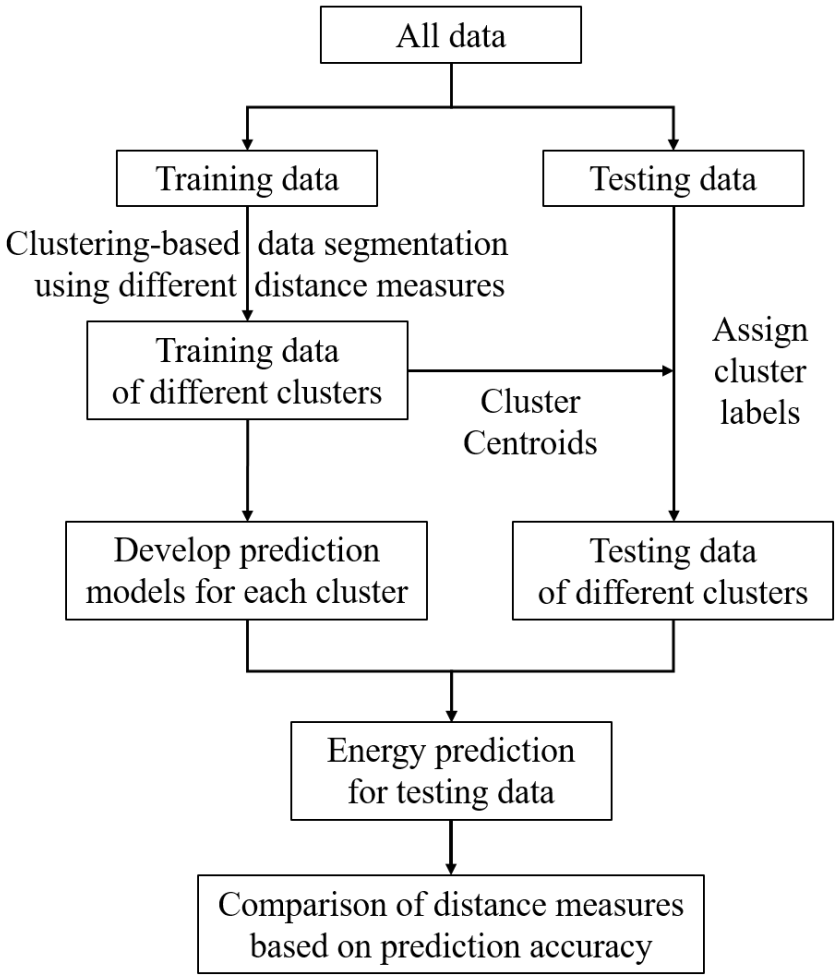
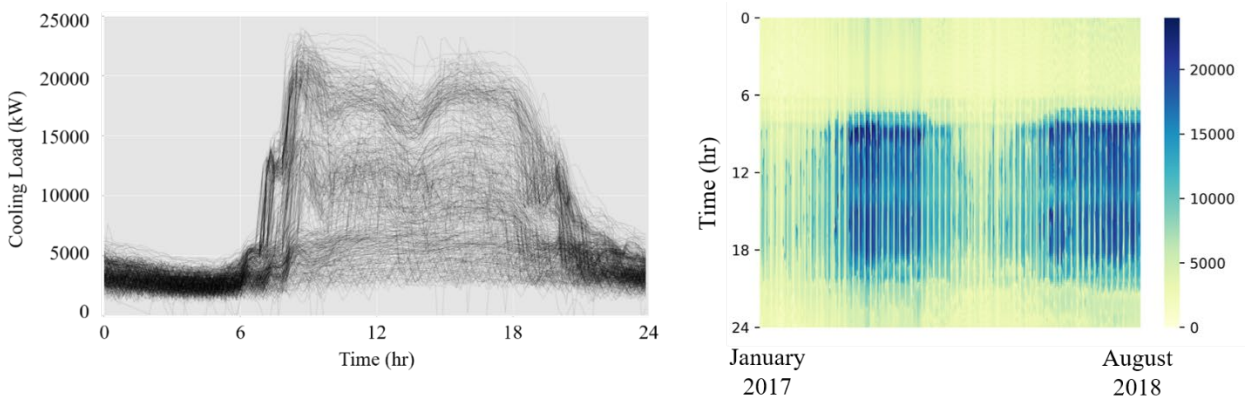


Figure 4. Comparison scheme of distance measures based on external clustering validation in Case II

The comparison scheme of different distance measures in Case II include both internal and external validation methods. The internal validation method uses the same two indexes, i.e., Dunn's index and S_Dbw index, as mentioned in Section 3.1. Additionally, an external clustering validation procedure based on model prediction accuracy is specially designed for Case II, as shown in Figure 4. The whole dataset is first separated into training data and testing data randomly. Clustering analysis (based on different distance measures) on weather data will be conducted to segment training data into different clusters. The centroids of all clusters are calculated and used to assign cluster labels for testing data. Afterwards, energy prediction models are trained by training data of each cluster. And the prediction accuracy of testing data will be used as the external clustering validation index which indicates the goodness of weather clustering-based data segmentation.

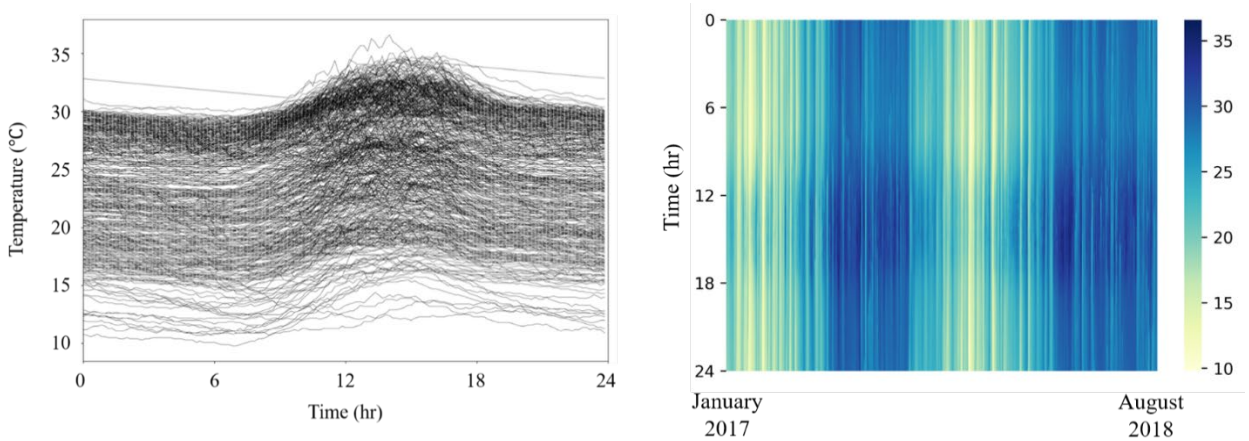
3.3 Data retrieved from the BAS of a real building

