Predicting hourly energy consumption in buildings using occupancy-related characteristics of end-user groups

Kwonsik Song[a], Nahyun Kwon[a], Kyle Anderson[b], Moonseo Park[a*], Hyun-Soo Lee[a], and SangHyun Lee[b]

[a]Department of Architecture and Architectural Engineering, Seoul National University, Room 39-425, Gwanak-ro 1, Gwanak-Gu, Seoul 151-742, South Korea

[b]Department of Civil and Environmental Engineering, University of Michigan, 2350 Hayward St., 2340 G.G. Brown Building, Ann Arbor, MI, United States of America

*Corresponding author at: Department of Architecture and Architectural Engineering, Seoul National University, Room 39-433, Gwanak-ro 1, Gwanak-Gu, Seoul 151-742, South Korea; email: mspark@snu.ac.kr

emails: woihj@snu.ac.kr, prideknh@snu.ac.kr, kyleand@umich.edu, mspark@snu.ac.kr, hyunslee@snu.ac.kr, shdpm@umich.edu

Abstract

Accurate predictions of energy consumption are essential to optimizing building energy use performance. To date, substantial efforts have been undertaken to improve prediction accuracy, specifically while focusing on occupants' presence in buildings. Unfortunately, two significant obstacles remain when predicting building energy consumption using occupancy data. First,

occupancy diversity among end-user groups is rarely considered during model development. Second, occupancy's correlation with energy consumption may be weak due to variances in occupant behavior. Therefore, this research aims to investigate how occupancy-related characteristics of end-user groups affect prediction performance. In order to achieve this objective, a data mining-based prediction model is constructed to mimic building thermal behaviors. The experimental results using the proposed prediction model make it evident that prediction accuracy is improved when considering diverse occupancy and its correlation with energy use. In addition, significant prediction accuracy is achieved using only a minimal amount of historical data. With the proposed prediction model, it is possible to obtain more detailed information about energy use patterns (e.g., load shape, the amount of energy use) for end-user groups. Thus, facility managers will be able to personalize the operation of energy-consuming equipment depending on end-user group for reducing energy consumption without compromising occupants' thermal comfort.

Keywords

Energy Saving; Energy Use Prediction; Data Mining Techniques; Occupancy Status; Occupants' Energy Use Behavior

1. Introduction

As buildings consume 40% of all energy globally, improving their performance remains a critical task in order to meet energy saving goals [1]. Accordingly, much effort has been made to reduce the energy use in buildings. Being able to accurately predict energy use in buildings is essential to optimize the operation of energy-using equipment during a buildings operation [2-4]. Once it is understood where energy is consumed within a building, it is possible to develop appropriate energy saving strategies. This enables facility managers to achieve energy saving in the following three ways: 1) efficiently set starting/finishing time of heating, ventilating, and air conditioning (HVAC) systems, 2) avoid reaching peak energy demand by pre-heating/cooling of buildings, and 3) adjust heating and cooling setpoints during the peak energy periods.

In the extensive literature on building energy use prediction, various influential factors have been considered due to their significant correlation with energy consumption. These factors include: weather, building characteristics, equipment, and occupant-related characteristics [5-8]. Among these factors, recent studies have emphasized the importance of occupancy since occupants interact with energy-consuming equipment and devices in the built environment [8-12]. When attempting to predict building energy use, many previous efforts have employed fixed occupancy schedules as an alternative to using actual building occupancy data for simplicity and due to lack of readily available data [8,9]. However, smart monitoring systems now allow us to automatically obtain high-resolution occupancy data (e.g., 1h time interval) and has begun to be used when constructing building energy use prediction models [10-13]. It has been found that considering occupancy as an input variable during model development improves prediction performance.

Unfortunately, despite the recent advancements in prediction accuracy, two significant obstacles remain when predicting energy consumption using occupancy data. First, diverse

63 occupancy, which refers to differences in occupancy status among end-user groups (EUGs), has
64 rarely been considered during model development. This is important because many buildings have
65 different EUGs which have different occupancy patterns. For instance, university buildings have
66 complex occupancy patterns due to the varying functions of rooms: administration, research,
67 lecture, and seminar [14]. Most studies [7-13] to date have eliminated occupancy diversity by
68 averaging the values of occupancy status at the building level, which may contribute to
69 discrepancies between actual and predicted energy use. Other studies have employed spatially
70 granular data (e.g., floor and unit level) and individual equipment level data for building energy
71 use prediction, but did not consider occupancy diversity during model development [2,15]. Second,
72 occupancy's correlation with consumption may be weak at time due to variances in occupant
73 behavior. If occupants fail to switch off their equipment and devices before leaving, energy will
74 be consumed while unoccupied and uncorrelated with occupancy [16-19]. As a consequence, it
75 can be difficult to ensure an improvement in prediction accuracy, i.e., the correlation effect. Most
76 studies to date have used occupancy as an input variable without considering its correlation with
77 energy use [11,12]. In rare studies that have investigated the correlation between energy use and
78 occupancy status, the correlation effect remains unclear because there was no attempt to compare
79 the performance of prediction models with different correlation levels [10,13].

80 Therefore, this research aims to investigate how occupancy-related characteristics of
81 EUGs affect energy use prediction performance. In order to achieve this objective, a data mining-
82 based prediction model is constructed since it facilitates to mimic building thermal behaviors [20]
83 and identifies representative EUGs within buildings [21,22]. The developed model will provide
84 more accurate information about daily peak demand and daily energy use in buildings.
85 Furthermore, it is expected that the model makes it possible to recognize energy use patterns for

EUGs. In turn, facility managers will be able to personalize the operation of energy-consuming equipment depending on EUG.

This paper is organized as follows. First, a literature review is presented on the application of occupancy data to building energy use prediction and prediction methodologies. Second, data mining techniques relevant to this work are introduced and discussed. Third, a data mining-based prediction model is developed by using real-world data collected from buildings in Seoul, South Korea. Next, the experimental results using the proposed prediction model are presented and the paper concludes with a discussion of the results followed by the conclusion.

2. Literature Review

2.1 Application of occupancy data to building energy use prediction

In the extensive literature on building energy use prediction, occupancy data is typically substituted with building or equipment schedules which indirectly reflects the behavior of occupants. Kwok et al. [6] proposed a multi-layer perceptron model for building energy use prediction using the power consumption of primary air-handling units (PAU) as an alternative to occupancy data to mimic occupants' presence in a building. Yezioro et al. [8] constructed an artificial neural network (ANN) prediction model which uses the occupancy schedule as an input variable. In the optimized ANN model for building energy forecasting suggested by Li et al. [9], the opening schedules of a library were used to represent the hourly occupancy of each reading room.

Due to recent advancements in technology, it has become possible to monitor occupancy in real-time. With this new capability, researchers have begun using actual occupancy data to simulate occupants' behavioral characteristics in temporal and spatial contexts. Sandels et al. [10]

109 presented a data analysis approach for conducting day-ahead predictions of electricity consumed
110 by appliances, ventilation systems, and cooling equipment in an office building floor. It is found
111 that the most significant predictor for the appliance load is the occupancy ratio. Virote and Neves-
112 Silva [11] produced reliable predictions for building energy use by integrating stochastic occupant
113 behavioral models with energy consumption models. Wang and Ding [12] proposed an occupancy-
114 based energy consumption prediction model. The prediction model used a time-varying indoor
115 occupancy rate which is obtained by using Monte Carlo simulation and Markov chain model. As
116 a method to quantify energy savings by measurements and verification, Liang et al. [13] developed
117 an energy baseline model using a short-time interval data on the number of occupants.

118 As mentioned above, a substantial number of studies investigated the effect of occupancy
119 on the performance of building energy use prediction. However, the occupancy data used in
120 previous studies was mostly simplified through aggregating occupancy status at the building level.
121 Furthermore, when constructing prediction models, there were limited attempts to investigate
122 occupancy's correlation with energy use. As a consequence of such significant obstacles, there still
123 remains a discrepancy between the actual and predicted energy use.

124 ***2.2 Prediction approaches***

125 Various building energy use prediction models have been proposed and can be categorized
126 as: engineering, statistical, or ANN [23]. Engineering models simulate building energy use based
127 on the physical and environmental factors [7-8]. While this approach has the advantage of being
128 able to calculate elaborate thermal dynamics at a building level, it is not without limitation. This
129 method can be a complicated difficult process which involves constructing a simulation model and
130 obtaining meaningful input data [24]. Statistical models predict building energy use by using
131 historical energy use data together with the measured input data [10,25,26]. While the structure of

this approach is well understood due to the simplicity of the model parameters, substantial effort is required to overcome autocorrelation and multi-collinearity problems [4]. Lastly, ANN operates training procedures using historical data and then predicts building energy use [4-9]. This approach is highly applicable for solving non-linear and complex problems [27]. Not without limitations as well, ANN face potential problems with the reliability and accuracy of prediction results since it relies on the training data and can be computationally intensive [20, 28].

As described thus far, each approach has its own advantages and disadvantages. Nevertheless, among these approaches, ANN prediction models are becoming increasingly more common in the field of building energy use prediction because the thermal behaviors of a building involve a non-linear problem [20].

3. Methodology

For this research, three data mining techniques are employed to predict the total amount of building energy use. First, *k*-means algorithm is used to investigate representative EUGs within buildings due to its ability to categorizes a set of objects into meaningful groups [29]. Second, artificial neural networks are constructed to predict energy use for the identified EUGs because it facilitates to mimic thermal behaviors of a building [20]. Third, *k*-nearest neighbor with simple classification capability is introduced to select an appropriate set of historical data for network training [30].

3.1 k-means algorithm

In order to predict building energy use, most studies averaged the values of occupancy status at the building level. However, in practice, there are various occupancy patterns depending

on EUG. Consequently, this can result in a discrepancy between the actual and predicted energy use. In order to consider occupancy diversity, this study conducts clustering analysis using *k*-means algorithm to investigate representative EUGs within buildings.

