A HYBRID SPACE FILLING DESIGN METHOD OF BUILDING PERFORMANCE DATABASE CONSTRUCTION FOR OFFICE BUILDING ENERGY PREDICTION

ABSTRACT

Building performance database (BPD) including building energy factors and the corresponding energy use data is an important research basis of building energy prediction for building design optimization and operation performance. Given the lack of a general building energy survey database, most of researches have chosen building simulation tool to obtain a targeted database as a basis for building energy prediction model development. Under the restriction of calculation time with building energy simulation tools, they can only focus on part of building factors to decrease the computational cost as much as possible. Moreover, current sampling and experiment design methods have limited sampling proportion. The efficiency and the dimensionality of these methods are not enough for the whole design space of building energy consumption with dozens of variables including weather conditions, building envelope, occupant behavior, HVAC systems, etc.

Given that, this study proposes a hybrid space filling design method combining the high-dimensional clustering method with the existing statistical sampling method to design the variables and cases for a BPD of office buildings. With this method, we can build a BPD only including about 10,000 cases but having the capability of representing the high-dimensional space constructed by 16 building energy variables at 3-6 levels. Relative to common statistical sampling method, the proposed method has higher sampling efficiency, and can help researches having many targeted variables to decrease the calculation cost and the following data mining complexity.

Based on the proposed cases design, we use jEPlus to conduct batch case calculation. The variables of massive cases and the corresponding outputs (building energy consumption) constituted the target BPD, that will contribute to effective building performance benchmark and assessment. In addition, the BPD can be used as the data basis of model development for office building energy prediction.

KEYWORDS

Building performance database, Case design, Space filling design, Sampling method, Parallel simulation

INTRODUCTION

Building energy prediction is the important basis of building energy efficiency design, energy performance optimization and energy retrofit evaluation, in which a variety of building energy prediction models are widely used in recent years. Nowadays the methods of building energy prediction mainly include forward simulation and data-driven models. In the complex application of the two methods, valid training database and suitable variables are essential to accurately predict building energy consumption. Strictly speaking, the best data comes from actual measurements, such as the Commercial Building Energy Consumption Survey (CBECS) (Lee SH et al. 2015). However, the available actual data is very limited. This situation conflicts with the high requirements of modelling data for existed methods, especially the data-driven method. Given this, many researches take the advantages of forward simulation tool to establish some targeted building performance databases (BPD) through massive calculations (Lee SH et al. 2015, Amiri SS et al. 2015).

A BPD is commonly comprised of large-scale building cases, and each of them is identified with multiple building variables and the simulated/measured building energy consumption. As the important basis of building energy assessment and prediction, a useful BPD needs to cover as many possibilities as possible to ensure the trained model performance. At present, there are two ways to construct a BPD to reflect the complex mapping relation between a variety of building factors and building energy consumption (Lee SH et al. 2015, Ivan K. 2013). The first way utilizes supercomputers to directly complete magnanimous calculation. Without consideration of efficient case design, this type of research commonly applied the exhaustive method to implement an ergodic process among the range of targeted building factors, which caused extremely large amounts of cases to simulate. The second way applies some experimental design or sampling methods to reduce the required number of cases examining the whole design space, in which all the possible combinations are distributed evenly within the test range. The commonly used sampling methods for BPD establishment include Orthogonal Experiment Design (Mao J et al. 2016), Monte Carlo method (Amiri SS et al. 2015, Kim Y et al. 2014.), and Latin Hypercube Sampling (Asadi E et al. 2014). This way selected certain combinations of building factors to represent the full factorial space so that the corresponding computational expense is acceptable. Obviously, the second way is more efficient and practical. While, it needs reasonable case design to ensure the representational capability of the built BPD.

Owing to the limited efficiency of the direct sampling using experimental design methods, existing BPDs in the building energy prediction researches mostly focused on partial building factors, such as a certain region (Yang L et al. 2015), a certain HVAC system (Neto AH et al. 2008) or a specific building. If not, the calculation cost is too high to implement easily. In this case, the research basis of predictive models varies for different case. The corresponding prediction models and conclusions are circumscribed and have little reference value to similar researches.