The data for this research's case studies were retrieved from the building automation system (BAS) of the tallest building in Hong Kong, the International Commerce Centre (ICC). This building is about 490 m high with a total floor area of approximately 321,000 m^2, consisting of a four-story basement, a six-story block building and a 98-story tower building. The building is served by a central chilling system consisting of six identical high-voltage centrifugal chillers that supply chilled water for air handling units. Each chiller has a rated cooling capacity of 7230 kW and a power consumption of 1270 kW. For analysis, a total of 463 days (from January 2017 to August 2018) of building operational data were retrieved. Figure 5 depicts the daily cooling load profile of the ICC during this period. The time interval of data collection is ten minutes. The climate data in the same period, including outdoor dry-bulb temperature (as shown in Figure 6), relative humidity, and solar radiation, were obtained from the Hong Kong Observatory.



(a) Line graph (b) Heat map

Figure 5. Daily cooling load profile of ICC



(a) Line graph (b) Heat map

Figure 6. Daily outdoor dry-bulb temperature profile of Hong Kong

As the data quality of BAS data is usually low due to measurement noise, sensor faults, transmission problems, and other factors. A data preprocessing procedure is used in this research to enhance data quality. The missing values are filled in using moving average method, while the outliers are identified with domain expertise and statistical criterion. Afterwards, min-max normalization is adopted to transform the data into a suitable scale for further analysis. Finally, the whole dataset is split into daily subsequences with hourly time interval. That is, the data samples for following clustering analysis are 24-d vectors.

4. Assessment Results

4.1 Case I: Building energy usage pattern recognition

In this study, to avoid unintuitive alignments, a window size constraint of 5 along the main diagonal on the envelope of the warping path ($|i_k - j_k| \leq 5, for\ k \in \{1,2,\dots,l\}$) is set in DTW application [66]. And a threshold $\varepsilon = 0.2$ is set in EDR algorithm. In Case I, the optimal cluster number is determined as 3, based on the value of internal clustering indexes under different cluster numbers in preliminary analysis.

Table 3 and Table 4 summarize the cross-test results of internal clustering indexes, i.e., S_Dbw index and DI, respectively. "C" in the header (i.e., the first row) represents the distance measure used in clustering analysis, while "V" in the first column represents the distance measure used in clustering validation. And the number (from 1 to 10) in the first line and column represent different distance measure, i.e., Euclidean, Cosine, DTW, DTW (window size=5), Mahalanobis, Pearson, Hausdorff, EDR, Chebyshev, and Manhattan distance, respectively. For example, the value of (C5, V1) in Table 3 is 1.087, which means that the Euclidean distance-based (V1) S_Dbw index equals to 1.087 for Mahalanobis distance-based (C5) clustering results. Smaller S_Dbw index, and larger DI indicate better clustering results. For example, it can be seen from the values of V1 Row (from (C1, V1) to (C10, V1)) in Table 3 that, Euclidean distance, EDR and Manhattan distance perform better than other distance measures when using Euclidean distance-based S_Dbw index in internal clustering validation. To make the results easier to comprehend, min-max normalization is conducted for each row. And the heat maps of normalized results for S_Dbw and DI are presented in Figure 7 and Figure 8, respectively. In each heat map, darker colors indicate better clustering results. As shown in the Figures, Euclidean (C1), EDR (C8) and Manhattan (C10) distances all

compete for the best distance measure for clustering, whereas Cosine (C2), Mahalanobis (C5), and Pearson (C6) distance perform relatively worse.

Table 3. Cross-test results measured by S_Dbw index

S_Dbw index: the smaller, the better, C: clustering distance, V: validation distance

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
C1	0.81	0.863	0.739	0.808	1.073	0.883	0.792	1.514	0.792	0.781
C2	1.163	0.994	1.306	1.306	1.646	1.03	1.168	1.979	1.168	1.186
C3	1.005	0.917	1.004	1.004	1.092	0.931	1.03	1.206	1.03	1.005
C4	0.969	0.882	0.987	0.987	1.076	0.878	0.977	1.273	0.977	0.948
C5	1.087	1.31	1.108	1.414	2.96	1.598	2.398	2.12	0.922	0.916
C6	1.448	1.26	1.465	1.465	2.134	1.321	1.614	2.484	1.614	1.465
C7	0.927	0.878	0.897	0.897	1.106	0.914	0.908	1.405	0.908	0.931
C8	0.802	0.876	0.772	0.76	1.092	0.897	0.77	1.368	0.754	0.789
C9	0.926	0.883	0.897	0.897	1.106	0.914	0.872	1.27	0.872	0.931
C10	0.811	0.84	0.808	0.808	1.07	0.884	0.792	1.423	0.792	0.808

(1: Euclidean; 2: Cosine; 3: DTW; 4: DTW (window size: 5); 5: Mahalanobis; 6: Pearson; 7: Hausdorff; 8: EDR; 9: Chebyshev; 10: Manhattan distance)