The *k*-means algorithm is a clustering method which categorizes a set of objects into meaningful groups [29]. Within one group, objects have high intra-class similarity and low inter-class similarity. This clustering algorithm consists of four steps: 1) arbitrarily determine *k* objects as the initial centroids of groups; 2) assign each object to clusters with the closest centroid based on the distance between an object and centroid for each cluster; 3) replace the current centroid with the object having mean value of each cluster; and 4) iterate step 2 and 3 until there are no more new assignment.

For the *k*-means algorithms, since the number of clusters cannot be known in given datasets, it is difficult to determine the best number of clusters. To overcome this difficulty, the Davies-Bouldin Index (DBI) is adopted as clustering evaluation criteria which is the mean value of a ratio of inter-cluster and intra-cluster distances [31]. Although many clustering evaluation criteria (e.g., Silhouette Index, Clustering Dispersion Indicator) have been proposed to investigate an optimal number of clusters in a given dataset, the DBI has a high capability regardless of data properties such as monotonicity, noise, density and skewed distributions [32]. In particular, as summarized by Chicco [33], the DBI shows a wide applicability in the field of end-user group identification. The DBI can be calculated by the following equation:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}\left\{\frac{\bar{d}_i + \bar{d}_j}{d_{ij}}\right\} \quad (1)$$

where k: the number of clusters; d_i: the average distance between all objects in the i_{th} cluster and the centroid of the i_{th} cluster; d_j: the average distance between all objects in the j_{th} cluster; and d_{ij}: the distance between the centroids of the i_{th} and j_{th} clusters. The minimum value of DBI corresponds to good clusters and is regarded as an optimal clustering solution.

3.2 Artificial neural network

In order to predict energy consumption for EUGs, an ANN is used which has a computational structure which mimics a biological neural system of human brain [34]. In general, the ANN consists of innumerous collections of neurons, which are linked with one another. Each individual neuron obtains multiple input values from other connected neurons to produce a single output value. According to this physical scheme, ANN learns and generalizes the relationships in the given datasets, and then extrapolates results for new datasets. Given this capability, ANN has been successfully applied to solve pattern recognition, classification, and forecasting problem.

Fig. 1 illustrates the structure of a typical three-layer feed forward neural network. The network consists of three types of layers in which the neurons are placed. The first layer, called input layer, obtains inputs from outside. The second one is the output layer, which produce the results evaluated by the network. Lastly, a hidden layer exists between the input and output layer. It should be noted that each neuron of a given layer is connected to other neurons of a previous layer by a weighted links. For a three-layer network, the mathematical function is defined as

$$Y = f\left(b_0 + \sum_{j=1}^{k} h\left(\varphi_j + \sum_{i=1}^{m} p_i w_{ij}\right) b_j\right) \tag{2}$$

where Y: the network outputs; $f(\cdot)$: nonlinear transfer function; p_i: the network inputs; b_0: the

197 output bias; b_j: the weight values from hidden layer to output layer; φ_j: the hidden layer biases;
198 w_{ij}: the weights from input layer to hidden layer; $h(\cdot)$: hidden layer activation function. In order
199 to produce better outputs, the gradient performance function is used through adjusting the weight
200 and bias parameters.
201
202 < **Fig. 1**. Structure of a three-layer feed forward neural network >
203
204 ***3.3 k-nearest neighbor***
205 In the process of constructing a prediction model, the collected historical data is used to
206 adjust the weight and bias parameters of the ANN. Once a large amount of historical data is
207 available, it will be convenient to mimic building thermal behaviors by optimizing the values of
208 these parameters with a global solution [28]. On the other hand, this can lead to inadequate
209 prediction performance for the following two reasons. First, a trained ANN model using a large
210 dataset cannot provide acceptable prediction accuracy since it does not always have useful
211 information to predict future energy consumption [15]. For example, if ANN models are trained
212 using a large dataset collected during summer months, it is impossible to predict heating energy
213 use because the historical data does not include records of building thermal behaviors in the other
214 seasons. Further, when randomly selected training datasets (e.g., weekdays) do not have similar
215 values for input variables to those of test datasets (e.g., weekends), prediction models perform
216 poorly. Second, network training using a large dataset can be a time-intensive endeavor since more
217 attempts should be made to investigate optimal network parameters (i.e., weight, bias) [35]. Further,
218 considering that there are additional efforts to improve the quality of historical data (e.g., data
219 preprocessing), ANN models suffer from longer computational time as the size of training datasets

increases [36].

In order to overcome these problems, an appropriate set of historical data will be selected and used by the *k*-nearest neighbor (KNN) in this research. The *k*-nearest neighbor searches for *k* similar historical datasets that are close to a given test dataset [30]. In this context, the closeness is calculated using distance measures such a Euclidean distance. If $p = (p_1, p_2, p_3,..., p_n)$ and $q = (q_1, q_2, q_3,..., q_n)$ are two objects, the Euclidean distance is defined as

$$dist(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{3}$$

Before calculating the distance between objects, it should be noted that each variable is normalized using normalization methods such as Min-Max normalization, Z-score normalization, and sigmoid normalization since the variables in larger units can dominate other variables. In this research, Min-Max normalization is used due to its applicability in the field of data mining. The normalized value of $v_{i\cdot A}^{'}$ for variable A is calculated as follow

$$v_{i\cdot A}^{'} = \frac{v_{i\cdot A} - v_{min\cdot A}}{v_{max\cdot A} - v_{min\cdot A}}\left(v_{max\cdot A}^{'} - v_{min\cdot A}^{'}\right) + v_{min\cdot A}^{'} \tag{4}$$

where $v_{i\cdot A}$: an existing value of variable A; $v_{max\cdot A}^{'}$: the new maximum value of variable A; $v_{min\cdot A}^{'}$: the new minimum value of variable A; $v_{max\cdot A}$: the existing maximum value of variable A; and $v_{min\cdot A}$: the existing minimum value of variable A. Based on the distance value obtained from Eq.

238 (3), the *k* corresponding closest datasets are obtained from the given historical dataset. More
239 importantly, to find the best number of the k variable, experiments need to be conducted by
240 adjusting the number of similar datasets, the results of which then need to be evaluated.
241
242 **4. Model Development**
243 Based on the three types of data mining techniques outlined above, a building energy use
244 prediction model is constructed in MATLAB with the Neural Network Toolbox. The proposed
245 prediction model considers diverse occupancy scenarios and the correlation of them with energy
246 consumption during model development. Data was collected by surveying seven university
247 dormitory buildings in Seoul, South Korea (see Table 1). The buildings have 250 single occupancy
248 and 1125 double occupancy rooms, which are allocated to students at the beginning of the spring
249 and fall semesters. In all the rooms smart metering systems measure the hourly electrical energy
250 consumed by under-floor heating systems, mini-refrigerators, lights, and other plug loads. In
251 addition, card entry systems provide a record of changes in occupancy status via entry and exit
252 data for each room.
253 From October 6, 2014 to March 1, 2015, historical data on energy use and occupancy was
254 collected hourly. Two hundred and twenty eight rooms are excluded in the construction of the
255 prediction model due to vacancy, reader malfunction, and erroneous consumption data. In addition,
256 due to system malfunctions no data was collected on November 28 and 29, 2014. Thus, the
257 historical dataset includes a total of 3480 time steps.
258
259 < **Table 1**. Overview of case buildings >
260

After the data collection and preprocessing are completed, we select the input variables to construct the prediction model. Then, the data-mining based prediction model is developed which considers occupancy-related characteristics of EUGs in buildings.

4.1 Input variable selection

Identifying significant determinants of building energy use is an important task because the prediction accuracy can be compromised using input variables that have a weak relation to energy use [37]. Previous studies have investigated the impact of various input variables on building energy use [38-40]. From their empirical analysis, it was recognized that the weather, the building's physical characteristics, equipment, and occupant-related characteristics have a significant impact on building energy use.

In this research, the proposed prediction model employs seven input variables (see Table 2). The first four weather variables (i.e., outside dry-bulb temperature, wind speed, relative humidity, and solar radiation) are selected due to their high correlation to energy use. As an occupant-related characteristic, occupancy rate refers to the ratio of occupied rooms compared to the total number of rooms at a certain time. An occupancy rate value close to 1 indicates a higher possibility for occupants to be present in their rooms. Additionally, since students consume different amounts of electrical energy over time, we account for the following two date-related input variables. The first input variable is the day of the week, which consists of weekdays and the weekend. Across all the given periods, more energy is consumed during weekdays than weekends (see Fig. 2). This difference can be expected since students living in dormitories tend to leave campus on weekends for their private life (e.g., visit to family). As a consequence of their weekend leave, dormitory buildings may consume less energy during weekends. Second, the course period is an indicator variable which denotes fall and winter semester. Fig. 3 shows that there is a

284 significant difference in average hourly energy use between fall and winter semester. These results
285 stems from the fact that most students move out of their rooms over winter semester.
286
287 < **Table 2**. Input variables for building energy use prediction >
288
289 < **Fig. 2**. Average hourly energy use by day of the week >
290
291 < **Fig. 3**. Average hourly energy use by course period >
292
293 ***4.2 Structure of data mining-based prediction model***
294 In order to examine the effect of occupancy-related characteristics on the prediction
295 performance, a data mining-based prediction model is developed which consists of four modules
296 named according to their role (see Fig. 4). As a beginning of predicting building energy use, the
297 entry module collects historical data concerning input and output variable. Based on the historical
298 datasets, the datasets selection module determines similar daily datasets to improve the prediction
299 performance. Next, in an effort to investigate diverse occupancy in buildings, the cluster
300 identification module investigates representative EUGs using the similar daily datasets. For the
301 identified EUGs, the prediction module then examines the correlation coefficient between energy
302 use and occupancy status, and constructs prediction sub-models.
303
304 **< Fig. 4**. Main structure of data mining-based energy use prediction model >
305
306 ***4.2.1 Data entry module***

307 The data entry module imports historical data on outside dry-bulb temperature, wind speed,
308 relative humidity, solar radiation, occupancy rate, day of the week, college year, and energy use.
309 For the weather and occupant-related variables, the historical data is collected every hour. The
310 daily data for date-related variables is also obtained. Following the data collection, data
311 preprocessing is carried out to improve the performance of building energy use prediction as
312 follows. First, if there are missing and abnormal values, this module excludes the data from
313 analysis. Second, the collected data from multiple sources is combined to form a dataset which
314 includes the values of seven input variables and an output variable. Third, the preprocessed dataset
315 is sent to the next module to investigate similar daily datasets for model development.