Given the above situation, a more reasonable case design process of the BPD construction is crucial for its general application in the building energy prediction and performance optimization field. Taking into account both efficiency and comprehensiveness of a BPD, this study focuses on the full-scale office building factors, and proposes a hybrid space filling design method, combing high-dimensional space metrics with common experimental design, to achieve the case design plan of an office BPD. Based on this plan, a parallel building energy simulation tool is used to generate and simulate almost 10,000 cases in batches. Within high-dimensional mixed building variables and the corresponding multiple time-scale building energy consumption, this database can basically reflect the complex mapping relation between building factors and building energy consumption. It can be utilized as a general data basis for building energy prediction, optimization design and benchmark evaluation of office buildings.

METHODS

The construction of the general BPD mainly includes three steps, primary variables design, case design, and the parallel simulation process, which are detailed below. Particularly, we proposed a hybrid space filling design method based on high-dimensional space metrics for the more reasonable and effective case design.

Primary variables design

To guarantee sufficient representativeness of the target BPD, the variables design needs to cover as many building factors as possible. After eliminating some factors with less flexibility or high co-correlation from overall factors of office buildings, we primarily summarized 16 building energy variables related to weather parameters, building shape, envelope, internal load, HVAC system and operation schedule as the primary variables of the BPD. For the convenience of the following case design, we divided them into two groups, numerical and non-numerical variables, see *Table 1*. The range of each variable is determined by ASHRAE or domestic codes and actual building situation.

Table 1. *Primary variables design for BPD*

Numerical variable	*Description*	*Range*	*Non-numerical variable*	*Description*	*Range*
v1_sat	Summer average temperature/°C	16.0~31.0	v13_HVAC	All zones:CAV	A0
v2_wat	Winter average temperature/°C	-11.0~23.2		All zones:VAV	A1
v3_tat	Transition average temperature/°C	4.5~24.9		All zones:FCU+OA	A2
v4_sarh	Summer average relative humidity	0.28~0.88		Inner :VAV Perimeter:FCU+OA	A3
v5_bsc	Building Shape Coefficient	0.10~0.50		All zones:VRV	A4
v6_wwr	Window wall ratio	0.10~1.00	v14_plant	CentiChiller & Boiler	P0
v7_ohtc	Overall Heat Transfer Coefficient, OHTC, w/m^2	5.0~35.0		ScrewChiller & Boiler	P1
v8_lpd	Lighting power density, w/m^2	10.0~20.0		Absorption chiller& Boiler	P2
v9_ppd	People power density, m^2/p	2.0~10.0		Ground source heat pump	P3
v10_epd	Equipment power density, w/m^2	10.0~20.0		Air source heat pump	P4
v11_sidt	Summer Indoor design temperature/°C	22.0~28.0		CentiChiller & Heat pump	P5
v12_widt	Winter Indoor design temperature/°C	15.0~22.0	v15_tspt	Variable speed pumps	Y/N
			v16_schd	Operation schedules	High/Std/Low usage

Case design

Taking inspiration from space filling design method, we introduced high-dimensional clustering as the deterministic part, and applied the common sampling as the randomized part. The two parts constituted the proposed hybrid space filling design method for the case design of the 12-Dimensional (12-D) numerical variables.

In the clustering process, high-dimensional space metrics are regarded as the key criterion for the clustering partition of the 12-D variable space. If the number of levels for each numerical variable is 3, the number of the overall cases by full permutation will be almost 530,000. We regarded a case having 12 variables as a data point in the high-dimensional space. All the cases construct a high-dimensional variable space. By means of space clustering, they are divided into 1,500 case clusters according to common sampling proportion.

Based on the case clusters, the Monte Carlo (MC) sampling method is used to stochastically select the numerical variable combination of the BPD case design from the cluster centres. Taken the centre of a cluster to represent its position in the high-dimensional space, we condensed the original space to a simplified space with lower resolution and the basic distribution features. By means of random sampling, about 65 numerical variable combination are achieved under the common sampling proportion of Monte Carlo method.

Simultaneously, given the discrete property of other four non-numerical, they are fully permutated along with the above 65 numerical variable combinations to obtain the final case design plan for the BPD, as shown in *Figure 1*.