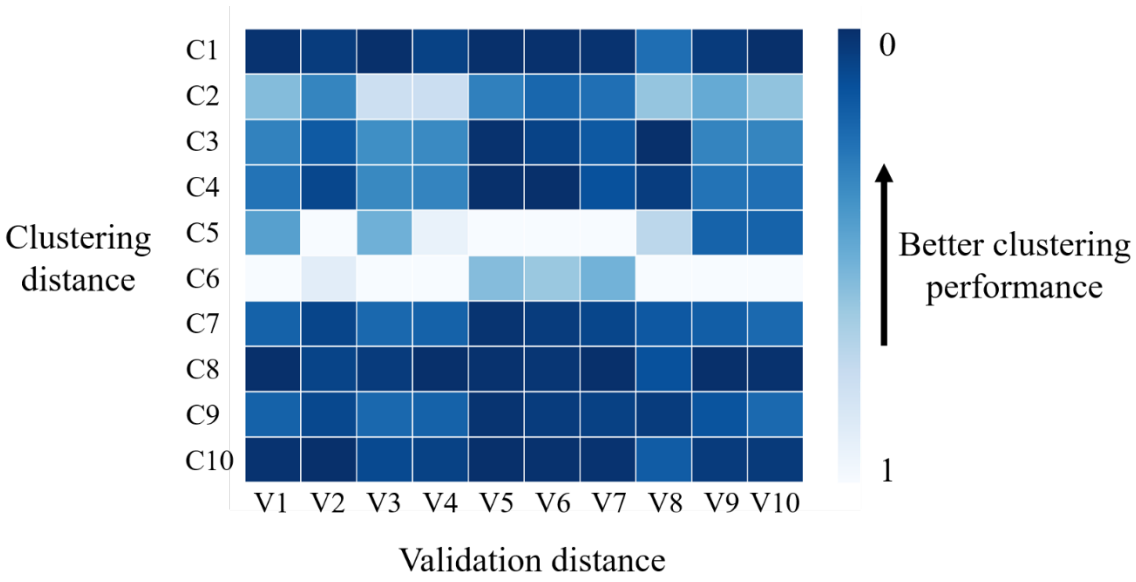


Figure 7. Heat map of cross-test results measured by normalized S_Dbw index

(The darker, the better clustering performance)

Table 4. Cross-test results measured by Dunn's Index

Dunn's Index: the bigger, the better, C: clustering distance, V: validation distance

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
C1	0.093	0.009	0.092	0.072	0.161	0.007	0.085	0.091	0.085	0.085
C2	0.037	0.019	0.038	0.031	0.142	0.017	0.047	0.077	0.047	0.036
C3	0.024	0.004	0.041	0.035	0.147	0.006	0.032	0.091	0.032	0.043
C4	0.058	0.008	0.089	0.07	0.142	0.006	0.07	0.091	0.07	0.044
C5	0.029	0.006	0.032	0.021	0.122	0.004	0.028	0.067	0.066	0.049
C6	0.015	0.013	0.012	0.01	0.102	0.016	0.022	0.067	0.022	0.014
C7	0.076	0.008	0.083	0.073	0.167	0.006	0.05	0.091	0.05	0.061
C8	0.121	0.01	0.091	0.074	0.187	0.007	0.065	0.083	0.085	0.081
C9	0.053	0.008	0.083	0.073	0.167	0.006	0.089	0.091	0.089	0.061
C10	0.098	0.009	0.095	0.074	0.166	0.007	0.092	0.091	0.092	0.079

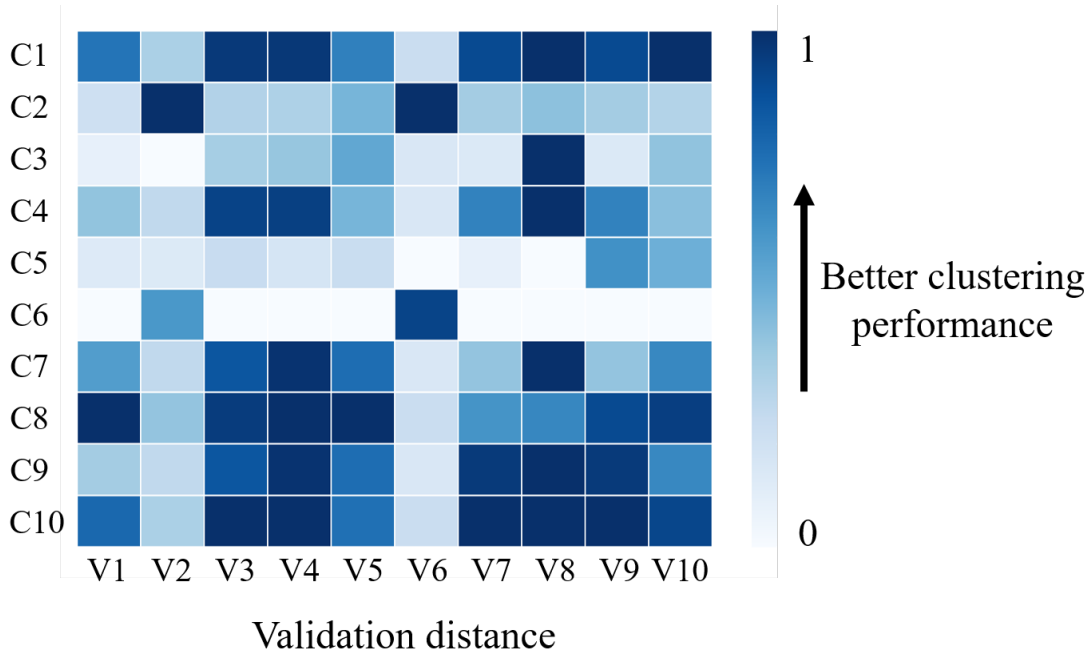


Figure 8. Heat map of cross-test results measured by normalized Dunn's Index

(The darker, the better clustering performance)

A closer inspection of the cross-test results reveals several groups/sets of distance measures that are analogous. As stated in Chapter 3.2, when an internal index based on a particular distance measure is utilized, the clustering algorithm based on the same distance measure should theoretically yield the best results, e.g., (C2, V2) in Table 4. By examining the definitions of various distance measures, it can be found that some of them are calculated in an analogous way, which is also reflected in the cross-test. For example, Euclidean distance and Manhattan distance are two different forms of Minkowski distance, with an order of 2 and 1, respectively. As illustrated in Table 3 and Table 4., there are slight differences between (C1, Vi) and (C10, Vi), for i ranging from 1 to 10. Similar phenomenon can also be observed for Cosine distance (C2) and Pearson (C6) distance. It can be explained by the fact that Cosine similarity is simply the cosine of the angle between two vectors, while Pearson correlation coefficient can be interpreted as a function of the angle between the two variable vectors [52].