316 ***4.2.2 Dataset selection module***

317 In the dataset selection module, a test dataset is determined to evaluate the prediction
318 accuracy. Then, in order to address the problems caused by a large amount of historical data, this
319 module determines k similar daily datasets to be used for model development. In this context, the
320 similar daily datasets are obtained by averaging the values of seven input variables imported from
321 the data entry module. After constructing the daily datasets, the k-nearest neighbor algorithm
322 searches for k daily datasets that have similar values for input variables to those in the
323 predetermined test dataset. In this process, similarity is determined by Euclidean distance
324 introduced in Eq. (3). Also, min-max normalization is used to minimize the scale difference among
325 the input variables. After investigating k similar daily datasets, these outputs are sent to the next
326 modules to find representative EUGs in buildings.

327 ***4.2.3 Cluster identification module***

328 The cluster identification module has two functional roles as follows. First, daily energy
329 use patterns for EUGs are identified by clustering analysis. In order to perform this role, the k-

means algorithm is applied to the similar daily datasets obtained in the dataset selection module. Clustering performance is then assessed using the DBI value to determine the best number of EUGs. The second role is to investigate daily occupancy patterns in buildings. For the identified EUGs, the average hourly occupancy rate is calculated from the following equation.

$$R^k(t) = \frac{1}{d} \cdot \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n} O_{ij}^k(t) \quad (5)$$

where $R^k(t)$: average occupancy rate for the k^{th} cluster at time t; d: the number of similar daily datasets; n: the number of rooms for the k^{th} cluster; and $O_{ij}^k(t)$: occupancy state (0: vacancy or 1: presence) of the i^{th} room for the k^{th} cluster at time t of the j^{th} similar dataset. After conducting these identification processes, representative daily profiles for the EUGs are exported to the next module.

4.2.4 Prediction module

The prediction module provides information about the amount of energy that will be consumed in buildings for the next few days. In an effort to improve the accuracy of building energy use prediction, multiple prediction sub-models using ANN are constructed depending on the EUG identified in the cluster identification modules. Also, for the identified EUGs, the correlation analysis between energy use and occupancy rate is performed to investigate occupants' behaviors during unoccupied periods. This is important because occupancy's correlation with consumption may be weak at time due to variances in occupant behavior [16-19]. As observed in previous studies [13], occupancy status is not significantly correlated with energy use when occupants leave on energy-consuming equipment in vacant rooms. In order to conduct this analysis, this module investigates the Pearson's correlation coefficient (r) between average hourly energy

use and average occupancy rate. If the correlation coefficient is higher than 0.5 (a threshold for determining a significant correlation [41]), the average occupancy rate is used as one of the input variables.

In the course of network training, the k similar daily datasets are randomly divided into two independent datasets for training and validation, respectively. After the network training is complete, the predetermined test dataset is used to predict building energy use by adding the results drawn from all the sub-models. Based on the predicted building energy use, the performance of the fully-trained predictor is assessed using the coefficient of variation of the root mean squared error (CV-RMSE). For this performance index, the CV-RMSE value is given by combining $RMSE$ and $\bar{Y}_i$ as follows:

$$CV - RMSE = \frac{RMSE}{\bar{Y}_n} \times 100\% \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_i^n \left(Y_i - \check{Y}_i\right)^2}{n}} \tag{7}$$

where $RMSE$: root mean squared error; $\bar{Y}_n$: average value of actual energy use during the prediction period n; Y_i: actual energy use at time i; and $\check{Y}_i$: predicted energy use at time i.

5. Results

5.1 Experimental Design

In order to investigate how occupancy-related characteristics of EUGs affect prediction

370 performance, comparative experiments are conducted using four different prediction models (see
371 Table 3). The first ANN model (BL-1) investigates the average value of occupancy rate at the
372 building level. For the second prediction model (BL-2), the average occupancy rate is employed
373 in case of its significant correlation with building energy use. Alternatively, the remaining models
374 (GL-1 and GL-2) use average values of occupancy rate for the EUGs. However, in the GL-2, the
375 average occupancy rate is used when its high correlation with energy use exists
376
377 < **Table 3**. Description of different prediction models >
378
379 Across all the prediction models, a parameter setting is identically performed as follows
380 (see Table 4). Given a test dataset (24 time steps), 10 similar daily datasets (240 time steps) are
381 selected and randomly divided into proportions of 80% (192 time steps) and 20% (48 time steps)
382 for training and validation, respectively. Using such a randomized division process, the network
383 training and validation is performed 10 times to find the best CV-RMSE values. In the cluster
384 identification model, the value of the cluster k variable varies from 2 to 10 in order to identify the
385 optimal number of clusters. Further, the number of hidden neurons is determined using Eq. (8),
386 adopted in related works [9,34].
387

$$N_h = \sqrt{N_i + N_o} + C \quad (8)$$

388
389 where N_h: number of hidden neurons; N_i: number of input variables; N_O: number of output
390 variable; C: an integer between 1 and 10. For this study, the proper number of hidden neurons is

391 within a range of 4 to 13. However, since prediction performance can vary depending on the
392 number of hidden neurons, a sensitivity analysis is conducted using the GL-2 with 4 to 13 hidden
393 neurons. As shown in Fig. 5, the best values of CV-RMSE are similar to one another. However,
394 the GL-2 with 10 to 13 hidden neurons provides a relatively low distribution of CV-RMSE values.
395 This indicates that when using more than 10 hidden neurons, the GL-2 performs better at predicting
396 energy use. In particular, considering the computation time increases with a larger number of
397 hidden neurons, 10 hidden neurons are optimal to fit the proposed ANN model [42].
398
399 < **Table 4**. Main parameters used for network training >
400
401 < **Fig. 5**. CV-RMSE by number of hidden neurons >
402
403 Based on this experimental design, the prediction results for the next day and five days
404 are comprehensively compared to investigate the effect of occupancy-related characteristics on
405 building energy use prediction. In the next section, we discuss the prediction results and suggest
406 an improvement for the proposed model.
407 ***5.2 Next Day Prediction Results***
408 In order to compare the prediction performance for the next day, experiments are
409 conducted using five test datasets (TD1: October 15, 2014; TD2: November 14, 2014; TD3:
410 December 22, 2014; TD4: January 12, 2015; TD5: February 5, 2015). As described in Table 5,
411 these test datasets list the different values for seven input variables to provide a basis for validating
412 the prediction results for the next day.
413

414 < **Table 5**. Average values of input variables for next day building energy use prediction >

415

416 Fig. 6 shows the DBI values by the number of EUGs. During all the given prediction
417 periods, the lowest values of DBI are found when two EUGs exist in the case buildings (TD1:
418 0.6733; TD2: 0.6066; TD3: 0.6487; TD4: 0.6709; TD5: 0.6991). As shown in Fig. 7, the identified
419 EUGs have different energy use patterns. For the EUG 1, it can be seen that a small amount of
420 electrical energy is consumed in vacant rooms. In contrast, the EUG 2 tends to consume a
421 significant amount of electricity regardless of occupancy status.

422

423 < **Fig. 6**. Davies-bouldin index by the number of clusters for next day building energy use
424 prediction >

425

426 < **Fig. 7**. Daily profiles of energy use and occupancy status for next day building energy use
427 prediction >

428

429 Table 6 presents the results of the correlation analysis of the EUGs and building. During
430 the given prediction periods, there is a significant correlation between energy use and occupancy
431 status for the EUG number 1. On the other hand, a weak correlation exists for the EUG number 2.
432 When investigating the correlation coefficient at the building level, building energy use is
433 generally significantly correlated with occupancy status.

434

435 < **Table 6**. Correlation coefficient between energy use and occupancy status for next day
436 building energy use prediction>

437

438 Comparing the prediction performance for the next day, the GL-2 provides the best values
439 of CV-RMSE at 14.4% for the TD1, 4.8% for the TD2, 3.9% for the TD3, 4.5% for the TD4, and
440 3.7% for the TD5 (see Fig. 8); a smaller CV-RMSE indicates better prediction performance. The
441 distribution of CV-RMSE values is relatively low in the GL-2. The execution time for training the
442 GL-1 and GL-2 is longer than the other prediction models. When examining the effect of
443 occupancy diversity on the CV-RMSE values, the GL-2 offers higher accuracy than both the BL-
444 1 and BL-2. However, the GL-1 does not always produce better prediction accuracy than both BL-
445 1 and BL-2. Looking closely at the CV-RMSE values by correlation effect, the BL-2 and GL-2
446 produce more accurate prediction results for the next day than both the BL-1 and the GL-1. In the
447 experiment using TD2 ($r < 0.5$ at the group and building level), the CV-RMSE values for the BL-
448 2 is more accurate than the BL-1 at 5.2% versus 5.4%. Also, compared to the CV-RMSE of 5.8%
449 for the GL-1, the relatively low value of 4.8% for the GL-2 is produced.

450

451 < **Fig. 8**. CV-RMSE and computation time for next day building energy use prediction >

452

453 Additionally, in order to investigate how the number of similar daily datasets affect the
454 prediction performance, the GL-2 is trained using 1 to 30 similar daily datasets. As shown in Fig.
455 9, the prediction accuracy of GL-2 decreases with a larger number of similar daily datasets. On the
456 other hand, training time does not significantly differ according to the number of similar daily
457 datasets.