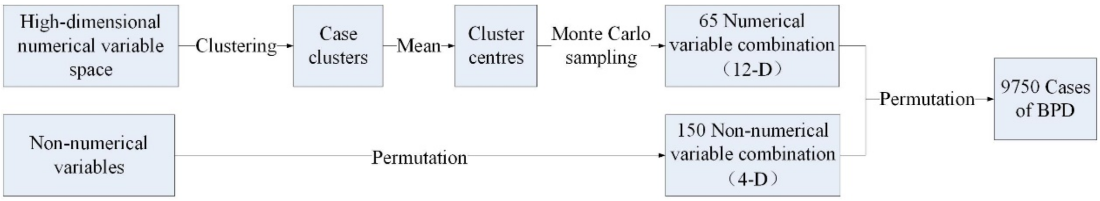
High-dimensional numerical variable space
Clustering
Case clusters
Mean
Cluster centres
Monte Carlo sampling
65 Numerical variable combination （12-D）
Permutation
9750 Cases of BPD
Non-numerical variables
Permutation
150 Non-numerical variable combination （4-D）

***Figure 1**. Case design plan for BPD*

Obviously, the clustering process is the key of hybrid case design. Clustering is the most important application of high-dimensional space metrics in the machine learning filed. Generally speaking, space metrics can be categorized as four types: (1) Norm distances, (2) Cosine distance, (3) Weighted distances, and (4) Other similarity distances, the algorithms of which are listed in *Table 2*. Due to the complexity of 12-D space distribution, it's difficult to directly choose a suitable metric for the clustering process. By comparing the different distribution of case clusters under different high-dimensional clustering metrics, the most suitable metric is determined to get reasonable 12-D space partition.

***Table 2.** Common high-dimensional space metrics*

Category	*No.*	*Metric*	*Algorithm*
Norm distance	1	Minkowski distance	$\mathrm{D_P}(\mathrm{a},\mathrm{b}) = \left[\sum_{i=1}^{s} \lvert \mathrm{a_i} - \mathrm{b_i} \rvert^{\mathrm{p}}\right]^{1/\mathrm{p}}$

			When P=1, named Manhattan distance When P=2, named Euclid distance When P=∞, named Maximum distance
	2	Linear combination distance (Wang Z et al. 2004)	$\alpha D_1(a,b) + \beta D_\infty(a,b)$
Cosine distance	3	Cosine distance	$r_{a,b} = \frac{\lvert \sum_{i=1}^{s} a_i \times b_i \rvert}{\sqrt{(\sum_{i=1}^{s} a_i{}^2) \times (\sum_{i=1}^{s} b_i{}^2)}}$
Weighted distance	4	Feature weighted metric with fuzzy weights (Wang J et al.2013)	$Wdist(a,b) = \sqrt{\sum_{h=1}^{s} w_h^\alpha (a_h - b_h)^2}$ $\alpha > 1,\ 0 < w_h < 1$
Similarity distance	5	Mahalanobis distance	$D_A(a,b) = (a-b)^T A^{-1}(a-b)$ A is the covariance matrix
	6	Lance&Williams distance	$D(a,b) = \sum_{i=1}^{s} \left\lvert \frac{a_i - b_i}{a_i + b_i} \right\rvert$

From the thought of the proposed case design method, it’s observed that the rationality of this plan depends enormously on performance of the clustering analysis. There are two kinds of evaluation index for clustering results. External index is a generally accepted reference partition, which is often impractical. Relatively, the more commonly used internal index directly focuses on the similarity of resulted clusters, including intra-cluster similarity (also called within-class divergence) and inter-cluster similarity (also called between-class variation). Good clustering makes birds of a feather flock together as much as possible, which means both the higher intra-cluster similarity and the lower inter-cluster similarity. Here, we used the most common K-means clustering method based on Euclid distance and the clusters similarity based on correlation coefficient separately as the external reference and internal index to illustrate the validity of the clustering process.

Parallel simulation process

Based on the above case design plan, we specified the 16-D building variables for each case in the targeted BPD. Besides, the building energy consumption for each case is the other crucial part of a complete BPD. In this study, we applied a java compiled tool-jEPlus for batch production and simulation of the almost 10,000 designed building cases to obtain the corresponding building energy consumption of the BPD.

JEPlus is an open source tool initially developed for managing complex parametric simulation using EnergyPlus (Zhang Y 2009). Parametric analysis using jEPlus provides a convenient and highly efficient way to perform optimisation for building design and operation. jEPlus uses a tree structure to organize the analysed parameters and their values, as well as the batch definition and generation of all the models. Then it drives EnergyPlus engine to execute the parallel multi-core computing and the building energy consumption collection afterwards.

RESULTS

According to the proposed BPD construction method, this section orderly implemented the four steps: (1) High-dimensional space clustering partition, (2) Numerical variable combination sampling, (3) Clustering performance evaluation, and (4) BPD simulation and construction.