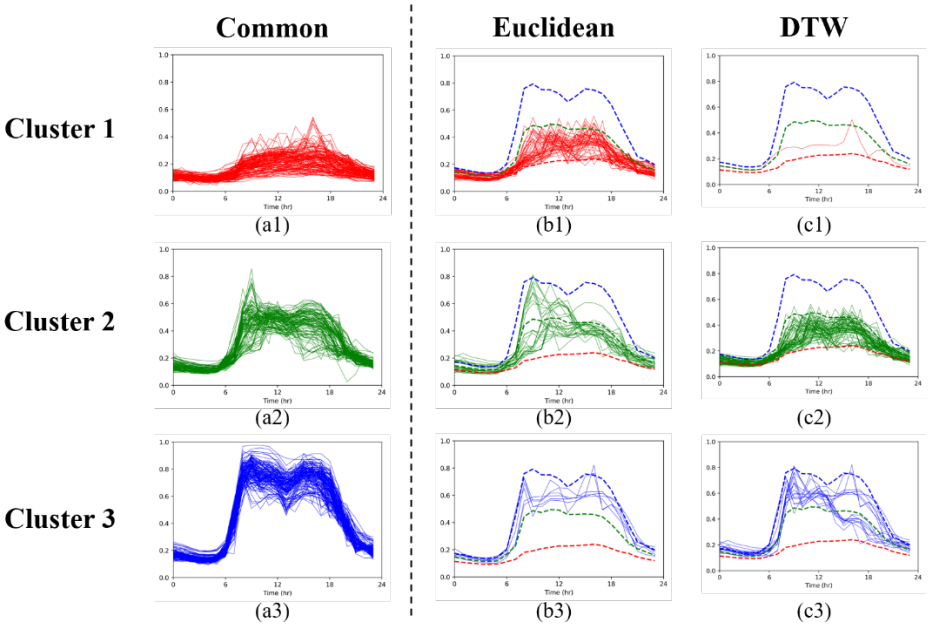


Figure 9. Comparison of clustering results using Euclidean distance and DTW (the curves of daily building energy consumption profile, 463 days in total)

To better explain the comparison, the clustering results using the Euclidean distance (C1) and DTW (C4) are illustrated in Figure 9. Figure 9 (a1), (a2) and (a3) show the common data samples (i.e., building energy consumption daily profiles) in the three clusters, which are assigned by the two distance measures to the same cluster. The rest data samples excluding the common samples in the corresponding cluster are depicted in Figure 9 (b1, b2, b3 clustered using the Euclidean distance) and (c1, c2, c3 clustered using the DTW). More specifically, Cluster 1 obtained by using the Euclidean distance consists of data samples in a1 and b1, while Cluster 1 obtained by using the DTW distance consists of data sample in a1 and c1. The two clusters have similar centroids and therefore both are named as Cluster 1. The similar way of expression is applied to Cluster 2 and Cluster 3 in Figure 9. The three dashed curves in the small figures denote the centroids of common data samples in each cluster, i.e., red, green, and blue lines for Cluster 1, 2, and 3, respectively. As can be observed, Figure 9 (b1) and (c2) essentially show almost the same batch of data samples. For this batch of daily energy consumption profiles, Euclidean distance assigns them into Cluster 1 (closer to the centroid of Cluster 1), while DTW assigns them into Cluster 2 (closer to the centroid of Cluster 2). This result clearly and intuitively demonstrates the different emphasizes of the two distance measures, i.e., DTW focuses more on the "shape" difference between two samples, e.g., the daily energy consumption profiles in this study. These distinctions in clustering results may have a significant impact on subsequent applications [10]. For example, clustering of key performance parameters is important in FDD applications. There are time lags in many processes (e.g., thermal response time in heat conduction process due to thermal inertia). The impact of certain key parameters may not be reflected immediately in the performance parameters. Under these circumstances, FDD methods based on the Euclidean distance may trigger false alarm, while the methods based on DTW will be more tolerant with these time lags. Generally speaking, it is

rather difficult to decide which of these distance measures provides the better/best result. Researchers should make decisions relying on the understanding of various of distance measures and the application scenarios.

4.2 Case II: Clustering-based weather data segmentation for the customized development of building energy prediction models

In Case II, 100 days are randomly selected from the whole dataset (479 days) as the testing data, with the remainder serving as the training data (379 days). Clustering analysis using different distance measures are conducted on the outdoor air temperatures, which divides the training dataset into three clusters. Two widely-used machine learning algorithms, i.e., Support Vector Machine (SVM) and Multiple Linear Regression (MLR) are adopted to develop building cooling load prediction models for each cluster. SVM can efficiently perform regression or classification tasks using kernel trick (i.e., mapping the inputs into high-dimensional feature spaces) and have been widely-used in building energy prediction. To prevent introducing excessive uncertainty, the input variables of prediction models are selected as the cooling load of previous 6 hours and the outdoor temperature. The time window is selected as 6 based on a preliminary results of autocorrelation analysis, as illustrated in Figure 10.

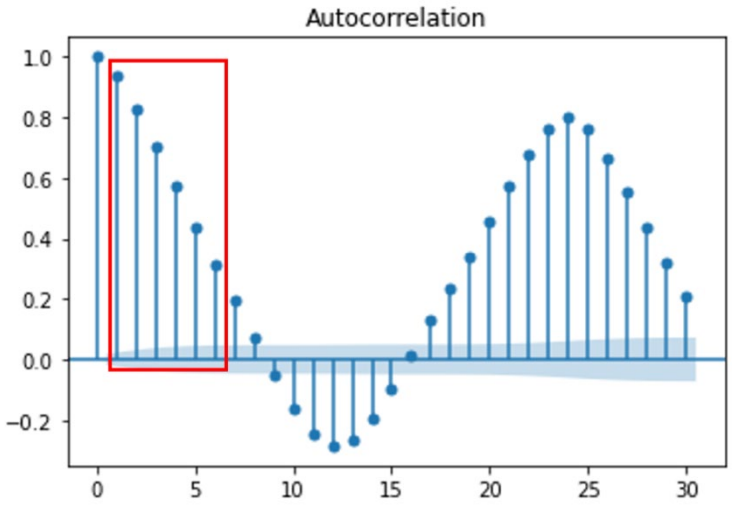


Figure 10. The autocorrelation analysis of cooling load time sequence

The performance indexes used in Case II include the mean squared error (MSE), and the mean absolute percentage error (MAPE). They are calculated based on Eq. (18) and (19), respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{18}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{\hat{y}_i}\right| \tag{19}$$

where y_i is the actual energy consumption, $\hat{y}_i$ is the predicted energy consumption, and N is the number of observations. MSE is scale-dependent, while MAPE is scale-independent which allows us to express the error of the model in a percentage.