458

459 < **Fig. 9**. Prediction performance for the next day by number of similar daily datasets >

460

461 ***5.3 Next Five Days Prediction Results***

462 In order to compare the prediction performance for the next five days, experiments are
463 performed using three test datasets (TD1: November 20-24, 2014; TD2: December 5-9, 2014; TD3:
464 January 24-28, 2015). As shown in Table 7, these test datasets encompass different values for seven
465 input variables. Therefore, it is possible to validate the prediction performance for the five days.

466

467 < **Table 7**. Average values of input variables for next five days building energy use prediction >

468

469 Fig. 10 represents the results of clustering analysis using 10 similar daily datasets. During
470 the given prediction periods, there are various EUGs in the case buildings (TD1: 11 EUGs; TD2:
471 12 EUGs; TD3: 16 EUGs). For all the identified EUGs, we conduct the correlation analysis
472 between energy use and occupancy status. As shown in Table 8, there are differences in correlation
473 coefficient values among the EUGs. When investigating the correlation coefficient at the building
474 level, it substantially varies within the prediction periods.

475

476 < **Fig. 10**. Davies-bouldin index by number of clusters for next five days building energy use
477 prediction >

478

479 < **Table 8**. Correlation coefficient between energy use and occupancy status for next five days
480 building energy use prediction >

481

482 Fig. 11 describes a comparison of prediction performance for the next five days. Based on

483 the lowest value of CV-RMSE, the GL-2 is the best model to predict building energy use for the
484 next five days (TD1: 12.6%; TD2: 7.5%; TD3: 7.2%). The difference between the maximum and
485 minimum values of CV-RMSE is relatively low in the GL-2. In addition, the GL-1 and GL-2
486 generally take a longer time for its network training. Looking closely at the prediction performance
487 by occupancy diversity, the lowest CV-RMSE values for the GL-2 are observed during all the
488 given prediction periods. However, the GL-1 does not always result in a higher value of CV-RMSE
489 than both the BL-1 and BL-2. When investigating the correlation effect on prediction performance,
490 the BL-2 and GL-2 are more accurate than the BL-1 and GL-1, respectively. In the experiments
491 using the TD2 ($r < 0.5$ at the group and building level), the CV-RMSE for the BL-2 is lower than
492 the BL-1 at 8.3% versus 9.1%. Also, compared to the CV-RMSE of 8.6% for the GL-1, the
493 relatively low value of 7.5% for the GL-2 is yielded.
494
495 < **Fig. 11**. CV-RMSE and computation time for next five days building energy use prediction >
496
497 In order to investigate the effect of the number of similar daily datasets on the prediction
498 performance, the GL-2 is trained within a range from 1 to 30 similar daily datasets. As shown in
499 Fig. 12, the CV-RMSE values increase as the number of similar daily datasets increases. However,
500 it is difficult to find a consistent tendency in computational time for training the GL-2.
501
502 < **Fig. 12**. Prediction performance for the next five days by number of similar daily datasets >
503
504 **6. Discussion**
505 Figs. 8 and 10 compare the prediction performance by CV-RMSE. Considering that a

smaller CV-RMSE indicates better prediction performance, it is observed that the GL-2 provides more reliable accuracy within acceptable tolerances (CV-RMSE, 30%) than the other prediction models do [43]. Although only a small gap in CV-RMSE values are observed among the prediction models, this is significant because the GL-2 shows relatively low variance in CV-RMSE values regardless of the random selection of training and validation datasets. Therefore, it can be inferred that occupancy-related characteristics of EUGs significantly contribute to the prediction performance.

More specifically, it appears that occupancy diversity affects the prediction performance. Compared to the BL-1 and BL-2, the GL-2 that has multiple prediction sub-models produces the lowest CV-RMSE values. These results can be expected since the GL-2 elaborates occupants' presence at the group level and aggregate the prediction results produced by sub-models. Considering that prediction performance increases with more granular data (e.g., floor, unit, equipment), it is apparent that the use of simplified occupancy data undermines the performance of building energy use prediction [2]. Additionally, the improved prediction accuracy is observed in the BL-2 and GL-2, which consider the correlation between energy use and occupancy status. Although they do not account for occupancy rate as an input variable, the CV-RMSE values are lower than those for the BL-1 and GL-1. These improvements can be expected since occupancy status is not always significantly correlated with energy use during the given prediction periods and thus could have a detrimental effect on the prediction accuracy. The lack of correlation could have resulted from poor occupant behavior, i.e., leaving on equipment while not in the room. As shown in Fig 7, the EUG without correlations tends to consume a significant amount of electrical energy while rooms are unoccupied, likely due to occupants not switching off the heating system before leaving. Further, the training time of GL-1 and GL-2 is longer than that the BL-1 and BL-

529 2 (see Figs. 8 and 11). This is because multiple prediction sub-models should be individually
530 trained according to the number of EUGs. These experimental results indicate that occupancy
531 diversity should be considered while network training. Furthermore, investigating the correlation
532 between energy use and occupancy status is an essential prerequisite to improve the prediction
533 accuracy.

534 Interestingly, in respect to training time, a distinct tendency is not observed regardless of
535 the number of similar daily datasets. These results do not concur with the previous opinions that
536 the computation process with a large data takes a longer time [20,25,28]. This gap in the literature
537 can be caused by the fact that as the amount of historical data increases, there are more
538 opportunities to use an inappropriate set of historical data for network training. As a consequence,
539 it would be difficult to update the gradient of the network performance in successive iterations, so
540 the network training will be terminated earlier than suggested in previous studies.

541 From the aforementioned findings, it can be seen that with the aid of highly granular
542 occupancy data energy prediction can be further improved. The ability to better recognize daily
543 peak demand and daily energy consumption during energy use prediction provides building
544 operators with a better guideline as to how to schedule the operation of HVAC systems at the
545 building level more efficiently (e.g., buildings' standard hours of air conditioning supply [6]).
546 Further, with the proposed prediction model, it is possible to obtain more detailed information
547 about energy use patterns (e.g., load shape, the amount of energy use) for EUG. Although several
548 studies attempted to predict energy consumption at the level of aggregation (e.g., floor level [2],
549 equipment [15]), identifying energy consumption for similar end-users has been limited due to the
550 previously used aggregation methods of occupancy data. Therefore, these improvements will allow
551 facility managers to personalize daily operations of HVAC systems depending on EUG to achieve

552 energy saving without compromising occupants' thermal comfort. For example, if various EUGs
553 exist in a smart metered building with a building energy management systems (BEMS), scheduling
554 different setback periods can be effective for occupants to maintain their desired levels of thermal
555 comfort after other occupants arrive. On the other hand, an issue might arise about the usability of
556 the proposed prediction model because not all buildings yet have smart metering or BEMS.
557 However, numerous efforts are under way to adopt such technologies in the field of building energy
558 use prediction and continue to become more widespread [2,10,11,26]. Further, considering that
559 demand response programs are an essential part in real-time building energy management, weather
560 forecasting data should be automatically imported into the BEMS because weather conditions
561 significantly vary occasionally [44]. In most countries, the forecasting data is provided by national
562 meteorological administrations and are available at the different temporal granularity levels (e.g.,
563 hourly, daily) using an open application programming interfaces (United States [45], South Korea
564 [46]). Therefore, the proposed model can extend its practical application for demand response
565 programs in an automated way.

566 Lastly, while these findings represent an advancement in the state-of-the-art of building
567 energy prediction modeling, this work is not without limitation and numerous avenues for future
568 work remain. For the developed model, additional efforts could include using additional types of
569 ANN since this affects prediction performance. Further, in this work, no attempt is made to
570 investigate the optimal network type for building energy use prediction; limitations in the
571 prediction accuracy thus remain. In order to address these issues, recurrent neural networks that
572 are trained using the current input variables as well as the previous input and output variables
573 should be considered since building energy use has sequence-dependent features. Moreover, it is
574 suspected that improved performance of building energy use prediction will be achieved by

575 incorporating in an occupancy prediction model. This incorporation has already been emphasized
576 in several studies, since contextual variables affect occupancy status as well as energy consumption
577 [47-49]. However, since this research uses the average values of occupancy rate observed in similar
578 daily datasets, there can be a discrepancy between the actual and predicted energy use. In order to
579 overcome this limitation, the future prediction models should be trained with the following steps
580 in mind. First, network training should be performed to predict occupancy status using its
581 significant determinants. Then, based on the predicted occupancy data, the future models should
582 be trained to predict building energy use.
583
584 **7. Conclusions**
585 With the increasing concern about energy saving in buildings, accurate predictions of
586 energy consumption are essential to optimize the operation of energy-consuming equipment during
587 a buildings operation. To date, substantial efforts have been undertaken to improve prediction
588 accuracy, specifically while focusing on occupants' presence in buildings. Unfortunately, despite
589 the recent advancements in prediction accuracy, two significant obstacles remain when predicting
590 energy consumption using occupancy data. First, occupancy diversity among EUGs has rarely
591 been considered during model development. Second, occupancy's correlation with energy
592 consumption may be weak at time due to variances in occupant behavior. Therefore, this research
593 investigated the effect of occupancy-related characteristics of EUGs on the prediction performance.
594 In order to achieve this objective, comparative experiments were conducted using a data
595 mining-based prediction model. The experiments produced two key findings. First, occupancy-
596 related characteristics of EUGs significantly contribute to the prediction performance. Across all
597 experiments, the GL-2 had the highest prediction accuracy, but took a longer time for its network

598 training. Second, the proposed prediction model provides acceptable prediction accuracy using a
599 minimal amount of historical data. All the prediction results were within the acceptable tolerance
600 range (CV-RMSE of 30%). In particular, the GL-2 produced higher accuracy when the network
601 training is performed using less than 10 similar daily datasets.