High-dimensional space clustering partition

This step compares five different high-dimensional metrics in the 12-D space clustering. From the distribution of case number included in clusters, as listed in *Table 3*, it's obvious that the clusters by LCDist have well-proportioned distribution with the lower root-mean-square error (RSME), skewness and kurtosis. So LCDist was selected as the suitable clustering metric. With it, most of (80%) clusters include 200~400 cases. The best clusters present slight right-skewed distribution.

Table 3. *Distribution of Case clusters*

Metrics	*Average*	*Medium*	*RSME*	*Skewness*	*Kurtosis*
Lance&Williams (LanDist)	304	281	191	0.71	0.22
Mahalanobis (MaDist)	304	282	151	0.74	0.23
Euclid (EuDist)	304	282	150	0.75	0.27
Cosine (CosDist)	304	276	174	0.76	0.46
Linear Combination (LCDist)	304	287	136	0.64	0.09

Numerical variable combination sampling

Appling the MC sampling to the cluster centres, the numerical variable combinations are obtained. To ensure the computability of BPD cases that have physical significance, the seasonal weather parameters in the obtained 12-D variable combinations are matched with all the practical cities in the northern hemisphere. Then the typical meteorological year (TMY) of the matched practical city can be used as the hourly weather data input of that case. After the matching process, it's observed that the obtained 12-D variable combinations can cover the main climate zones defined by ASHRAE, see *Table 4.* And that, the deviations of the weather data between the BPD cases and their matched cities is less than 0.8ºC, which is acceptable.

Table 4. *Climate zone coverage of BPD cases*

Climate zones	*Representative city*	*Climate feature*	*Matched cases*
2A	Jacksonville, Florida, USA	Hot-humid	8
2B	Cairo, Egypt	Hot-dry	9
3A	Shanghai, China	Warm-humid	11
3B	March. AFB, California, USA	Warm-dry	6
3C	Kunming, Yunnan, China	Warm-marine	7
4A	Lyon, France	Mixed-humid	8
4C	Zhengzhou, Henan, China	Mixed- marine	8
5A	Shenyang, Liaoning, China	Cool-humid	11
5B	Denver, Colorado, USA	Cool-dry	5
6A	Toronto, Canada	Cold-humid	5
6B	Urumqi, Xinjiang, China	Cold- dry	3

Clustering performance evaluation

The external reference for clustering performance evaluation is the common K-means clustering partition with EuDist as the space metric. To test its stability, we did the K-means clustering twice with different initial centres for its iterative process. The comparison of the proposed LCDist clustering and the twice K-means clustering showed that, see *Figure 2*, the EuDists of the three batch of cluster centres are distributed similarly and have a narrow range (mainly 0.3~0.5), which means the differences of them are acceptable and the proposed clustering results are accurate.

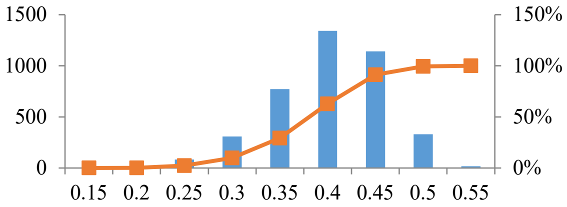

Figure 2*. Distance distribution of cluster centres among the proposed and reference clustering partition*

The internal index is the correlation coefficient of 12-D variable combinations. The correlations among 12-D cases included in the same cluster reflect the intra-cluster similarity, and those among 12-D cases included in the different clusters reflect the inter-cluster similarity. We found that the intra-cluster similarity of most clusters are 0.7~0.8, highly relevant, and the most inter-cluster similarity are less than 0.5, basically irrelevant. It explained that the proposed clustering process is validate.

Office BPD simulation and construction

The obtained 65 12-D numeric variable combinations using the hybrid space filling design method are then fully permutated with other 4 non-numeric variables to get the 16-D building inputs of the BPD models. After the two months' parallel simulation by jEPlus, this study constructed the BPD comprised of 9,750 building cases having high-dimensional mixed building variables and the corresponding multiple time-scale building energy consumption. Every case is uniquely identified with an ID. Its building variables include weather parameters, building shape, envelope, internal load, HVAC system and operation schedule. The results include whole building energy consumption, sub-metering consumption (e.g. cooling, heating and lighting etc.) and daily consumption. The high-dimensional space forming by the BPD basically reflects the complex relation between building factors and energy consumption of office buildings.