Table 5. The internal and external clustering validation results of Case II (cluster number: 3)

Validation	**Metric/Index**		**Euclidean**	**Cosine**	**DTW**	**DTW(w=5)**	**Mahalanobis**
Internal	S_Dbw		0.376	1.472	0.715	0.376	1.189
	DI		0.06	0.014	0.036	0.06	0.013
External	SVM	MSE	0.0043	0.0038	0.0042	0.0042	0.0043
		MAPE	13.99%	13.22%	13.89%	13.90%	14.06%
	MLR	MSE	0.0028	0.0027	0.0028	0.0027	0.0025
		MAPE	11.10%	10.99%	11.20%	11.00%	10.05%

Validation	Metric/Index		Pearson	Hausdorff	EDR	Chebyshev	Manhattan
Internal	S_Dbw		2.35	0.72	0.729	0.717	0.376
	DI		0.016	0.05	0.021	0.05	0.06
External	SVM	MSE	0.0037	0.0042	0.0043	0.0042	0.0043
		MAPE	12.91%	13.95%	13.99%	13.95%	13.99%
	MLR	MSE	0.0024	0.0028	0.0028	0.0028	0.0028
		MAPE	10.25%	11.24%	11.16%	11.18%	11.1%

Table 5 summarizes both the internal and external clustering validation results of Case II. For internal validation, the distance used to calculate the internal indexes is Euclidean distance. In general, the prediction accuracy of the same modeling method (i.e., SVR or MLR) but adopting different distance measures varies within a small range. However, it is interesting to note that, whereas Cosine and Pearson distance perform poorly in internal clustering validation, they achieve slightly better prediction performance on the testing data.

To better understand how the clustering results, influence the prediction results, the clustering results using the Euclidean distance and Cosine distance are illustrated in Figure 11 and Figure 12, respectively. As the clustering results of Pearson distance are similar to the Cosine distance, this paper selected the results of the Cosine distance as an example to illustrate and analyze the findings.) The upper parts of Figure 11 and Figure 12 represent the outdoor temperature profiles in each cluster, while the lower parts represent the normalized building energy consumption profiles corresponding to each cluster. Figure 11 shows that the clustering using the Euclidean distance segments daily outdoor air temperature profiles into three clusters with very different mean values, which are roughly corresponding to cool, mild and hot seasons in Hong Kong. The widely-adopted intuitive calendar-based data segmentation methods (e.g., dividing the whole dataset into subsets according to seasons) cluster the data in the similar manner as the Euclidean-

based clustering method and hence can get similar clustering results. The lower average daily outdoor temperature corresponds to lower average daily building energy consumption, which well agrees with the domain knowledge. However, as illustrated in Figure 12, the outdoor temperature profiles have larger variations in each cluster when the data are segmented based on the Cosine distance, while the mean values of the outdoor air temperature in each cluster don't exhibit obvious difference. The same observation can be obtained for the cooling load profiles in each cluster It is interesting to see that the models developed using sub-dataset with larger variations perform even better than those developed using sub-datasets with smaller variations.

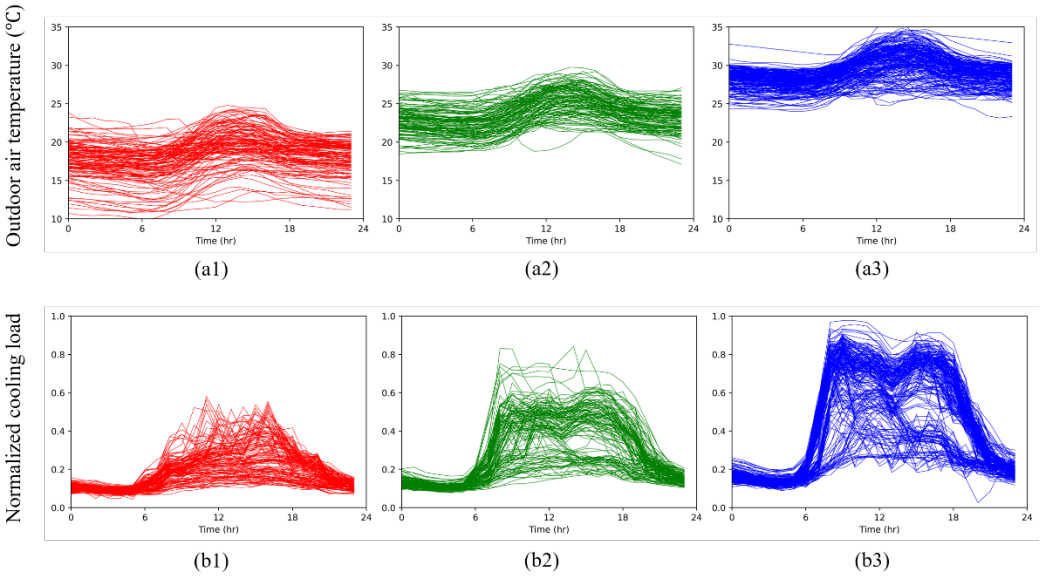


Figure 11. (a) The clustering results of outdoor temperature daily profiles using Euclidean distance (three clusters); (b) Building cooling load profiles corresponding to each cluster