602 This research contributes to the literature by enhancing our knowledge of how occupancy-
603 related characteristics of EUGs affect energy use prediction performance. In addition, this research
604 develops a data mining-based prediction model that facilitates the recognition of the amount of
605 energy being consumed by EUGs. With this information, facility managers can personalize the
606 operation of energy-consuming equipment based on EUG. Further, this research is significant
607 because the developed model provides practical solutions to achieve acceptable prediction
608 accuracy using minimal historical data. Future research efforts should explore the following
609 avenues. First, exploring whether or not prediction performance can be improved using alternate
610 types of ANN (e.g., recurrent neural networks). Second, the proposed prediction model should be
611 incorporated with occupancy prediction models to provide more accurate information about
612 building energy consumption.

613

614 **Acknowledgement**

615 The authors wish to acknowledge financial support by Ministry of Land, Infrastructure
616 and Transport of Korean government from Infrastructure and Transportation Technology
617 Promotion Program (15CTAP-B080352-02), the Integrated Research Institute of Construction and
618 Environmental Engineering at Seoul National University research which is funded by the South
619 Korean Ministry of Education & Human Resources Development, and a National Science
620 Foundation Award (No. CBET-1705273).

621

622 ## References

623 [1] B. Metz, O.R. Davidson, P.R. Bosch, R. Dave, L.A. Meyer, Contribution of working group III to the
624 fourth assessment report of the intergovernmental panel on climate change, Cambridge University
625 Press, United Kingdom and New York, 2007.

626 [2] R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor, Forecasting energy consumption of multi-family
627 residential buildings using support vector regression: Investigating the impact of temporal and spatial
628 monitoring granularity on performance accuracy, Appl. Energy 123 (2014) 168-178.

629 [3] D. Monfet, M. Corsi, D. Choinière, E. Arkhipova, Development of an energy prediction tool for
630 commercial buildings using case-based reasoning, Energy Build. 81 (2014) 152-160.

631 [4] J. Yang, H. Rivard, R. Zmeureanu, On-line building energy prediction using adaptive artificial neural
632 networks, Energy Build. 37 (2005) 1250-1259.

633 [5] R.Ž. Jovanović, A.A. Sretenović, B.D. Živković, Ensemble of various neural networks for prediction
634 of heating energy consumption, Energy Build. 94 (2015) 189-199.

635 [6] S.S.K. Kwok, R.K.K. Yuen, E.W.M. Lee, An intelligent approach to assessing the effect of building
636 occupancy on building cooling load prediction, Build. Environ. 46 (2011) 1681-1690.

637 [7] A.H. Neto, F.A.S. Fiorelli, Comparison between detailed model simulation and artificial neural
638 network for forecasting building energy consumption, Energy Build. 40 (2008) :2169-2176.

639 [8] A. Yezioro, B. Dong, F. Leite, An applied artificial intelligence approach towards assessing building
640 performance simulation tools, Energy Build. 40 (2008) 612-620.

641 [9] K. Li, C. Hu, G. Liu, W. Xue, Building's electricity consumption prediction using optimized artificial
642 neural networks and principal component analysis, Energy Build. 108 (2015) 106-113.

643 [10] C. Sandels, J. Widén, L. Nordström, E. Andersson, Day-ahead predictions of electricity consumption
644 in a Swedish office building from weather, occupancy, and temporal data, Energy Build. 108 (2015)
645 279-290.

646 [11] J. Virote, R. Neves-Silva, Stochastic models for building energy prediction based on occupant
647 behavior assessment, Energy Build. 53 (2012) 183-193.
648 [12] Z. Wang, Y. Ding, An occupant-based energy consumption prediction model for office equipment,
649 Energy Build. 109 (2015) 12-22.
650 [13] X. Liang, T. Hong, G.Q. Shen, Improving the accuracy of energy baseline models for commercial
651 buildings with occupancy data, Appl. Energy 179 (2016) 247-260.
652 [14] M.S. Gul, S. Patidar, Understanding the energy consumption and occupancy of a multi-purpose
653 academic building, Energy Build. 87 (2015) 155-165.
654 [15] G. Escrivá-Escrivá, C. Álvarez-Bel, C. Roldán-Blay, M. Alcázar-Ortega, New artificial neural
655 network prediction method for electrical consumption forecasting based on building end-uses,
656 Energy Build. 43 (2011) 3112-3119.
657 [16] K. Anderson, K. Song, S. Lee, H. Lee, M. Park, Energy consumption in households while unoccupied:
658 Evidence from dormitories, Energy Build. 87 (2015) 335-341.
659 [17] N. Brown, A.J. Wright, A. Shukla, G. Stuart, Longitudinal analysis of energy metering data from
660 non-domestic buildings, Build. Res. Inf. 38 (2010) 80-91.
661 [18] O.T. Masoso, L.J. Grobler, The dark side of occupants' behaviour on building energy use, Energy
662 Build. 42 (2010) 173-177.
663 [19] M.S. Gul, S. Patidar, Understanding the energy consumption and occupancy of a multi-purpose
664 academic building, Energy Build. 87 (2015) 155-165
665 [20] G. Mustafaraj, G. Lowry, J. Chen, Prediction of room temperature and relative humidity by
666 autoregressive linear and nonlinear neural network models for an open office, Energy Build. 43 (2011)
667 1452-1460.
668 [21] G. Chicco, Overview and performance assessment of the clustering methods for electrical load
669 pattern grouping, Energy 42 (2012) 68-80.

[22] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Autom. Constr. 49 (2015) 1-17.

[23] H. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (2012) 3586-3592.

[24] Y. Sun, S. Wang, F. Xiao, Development and validation of a simplified online cooling load prediction strategy for a super high-rise building in Hong Kong, Energy Convers. Manage. 68 (2013) 20-27.

[25] K. Yun, R. Luck, P.J. Mago, H. Cho, Building hourly thermal load prediction using an indexed ARX model, Energy Build. 54 (2012) 225-233.

[26] R. Sevlian, R. Rajagopal, Short Term Electricity Load Forecasting on Varying Levels of Aggregation, 2014 arXiv:1404.0058.

[27] P.A. González, J.M. Zamarreño, Prediction of hourly energy consumption in buildings based on a feedback artificial neural network, Energy Build. 37 (2005) 595-601.

[28] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Appl. Energy 127 (2014) 1-10.

[29] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, Waltham, 2012.

[30] T.M. Cover and P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory, 13 (1967) 21-27.

[31] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1979) 224-227.

[32] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: Proceedings of the 2010 IEEE International Conference on Data Mining, IEEE Computer Society, 2010, 911–916.

[33] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy 42 (2012) 68-80.

695 [34] R.J. Schalkoff, Artificial Neural Networks, McGraw-Hill, NewYork, 1997.
696 [35] M.A. El-Sharkawi, R.J. Marks, Artificial neural networks as operator aid for on-line static security
697 assessment of power systems, 10th PSCC, 1990, 895-901.
698 [36] Y.S. Kim, Toward a successful CRM: variable selection, sampling, and ensemble, Decis. Support
699 Syst. 41 (2006) 542-553.
700 [37] R. May, G. Dandy, H. Maier, Review of input variable selection methods for artificial neural
701 networks, Artif. Neural Networks-Methodol. Adv. Biomed. (2011) Appl., 19-44.
702 [38] O.G. Santin, L. Itard, H. Visscher, The effect of occupancy and building characteristics on energy
703 use for space and water heating in Dutch residential stock, Energy Build. 41 (2009) 1223-1232.
704 [39] J. Chena, X. Wang, K. Steemers, A statistical analysis of a residential energy consumption survey
705 study in Hangzhou, Chia, Energy Build. 66 (2013) 193-202.
706 [40] G.Y. Yun, K. Steemers, Behavioural, physical and socio-economic factors in household cooling
707 energy consumption, Appl. Energy 88 (2011) 2191-2200.
708 [41] K. Suomalainen, G. Pritchard, B. Sharp, Z. Yuan, G. Zakeri, Correlation analysis on wind and hydro
709 resources with electricity demand and prices in New Zealand, Appl. Energy 137 (2015) 445-462.
710 [42] G. Panchal, A. Ganatra, Y.P. Kosta, D. Panchal, Behaviour analysis of multilayer perceptrons with
711 multiple hidden neurons and hidden layers, Int. J. Comput. Theory Eng. 3 (2011) 332-337.
712 [43] ASHRAE, ASHRAE guideline 14-2002, American Society of Heating, Refrigerating, and Air-
713 Conditioning Engineers Inc., Atlanta, 2002.
714 [44] W. Shengwei, X. Xue, Y. Chengchu, Building power demand response methods toward smart grid,
715 HVAC&R Res. 20 (2014) 665-687.
716 [45] National Oceanic and Atmospheric Administration (NOAA), NDFD SOAF Service,
717 https://graphical.weather.gov/xml/ (accessed 02.08.17).
718 [46] Korea Meteorological Administration (KMA), Weather Information,
719 http://www.kma.go.kr/weather/main.jsp, 2017 (accessed 02.08.17).