DISCUSSION

With consideration of the full-scale building factors, this study aims to establish a more general BPD for office building energy prediction and performance optimization by proposing an effective case design method. Given the high-dimensional and mixed features of building variables, several building variables are certainly simplified to balance the construction complexity and application universality of the BPD. The range of building variables including building shape, operation schedules, and energy system parameters just cover some typical types and couldn't include more segmentation.

The building shape in the BPD have two common types of rectangle and square office buildings. By comparing building energy models of different shapes, we found that the error of building energy consumption driven by other building shapes beyond the BPD scope, like L-shape, is about 5%.

According to existing standards, the BPD includes three usage levels of hourly operative schedules for different internal loads, like occupancy, lighting, equipment, and HVAC indoor set-point, as three operation scenarios. Assuming no obvious difference on commuter time of general office building, the difference of three scenarios is mainly reflected in the hourly usage profile in the working period. This simplification couldn't cover the complex effects of practical occupant behavior on building energy consumption.

As to the energy system variables including HVAC and plant system, the case design process of the BPD models just specified anyone of common system types. In the parallel simulation process, the detailed parameters of each system, like capacity and efficiency, basically use the default or general values. For the capacity of energy sources, certain adjustment of the BPD models is conducted to decrease the zone not-meet hours to less than 300, so that the BPD cases are reasonable and close to actual buildings as much as possible.

CONCLUSION AND IMPLICATIONS

Given the limitation of existing BPDs and their construction methods, this study proposed a hybrid space filling design method with combination of high-dimensional space clustering and stochastic sampling to develop an efficient case design plan for a comprehensive BPD of office building. Having high-dimensional mixed building variables and the corresponding multiple time-scale building energy consumption, the BPD can be utilized as the data basis for building energy prediction, optimization design and benchmark evaluation of office buildings. Besides, the proposed case design method integrated high-dimensional space metrics, space filling design with traditional experimental design to achieve more efficient BPD construction. Theoretically speaking, the method is suitable for other simulated databases with restriction on both variable dimensionality and computation scale, to decrease the calculation cost and the following data mining complexity.

An important note about this study is that it is part of the research on exploring necessary building variables for building energy prediction models of office building. As the main result of this study, the constructed BPD is going to be the data basis for the next part of the mentioned research. The BPD can provide the analysis basis for figuring out the complex relationship between building variables and energy consumption used in building energy prediction models.

REFERENCES

Lee SH, Hong T, Mary AP, Sarah CTL. 2015. Energy retrofit analysis toolkits for commercial buildings: A review, Energy. 89:1087-1100.

Lee SH, Hong T, Mary AP, Geof S, Chen Y, Sarah CTL. 2015. Accelerating the energy retrofit of commercial buildings using a database of energy efficiency performance, Energy. 90:738-747.

Amiri SS, Mottahedi M, Asadi S. 2015. Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the U.S., Energy and Buildings.109:209-216.

Ivan K. 2013. Regression models for predicting UK office building energy consumption from heating and cooling demands, Energy and Buildings.59:214-227.

Mao J, Pan Y, Fu Y. 2016. Towards fast energy performance evaluation: A pilot study for office buildings, Energy and Buildings. 121:104-113.

Kim Y, Ahn K, Park C. 2014. Decision making of HVAC system using Bayesian Markov chain Monte Carlo method, Energy and Buildings.72:112-121.

Asadi E, Da SMG, Antunes CH, Dias L, Glicksman L. 2014. Multi-objective optimization for building retrofit: A model using genetic algorithm and artificial neural network and an application, Energy and Buildings. 81:444-456.

Yang L, Hou L, Li H, Xu X, Liu J. 2015. Regression models for energy consumption prediction in air-conditioned office building, J. Xi'an Univ. of Arch. & Tech. (Natural Science Edition). 47(5):707-711.

Neto AH, Fiorelli FAS. 2008. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption, Energy and Buildings. 40(12):2169-2176.

Wang Z, Li H, Zhou C. 2004. On computing for replacing Eulideans distance by linear combination of Cityblock and Chessboard distances in high dimensional space, Mini-micro systems, 25(12) :2120-2125.

Wang J, Wang S, Chung F, Deng Z. 2013. Fuzzy partition based soft subspace clustering and its applications in high dimensional data, Inform Sciences.246:133-154.

Zhang Y. 2009. “Parallel” EnergyPlus and the development of a parametric analysis tool, *Building Simulation 2009: Eleventh International IBPSA Conference*, Glasgow, Scotland.