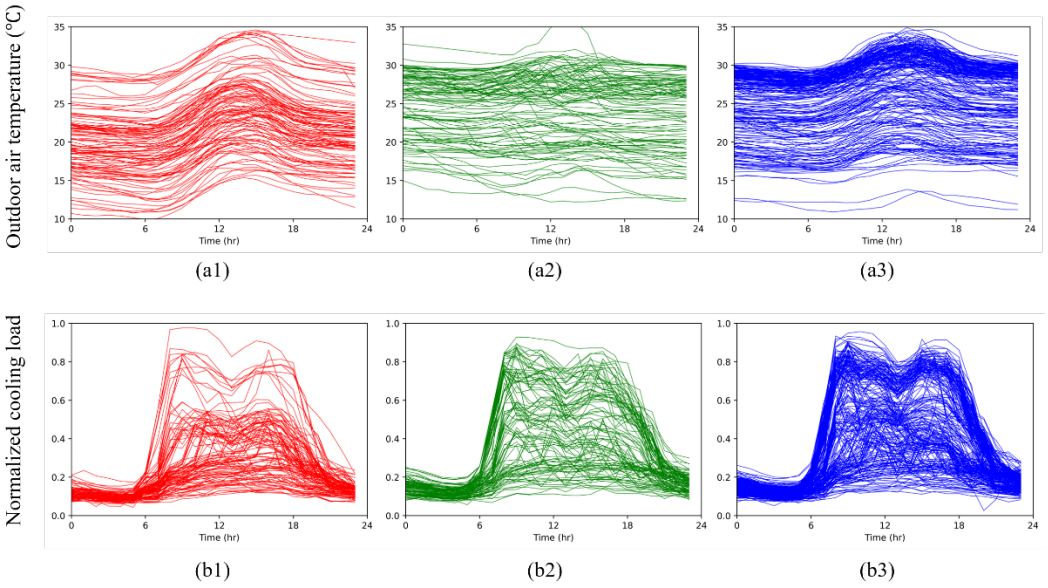


Figure 12. (a) The outdoor temperature daily profiles obtained using Cosine distance-based clustering (three clusters); (b) Building cooling load profiles corresponding to each cluster

5. Discussions

Based on the above review and comparison study, the pros and cons of different distance measures, and how to choose distance measures for a specific task are discussed in this section.

The distance measures are different in terms of mathematical functions, physical meanings, focuses, robustness, computation time and applicable conditions. Some distance measures have strong physical meanings, e.g., Euclidean, Cosine and Pearson distance, which is beneficial to matching them with different tasks Some distance measures are specially designed for measuring the distance of time series/sequences, e.g., DTW and EDR. The Hausdorff distance can also be used to measure the difference between two subsets regardless of the sequence of data samples in the datasets, which can be used for data/dimension reduction and transfer learning-related applications [67]. Some distance measures, e.g., DTW, EDR and Hausdorff distance, loosen the alignment requirements on the data pairs to some extent, which could improve loss function design [68] to overcome the “double penalty” effect of traditional error metric (e.g., the widely adopted

mean squared error, which is equivalent to the Euclidean distance). The Mahalanobis distance addresses the problem caused by different distributions of each feature/element of the data samples, using the covariance matrix. Some distance measures are more sensitive to outliers, e.g., the Pearson distance and Mahalanobis distance. They may not be suitable for the data exploratory tasks (e.g., pattern recognition in Case I), but perform better in FDD applications [31,42]. Using the Euclidean distance or an appropriate distance measure for a special data type (e.g., using DTW for time series) could be a safe decision when there is no strong evidence to support the selection of a distance measure.

It is possible that individual distance measure does not well fit particular data or applications. In this case, a new complex distance measure, which combine several distance measures in a complementary way [43,44], may be defined. Distance metric learning is the other promising method to define new distance measures [6], which is a branch of machine learning that aims to automatically learn/construct task-specific distances from the data with labels. It learns a distance measure that can put data samples with the same label together while push away samples with different labels. The learned distance can then be used to perform various tasks (e.g., k-NN classification, clustering, information retrieval). Reviews on distance metric learning [2,6] are recommended for readers interested in it.

There are two major findings from the case studies, which indicates that it is worthy of reflections on previous understandings and selection of distance measures. First, distance measures perform better according to internal validation don't necessarily lead to better performance in the application tasks (e.g., predictive modeling). Most previous studies adopting clustering analysis concerning different distance measures only compared the measures using internal validation. The distance measure with the best clustering performance was usually chosen to segment a large

dataset to several small sub-datasets for the subsequent tasks, such as predictive modeling and FDD. This study shows that this may not be the optimal way to choose a suitable distance measure as the ultimate purpose is the application tasks rather than segmenting data. Second, Case II shows that, in the same predictive modeling task using the same dataset, the models developed using sub-dataset with larger variations perform even better than those developed using sub-dataset with smaller variations. It is a common understanding in building informatics that data segmentation, as a data-preprocessing step of data-driven modeling, can improve modeling accuracy because the relationships in a sub-dataset are closer which is beneficial to modeling/learning. Clustering based on the Euclidean distance may ensure that data variances in a sub-dataset are smaller or the scope of operation condition covered by the sub-dataset is smaller, but not necessarily lead to the closer relationships in the sub-dataset. It is worthwhile to take a close look at the clustering results and evaluate the distance-based methods till the last step of a specific task.

Limited by space, this research assesses the distance measures by conducting s comparison study on only two typical tasks in the building energy field. The pros and cons of the distance measures may not be adequately revealed. It should be stated that, although the conclusions and insights obtained are interesting and inspiring, they are not complete. Hopefully, this study will enlighten more comprehensive and in-depth studies on distance measures as they are critically important to the rapidly growing R&D in building informatics for smart and energy-efficient buildings.

6. Conclusion

Distance-based algorithms, especially clustering algorithms, have been widely used in all kinds of building informatics applications oriented for improving building energy performance, including energy consumption prediction, fault detection and diagnosis, energy usage pattern recognition, and building energy consumption benchmarking. However, the complexity and diversity of

distance measurement objects and application scenarios bring considerable difficulties to the distance selection or definition, which significantly influences distance-based information retrieval, classification, clustering and other subsequent data mining procedures (in building energy related applications).

This study reviews nine typical distance measures in building informatics. A systematic comparison study for assessing the distance measures is conducted on two typical tasks, i.e., building energy usage pattern recognition (Case I), and clustering-based weather data segmentation for customized energy prediction model development (Case II). In total, nine widely-used distance measures are investigated. The comparison of distance measures adopts both internal and external clustering validation approaches. The traditional internal clustering validation approach is improved based on "cross-test" concept, and an external validation approach based on prediction accuracy is specially designed for Case II. The research results indicate that the Euclidean distance, Manhattan distance and EDR perform better in building energy usage pattern recognition, while Cosine and Pearson distances work better in clustering data for building cooling load prediction. The research results and insights obtained can be utilized to guide future distance-based research in building informatics perspective, improve building data management and building energy performance.