720 [47] J. Page, D. Robinson, N. Morel, J.L. Scartezzini, A generalised stochastic model for the simulation
721 of occupant presence, Energy Build. 40 (2008) 83-98.
722 [48] E. McKenna, M. Krawczynski, M. Thomson, Four-state domestic building occupancy model for
723 energy demand simulations, Energy Build. 96 (2015) 30-39.
724 [49] C.M. Stoppel, F. Leite, Integrating probabilistic methods for describing occupant presence with
725 building energy simulation models, Energy Build. 68 (2014) 99-107.
726

727 **Table 1.** Overview of case buildings

Building characteristics	Case buildings
Construction year	2010
Building function	Residential
Number of floors	7 or 8
Room size	250 single rooms, 1125 double rooms
Occupants	Graduate and undergraduate student
Heating system	Radiant floor heating with individual control, electric
Available periods for heating	January to March, October to December

728

729 **Table 2.** Input variables for building energy use prediction

Variables	Unit	Value
Outside dry-bulb temperature (OT)	℃	Continual
Wind speed (WS)	m/s	Continual
Relative humidity (RH)	%	Continual
Solar radiation (SR)	MJ/㎡	Continual
Occupancy rate (OR)	-	Continual
Day of the week (DW)	day	Categorical: Weekday(0), Weekend(1)
Course period (CP)	-	Categorical: Fall(0), Winter(1)

730

731

Table 3. Description of different prediction models

Model	Occupancy-related Characteristics	
	Occupancy Diversity	Correlation Effect
BL-1		
BL-2		√
GL-1	√	
GL-2	√	√

732

733 **Table 4.** Main parameters used for network training

Parameter	Value
Number of similar daily datasets	10
Clustering algorithm	*k*-means algorithm
Network type	Feed forward neural network
Number of hidden layers	1
Number of nodes in hidden layer	10
Number of epochs	500
Minimum gradient of performance	1e-07
Maximum number of validation checks	50

734

735 **Table 5.** Average values of input variables for next day building energy user prediction

Test Dataset	Input Variable						
	OT	WS	RH	SR	OR	DW	CY
TD1	13.72	2.36	61.38	0.55	0.6	1	1
TD2	4.06	2	41.04	0.43	0.6	2	1
TD3	-5.03	2.23	64.38	0.29	0.43	1	2
TD4	-2.61	1.84	37.5	0.42	0.44	1	2
TD5	-0.07	2.93	58.08	0.43	0.43	1	2

736

737 **Table 6.** Correlation coefficient between energy use and occupancy status for next day building
738 energy user prediction

Test Dataset	Group Level		Building Level
	EUG 1	EUG 2	
TD1	0.7271	0.4124	0.6047
TD2	0.6592	0.3039	0.4198
TD3	0.8171	0.4718	0.7299
TD4	0.8230	0.4377	0.7482
TD5	0.8151	0.4037	0.7543

739

740 **Table 7.** Average values of input variables for next five days building energy user prediction

Test Dataset		Input Variable						
		OT	WS	RH	SR	OR	DW	CY
TD1	TD1-1	6.27	2.5	77.08	0.26	0.56	1	1
	TD1-2	5.02	2.17	55.08	0.46	0.56	1	1
	TD1-3	5.5	1.8	55	0.42	0.55	1	1
	TD1-4	7.74	1.79	61.67	0.38	0.57	1	1
	TD1-5	10.82	2.84	62.58	0.31	0.55	1	1
TD2	TD2-1	-6.55	3.54	44.42	0.42	0.56	1	1
	TD2-2	-5.58	2.83	49.08	0.4	0.59	2	1
	TD2-3	-3.84	1.6	43.88	0.23	0.59	2	1
	TD2-4	-1.34	2.4	48.38	0.39	0.57	1	1
	TD2-5	-2.6	1.61	45.54	0.39	0.55	1	1
TD3	TD3-1	3.54	1.69	62.21	0.35	0.5	2	2
	TD3-2	2.8	3.51	75.92	0.09	0.45	2	2
	TD3-3	3.88	2.7	89.5	0.13	0.42	1	2
	TD3-4	-2.9	3.85	57.04	0.49	0.4	1	2
	TD3-5	-4.3	2.18	48.88	0.46	0.43	1	2

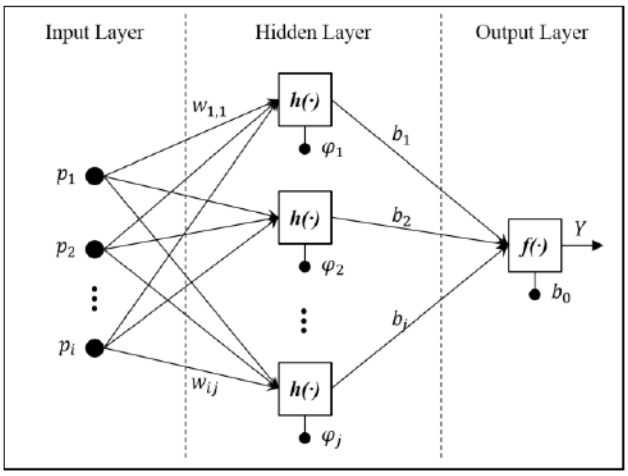
741

742 **Table 8.** Correlation coefficient between energy use and occupancy status for next five days
743 building energy user prediction

Test Dataset		Group Level				Building Level
		EUG 1	EUG 2	EUG 3	EUG 4	
TD1	TD1-1	0.7131	0.6579	0.1794	-	0.5215
	TD1-2	0.7765	0.2863	-	-	0.6153
	TD1-3	0.7569	0.4563	-	-	0.6017
	TD1-4	0.7863	0.3333	-	-	0.7048
	TD1-5	0.7737	0.5286	-	-	0.7141
TD2	TD2-1	0.8153	0.3582	-	-	0.7637
	TD2-2	0.6774	0.5294	0.1776	-	0.2549
	TD2-3	0.7831	0.4367	0.0966	-	0.2493
	TD2-4	0.8019	0.2826	-	-	0.6264
	TD2-5	0.7780	0.4874	-	-	0.6195
TD3	TD3-1	0.8459	0.3578	-	-	0.7033
	TD3-2	0.8913	0.7772	0.3970	0.3541	0.8403
	TD3-3	0.8149	0.8046	0.7792	0.4615	0.6911
	TD3-4	0.8324	0.8224	0.3558	-	0.7835
	TD3-5	0.8338	0.8227	0.3227	-	0.7094

744

745



746

747 **Fig. 1.** Structure of a three-layer feed forward neural network

748

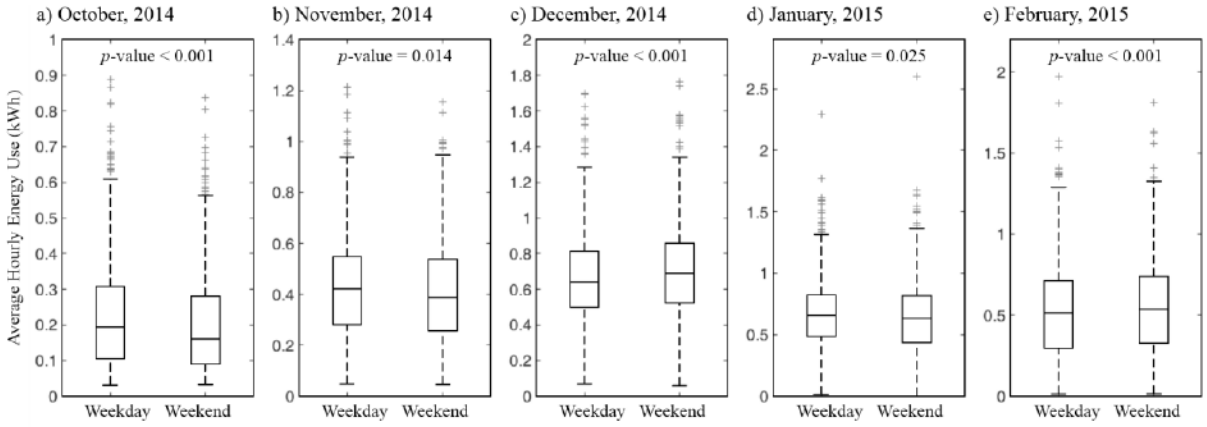
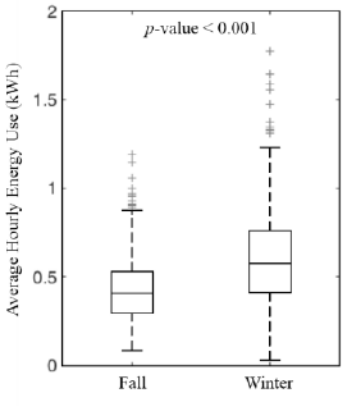