Acknowledgement

The authors gratefully acknowledge the support of this research by National Key Research and Development Program of China (2021YFE0107400) and the Research Grants Council of the Hong Kong SAR (152133/19E).

References

1. Xu, J., Kang, X., Chen, Z., Yan, D., Guo, S., Jin, Y., ... & Jia, R. (2021, February). Clustering-based probability distribution model for monthly residential building electricity consumption analysis. In Building Simulation (Vol. 14, No. 1, pp. 149-164). Tsinghua University Press.
2. Kulis, B. (2012). Metric learning: A survey. Foundations and trends in machine learning, 5(4), 287-364.
3. Volk, R., Stengel, J., & Schultmann, F. (2014). Building Information Modeling (BIM) for existing buildings—Literature review and future needs. Automation in construction, 38, 109-127.
4. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
5. Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., & Livingood, W. (2021). A review of machine learning in building load prediction. Applied Energy, 285, 116452.
6. Suárez, J. L., García, S., & Herrera, F. (2021). A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges.Neurocomputing, 425, 300-322.
7. Wahid, F., & Kim, D. (2016). A prediction approach for demand analysis of energy consumption using k-nearest neighbor in residential buildings. International Journal of Smart Home, 10(2), 97-108.
8. Tran, D. A. T., Chen, Y., Chau, M. Q., & Ning, B. (2015). A robust online fault detection and diagnosis strategy of centrifugal chiller systems for building energy efficiency.Energy and Buildings, 108, 441-453.
9. Li, D., Hu, G., & Spanos, C. J. (2016). A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis. Energy and Buildings, 128, 519-529.
10. Iglesias, F., & Kastner, W. (2013). Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies, 6(2), 579-597.
11. Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., ... & Tham, K. W. (2017). k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy and Buildings, 146, 27-37.
12. Sala, J., Li, R., & Christensen, M. H. (2021, February). Clustering and classification of energy meter data: A comparison analysis of data from individual homes and the aggregated data from

multiple homes. In Building Simulation (Vol. 14, No. 1, pp. 103-117). Tsinghua University Press.

13. Aerts, D., Minnen, J., Glorieux, I., Wouters, I., & Descamps, F. (2014). A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison. Building and environment, 75, 67-78.
14. Katipamula, S., Pratt, R. G., Chassin, D. P., Taylor, Z. T., Gowri, K., & Brambley, M. R. (1999). Automated fault detection and diagnostics for outdoor-air ventilation systems and economizers: Methodology and results from field testing.Transactions-American Society Of Heating Refrigerating And Air Conditioning Engineers, 105, 555-567.
15. Santos, J., Brightbill, L., & Lister, L. (2000, May). Automated diagnostics from DDC data–PACRAT. In Proceedings of the 8th National.
16. Schaefer, A., & Ghisi, E. (2016). Method for obtaining reference buildings. Energy and buildings, 128, 660-672.
17. Bre, F., Silva, A. S., Ghisi, E., & Fachinotti, V. D. (2016). Residential building design optimisation using sensitivity analysis and genetic algorithm. Energy and Buildings, 133, 853-866.
18. Zhao, Y., Li, T., Fan, C., Lu, J., Zhang, X., Zhang, C., & Chen, S. (2019). A proactive fault detection and diagnosis method for variable-air-volume terminals in building air conditioning systems. Energy and Buildings, 183, 527-537.
19. Paudel, S., Elmitri, M., Couturier, S., Nguyen, P. H., Kamphuis, R., Lacarrière, B., & Le Corre, O. (2017). A relevant data selection method for energy consumption prediction of low energy building based on support vector machine. Energy and Buildings, 138, 240-256.
20. Paudel, S., Nguyen, P. H., Kling, W. L., Elmitri, M., Lacarrière, B., & Corre, O. L. (2015). Support vector machine in prediction of building energy demand using pseudo dynamic approach. arXiv preprint arXiv:1507.05019.
21. Berthou, T., Stabat, P., Salvazet, R., & Marchio, D. (2014). Development and validation of a gray box model to predict thermal behavior of occupied office buildings. Energy and Buildings, 74, 91-100.
22. Ashouri, M., Haghighat, F., Fung, B. C., Lazrak, A., & Yoshino, H. (2018). Development of building energy saving advisory: A data mining approach. Energy and Buildings, 172, 139-151.

23. Ma, M., Ma, X., Cai, W., & Cai, W. (2019). Carbon-dioxide mitigation in the residential building sector: a household scale-based assessment. Energy Conversion and Management, 198, 111915.
24. Sütterlin, B., Brunner, T. A., & Siegrist, M. (2011). Who puts the most energy into energy conservation? A segmentation of energy consumers based on energy-related behavioral characteristics. Energy Policy, 39(12), 8137-8152.
25. Capozzoli, A., Lauro, F., & Khan, I. (2015). Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Systems with Applications, 42(9), 4324-4338.
26. Yilmaz, S., Chambers, J., Cozza, S., & Patel, M. K. (2019, November). Exploratory study on clustering methods to identify electricity use patterns in building sector. In Journal of Physics: Conference Series (Vol. 1343, No. 1, p. 012044). IOP Publishing.
27. Al-Wakeel, A., Wu, J., & Jenkins, N. (2017). K-means based load estimation of domestic smart meter measurements. Applied energy, 194, 333-342.
28. Leprince, J., & Zeiler, W. (2020, September). A Robust Building Energy Pattern Mining Method and its Application to Demand Forecasting. In 2020 International Conference on Smart Energy Systems and Technologies (SEST) (pp. 1-6). IEEE.
29. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 1(2), 1542-1552.
30. Bode, G., Schreiber, T., Baranski, M., & Müller, D. (2019). A time series clustering approach for building automation and control systems. Applied energy, 238, 1337-1345.
31. Westermann, P., Deb, C., Schlueter, A., & Evins, R. (2020). Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data.Applied Energy, 264, 114715.
32. Habib, U., Hayat, K., & Zucker, G. (2016). Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering. Complex Adaptive Systems Modeling, 4(1), 1-20.
33. Liu, B., Luan, W., & Yu, Y. (2017). Dynamic time warping based non-intrusive load transient identification. Applied energy, 195, 634-645.

34. Zhang, C., Zhao, Y., Fan, C., Li, T., Zhang, X., & Li, J. (2020). A generic prediction interval estimation method for quantifying the uncertainties in ultra-short-term building cooling load prediction. Applied Thermal Engineering, 173, 115261.
35. Chelmis, C., Kolte, J., & Prasanna, V. K. (2015, October). Big data analytics for demand response: Clustering over space and time. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 2223-2232). IEEE.
36. Yang, L., Lyu, K., Li, H., & Liu, Y. (2020). Building climate zoning in China using supervised classification-based machine learning. Building and Environment, 171, 106663.
37. Chang, C., Zhu, N., Yang, K., & Yang, F. (2018). Data and analytics for heating energy consumption of residential buildings: The case of a severe cold climate region of China.Energy and Buildings, 172, 104-115.
38. Rahman, S., Alam, M. G. R., & Rahman, M. M. (2019, December). Deep Learning based Ensemble Method for Household Energy Demand Forecasting of Smart Home. In2019 22nd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.
39. Kar, P., Shareef, A., Kumar, A., Harn, K. T., Kalluri, B., & Panda, S. K. (2019). ReViCEE: A recommendation based approach for personalized control, visual comfort & energy efficiency in buildings. Building and Environment, 152, 135-144.
40. Xu, Y., Zhang, M., Ye, L., Zhu, Q., Geng, Z., He, Y. L., & Han, Y. (2018). A novel prediction intervals method integrating an error & self-feedback extreme learning machine with particle swarm optimization for energy consumption robust prediction.Energy, 164, 137-146.
41. Jahromi, K. G., Gharavian, D., & Mahdiani, H. (2020). A novel method for day-ahead solar power prediction based on hidden Markov model and cosine similarity. Soft Computing, 24(7), 4991-5004.
42. Monedero, I., Biscarri, F., León, C., Guerrero, J. I., Biscarri, J., & Millán, R. (2012). Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. International Journal of Electrical Power & Energy Systems, 34(1), 90-98.
43. Li, K., Ma, Z., Robinson, D., Lin, W., & Li, Z. (2020). A data-driven strategy to forecast next-day electricity usage and peak electricity demand of a building portfolio using cluster analysis, Cubist regression models and Particle Swarm Optimization. Journal of Cleaner Production, 273, 123115.

44. Li, K., Yang, R. J., Robinson, D., Ma, J., & Ma, Z. (2019). An agglomerative hierarchical clustering-based strategy using Shared Nearest Neighbours and multiple dissimilarity measures to identify typical daily electricity usage profiles of university library buildings. Energy, 174, 735-748.
45. Fan, C., Yan, D., Xiao, F., Li, A., An, J., & Kang, X. (2020, October). Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches. In Building Simulation (pp. 1-22). Tsinghua University Press.
46. Haben, S., Ward, J., Greetham, D. V., Singleton, C., & Grindrod, P. (2014). A new error measure for forecasts of household-level, high resolution electrical energy consumption. International Journal of Forecasting, 30(2), 246-256.
47. Xiong, J., Yao, R., Grimmond, S., Zhang, Q., & Li, B. (2019). A hierarchical climatic zoning method for energy efficient building design applied in the region with diverse climate characteristics. Energy and Buildings, 186, 355-367.
48. Mahony, C. R., Cannon, A. J., Wang, T., & Aitken, S. N. (2017). A closer look at novel climates: new methods and insights at continental to landscape scales. Global Change Biology, 23(9), 3934-3955.
49. De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. Chemometrics and intelligent laboratory systems, 50(1), 1-18.
50. de la Hermosa González, R. R. (2018). Wind farm monitoring using Mahalanobis distance and fuzzy clustering. Renewable energy, 123, 526-540.
51. Pearson, K. (1920). Notes on the history of correlation. Biometrika, 13(1), 25-45.
52. Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1), 59-66.
53. Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C. C. M., Zhu, Y., Gharghabi, S., ... & Keogh, E. (2019). The UCR time series archive. IEEE/CAA Journal of Automatica Sinica, 6(6), 1293-1305.
54. Sakurai, Y., Yoshikawa, M., & Faloutsos, C. (2005, June). FTW: fast similarity search under the time warping distance. In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 326-337).

55. Yi, B. K., Jagadish, H. V., & Faloutsos, C. (1998, February). Efficient retrieval of similar time sequences under time warping. In Proceedings 14th International Conference on Data Engineering (pp. 201-208). IEEE.
56. Keogh, E. J., & Pazzani, M. J. (2001, April). Derivative dynamic time warping. In Proceedings of the 2001 SIAM international conference on data mining (pp. 1-11). Society for Industrial and Applied Mathematics.
57. Ratanamahatana, C. A., & Keogh, E. (2004, April). Making time-series classification more accurate using learned constraints. In Proceedings of the 2004 SIAM international conference on data mining (pp. 11-22). Society for Industrial and Applied Mathematics.
58. Fu, T. C. (2011). A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1), 164-181.
59. Jurafsky, D. (2000). Speech & language processing. Pearson Education India.
60. Chen, L., Özsu, M. T., & Oria, V. (2005, June). Robust and fast similarity search for moving object trajectories. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (pp. 491-502).
61. Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India.
62. Zhao, Y., & Karypis, G. (2002, November). Evaluation of hierarchical clustering algorithms for document datasets. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 515-524).
63. Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of cybernetics, 4(1), 95-104.
64. Halkidi, M., & Vazirgiannis, M. (2001, November). Clustering validity assessment: Finding the optimal partitioning of a data set. In Proceedings 2001 IEEE international conference on data mining (pp. 187-194). IEEE.
65. Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010, December). Understanding of internal clustering validation measures. In 2010 IEEE international conference on data mining (pp. 911-916). IEEE.
66. Jia, M., Komeily, A., Wang, Y., & Srinivasan, R. S. (2019). Adopting Internet of Things for the development of smart buildings: A review of enabling technologies and applications. Automation in Construction, 101, 111-126.Sakoe, H., & Chiba, S. (1978). Dynamic

programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1), 43-49.

67. Kumar, S., Shukla, A. K., Muhuri, P. K., & Lohani, Q. D. (2016, July). Atanassov Intuitionistic Fuzzy Domain Adaptation to contain negative transfer learning. In 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 2295-2301). IEEE.
68. Guen, V. L., & Thome, N. (2019). Shape and time distortion loss for training deep time series forecasting models. arXiv preprint arXiv:1909.09020.