Fig. 2. Average hourly energy use by day of the week



753

754 **Fig. 3.** Average hourly energy use by course periods

755

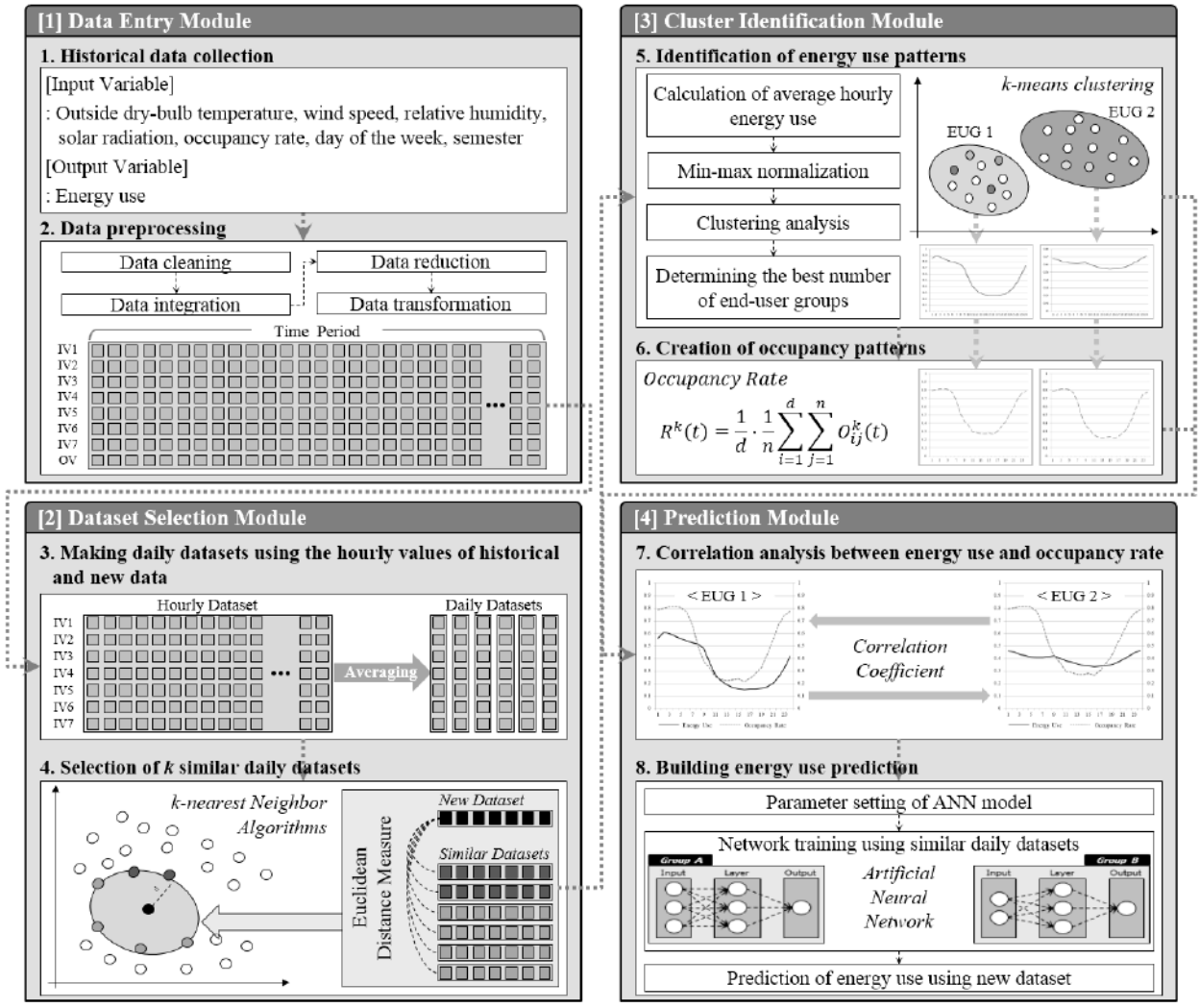
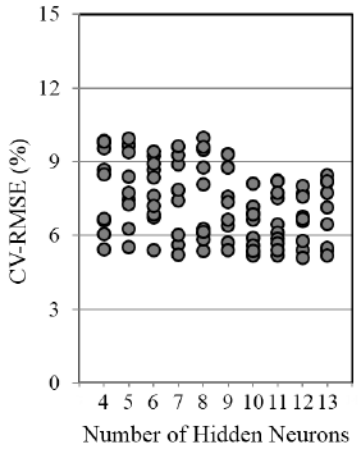


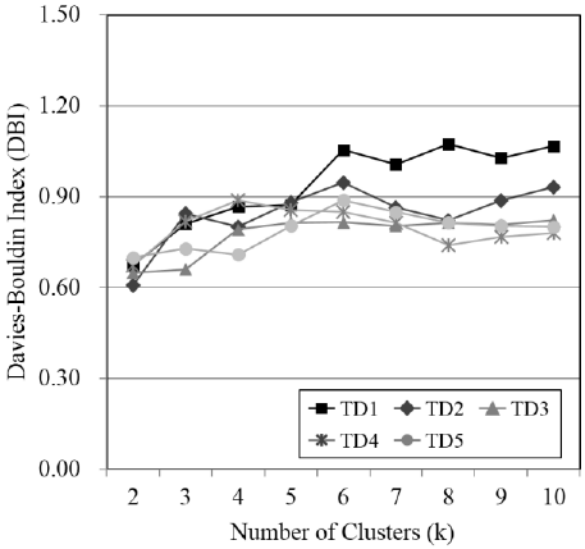
Fig. 4. Main structure of data mining-based energy use prediction model



759

760 **Fig. 5.** CV-RMSE by number of hidden neurons

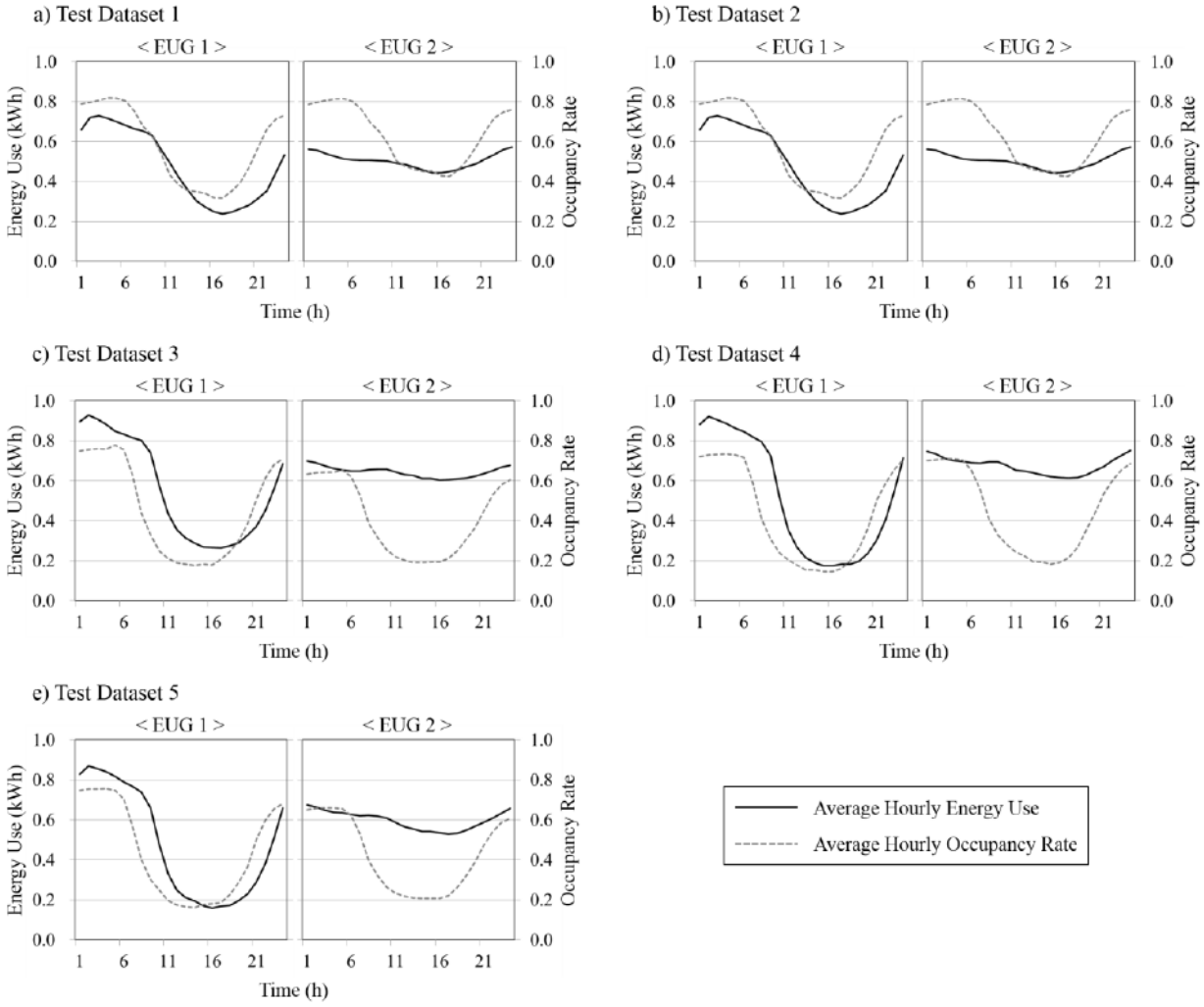
761



762

763 **Fig. 6.** Davies-Bouldin index by number of clusters for next day building energy use prediction

764



765

766 **Fig. 7.** Daily profiles of energy use and occupancy status for next day building energy use

767 prediction

768

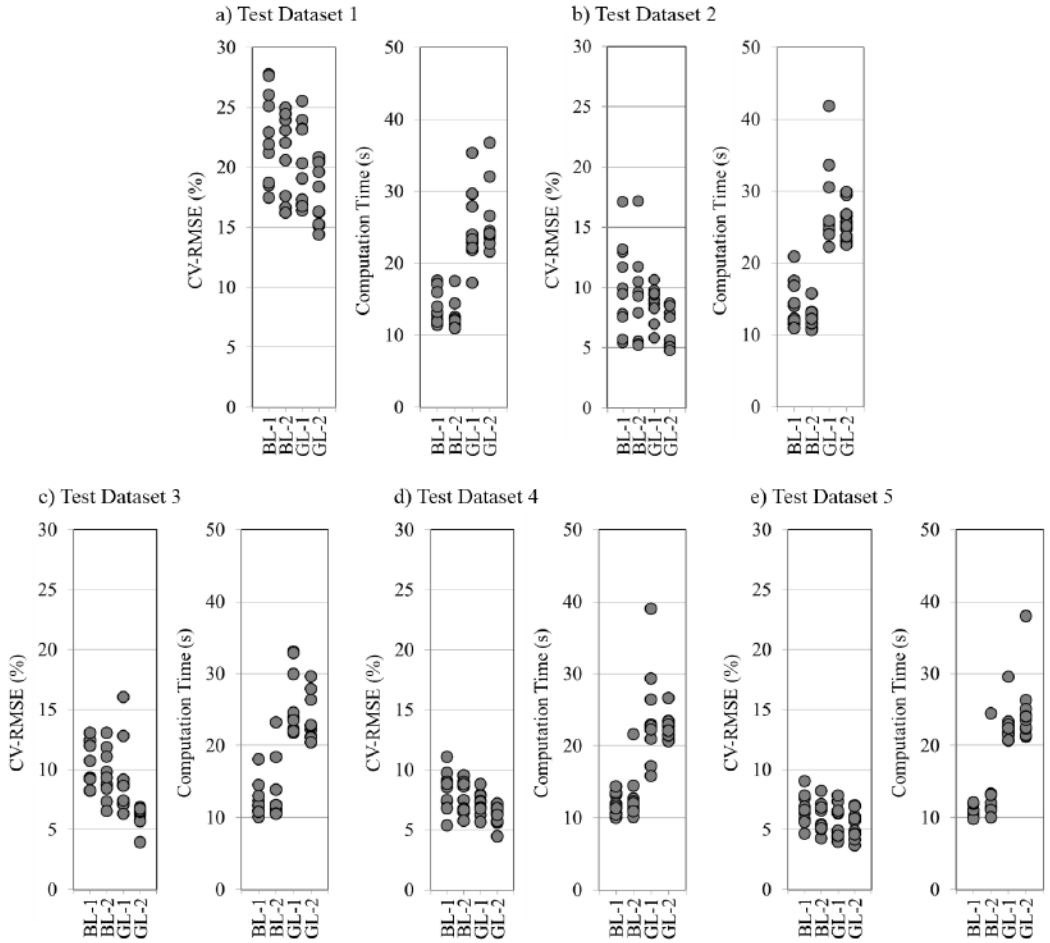
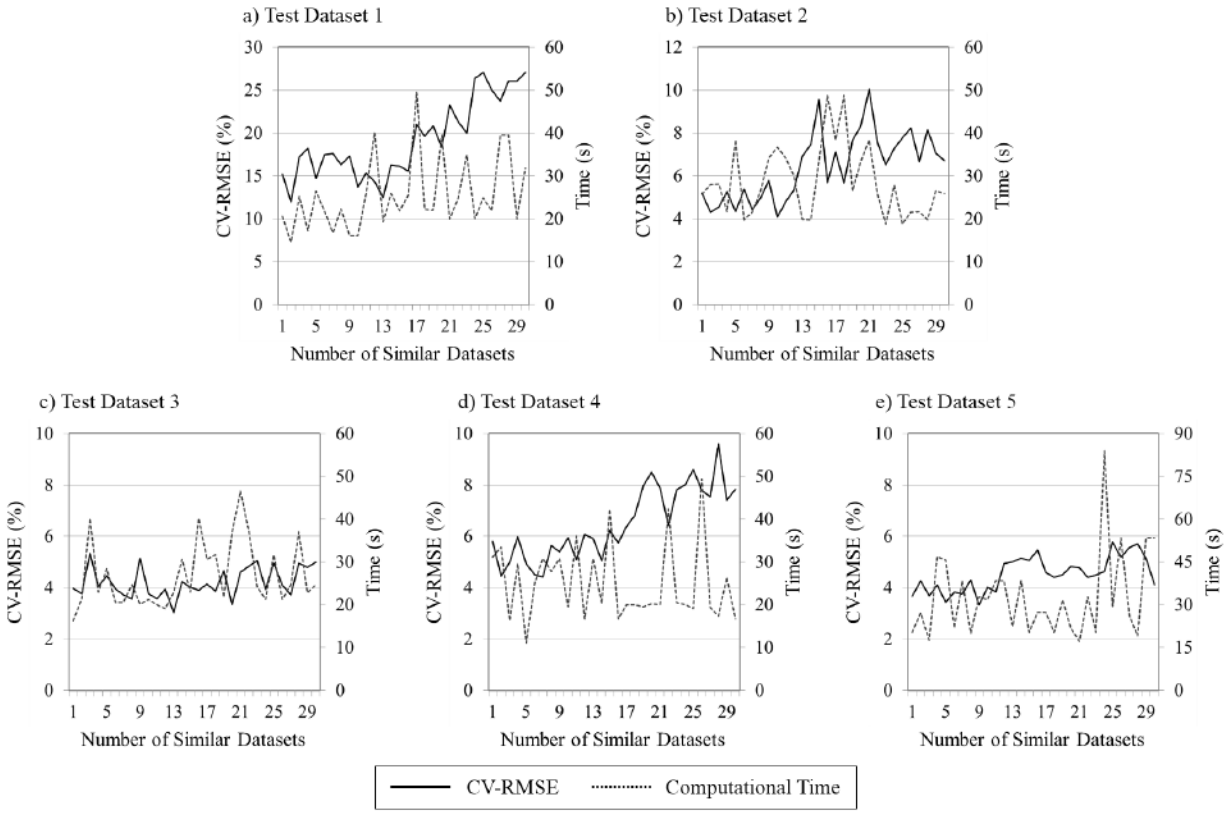


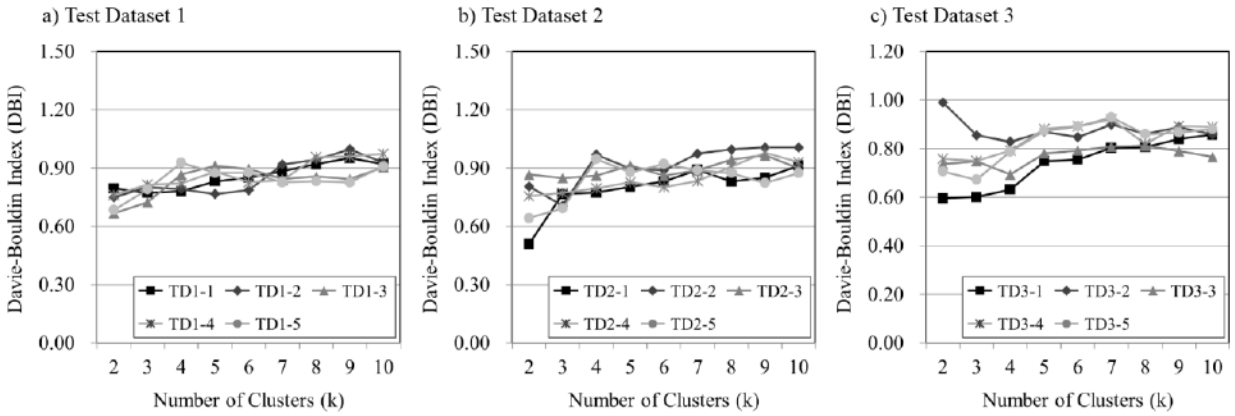
Fig. 8. CV-RMSE and computation time for next day building energy use prediction



772

773 **Fig. 9.** Prediction performance for the next day by number of similar daily datasets

774

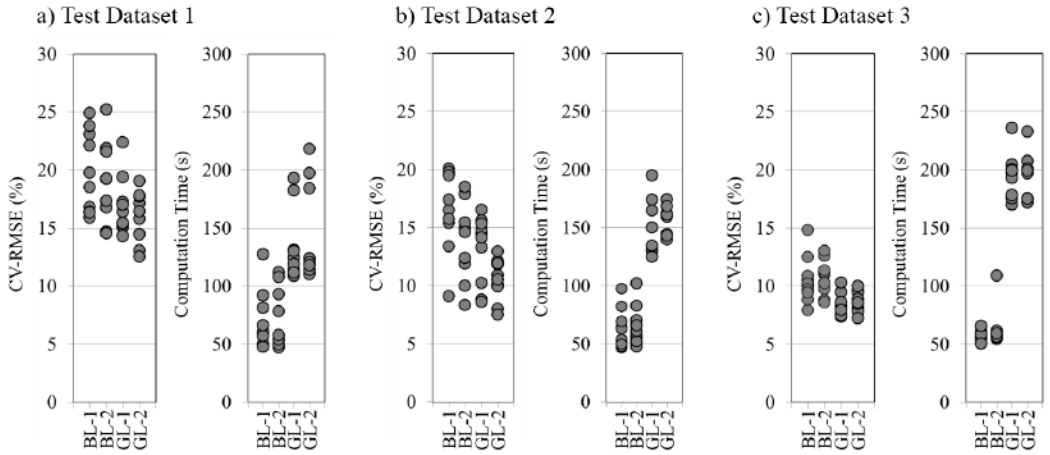


775

776 **Fig. 10.** Davies-Bouldin index by number of clusters for next five days building energy use

777 prediction

778



779

780 **Fig. 11.** CV-RMSE and computation time for next five days building energy use prediction

781

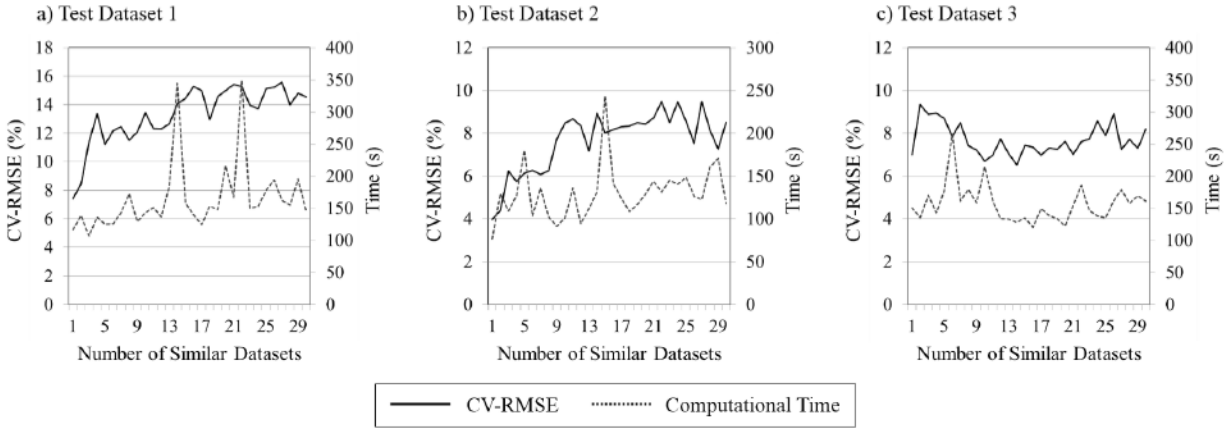


Fig. 12. Prediction performance for the next five days by number of similar daily datasets