energies

MDPI

Article

Combine Clustering and Machine Learning for Enhancing the Efficiency of Energy Baseline of Chiller System

Chun-Wei Chen, Chun-Chang Li * and Chen-Yu Lin

Intelligent Machining Division, Taiwan Instrument Research Institute, NARL, Hsinchu City 300, Taiwan; rich@narlabs.org.tw (C.-W.C.); chenyulin@narlabs.org.tw (C.-Y.L.)

* Correspondence: 1709873@narlabs.org.tw; Tel.: +886-3-5779911 (ext. 171)

Received: 13 July 2020; Accepted: 20 August 2020; Published: 24 August 2020

Abstract: Energy baseline is an important method for measuring the energy-saving benefits of chiller system, and the benefits can be calculated by comparing prediction models and actual results. Currently, machine learning is often adopted as a prediction model for energy baselines. Common models include regression, ensemble learning, and deep learning models. In this study, we first reviewed several machine learning algorithms, which were used to establish prediction models. Then, the concept of clustering to preprocess chiller data was adopted. Data mining, K-means clustering, and gap statistic were used to successfully identify the critical variables to cluster chiller modes. Applying these key variables effectively enhanced the quality of the chiller data, and combining the clustering results and the machine learning model effectively improved the prediction accuracy of the model and the reliability of the energy baselines.

Keywords: energy baselines; machine learning; clustering

1. Introduction

With the popularity of sustainable development concepts, an increasing number of enterprises are adopting energy conservation and carbon reduction as a significant aspect of corporate development. In most current enterprises, air-conditioning systems are the most energy-intensive equipment. Subsequently, chiller system are the most energy-intensive subsystems in air-conditioning systems. Therefore, improving the energy efficiency of chiller system can significantly reduce the energy consumption of entire systems.

Once the energy efficiency of chiller system is improved, our next focus is the effectiveness and benefits of the improvement methods. In this stage, accurately assessing the energy efficiency of improvement methods becomes a critical topic. Currently, the most widely used method is the establishment of energy baselines. An energy baseline refers to the collection of data within a time period before equipment improvement. The collected data can then be used to establish the mathematical equations that can describe the operation modes of equipment. This process is known as baseline modeling. Then, data are collected within a time period after equipment improvement to determine the prediction values of the post-improvement data in the baseline model. Finally, energy efficiency can be calculated by comparing the prediction values and post-improvement data.

Because energy baselines are an essential approach for assessing the improvement performance of chiller system, many studies have focused on developing chiller prediction models. The models can be predominantly classified into semi-empirical models and empirical models. Semi-empirical models refer to the use of equations derived from relevant laws of physics to describe performance of chiller system. For example, Lee and Reddy developed regression models to predict the coefficient of

performance (COP) of screw chillers and centrifugal chillers [1,2]. Empirical models are data-oriented models. Equations that describe chiller performance can be established without having to collected chiller-related system data. For example, Adnan et al. combined artificial neural network (ANN) models of different structures and used three variables, specifically refrigeration ton, inlet temperature, and outlet temperature, to create a chiller prediction model [3]. Kim et al. used different combinations of input variables to identify the ANN model with the highest prediction accuracy [4]. Yu et al. used random forest model to predict the operating parameters that maximize chiller COP under different working conditions [5,6].

The development of prediction models can effectively enhance the accuracy of energy baseline predictions. Nonetheless, chiller system are intricate pieces of equipment. Many operating parameters must be collected, and operating modes may vary depending on the setting. Appropriately preprocessing data can facilitate overall analysis efficiency. Clustering is an excellent data preprocessing approach. It functions by calculating the relationships between data points and identifying hidden data structures. Malinao et al. applied the X-means clustering method to cluster chiller system and identify different operating modes [7]. Habib et al. used a two-layer K-means algorithm to cluster chiller system and identify and remove outliers to enhance energy analysis efficiency [8]. Habib et al. combined K-means, BoWR, and hierarchical clustering to preprocess chiller data. The researchers proposed a model to automatically detect the energy systems of different constructs. The model can be used for fault detection and diagnosis [9].

The operating modes in different conditions can be identified by clustering chiller data. This process enhances data quality and usability, thereby improving analysis efficiency. However, existing studies mostly used clustering for fault detection and diagnosis and rarely used preprocessed data in the development of prediction models. Therefore, using the COP of chiller system as the target of research, we applied a clustering method to preprocess chiller data and identify the operating modes of chiller system in different settings. In addition, a machine learning method was used to create prediction models for various operating modes.

The contribution of this paper is the proposal of a methodology for improving the prediction accuracy of chiller system. The chiller system examined in this study was a 230RT air-conditioning chiller equipped with a variable-frequency, centrifugal compressor. The methodology first selected K-means as clustering method based on characteristics of data. Then, we used data mining and statistical techniques to identify the critical variables for clustering method. After successfully identifying the critical variables, we applied K-means clustering and gap statistic to cluster chiller modes. For finding the best prediction accuracy of chiller system, the optimal number of clusters was calibrated, if needed. Finally, we combined the clustering results and machine learning models to establish a prediction model of chiller system. The simulation showed that the error rate of prediction model was successfully reduced and the prediction accuracy of chiller energy baselines without excessively increasing computational cost was enhanced.

The structure of this paper is as follows. In Section 2, we introduce commonly used chiller-related prediction models, such as regression models, ANN models, and random forest models. Extreme gradient boosting model is compared, which has gained considerable popularity in recent data analysis competitions. In Section 3, data, modeling, and model assessment criteria are discussed. In Section 4, a prediction simulation on the data is performed and we discuss the results. In Section 5, a conclusion to this study is provided.

2. Review of Machine Learning Algorithm

In this section, we review several machine learning algorithms which were used to establish prediction models of chiller system or related work. Here, we briefly review the final mathematical form of each model, and a detailed formulation is described in Appendix A.

2.1. Regression Model

2.1.1. Lee Simplified Model

Lee combined law of thermodynamics and heat exchanger to develop a prediction model of screw chillers [1]. Equation (1) describes the prediction model of coefficient of performance (COP):

$$\frac{1}{\text{cop}} = -1 + \frac{T_{ci}}{T_{wi}} + \frac{1}{Q_e}\left[-A_0 + A_1 T_{ci} - A_2 \frac{T_{ci}}{T_{wi}}\right] \quad (1)$$

where A_0, A_1, and A_2 are coefficients of model and can be derived by regression analysis (see Appendix A).

2.1.2. Multivariate Polynomial Regression Model

Reddy and Andersen used three variables, specifically cooling capacity, cooling water inlet temperature, chilled water outlet temperature, and their interaction, to create a multivariate regression model of centrifugal chillers [2]. Equation (2) describes the prediction model:

$$\text{COP} = \beta_0 + \beta_1 Q_e + \beta_2 T_{wi} + \beta_3 T_{ci} + \beta_4 {Q_e}^2 + \beta_5 {T_{wi}}^2 + \beta_6 {T_{ci}}^2 + \beta_7 Q_e T_{wi} + \beta_8 Q_e T_{ci} + \beta_9 T_{wi} T_{ci} \quad (2)$$

2.2. Artificial Neural Networks

A basic ANN framework is illustrated in Figure 1. Blue circles mark the neurons. They are responsible for recording values. The arrows illustrate the neural connections and the direction of data transfer. The framework can be broadly categorized into an input layer, hidden layer, and output layer. The hidden layer is responsible for receiving and converting data from the input layer and transferring the converted data to the output layer to derive a solution. The structure of the hidden layer and the data conversion method influence the quality of the overall ANN.

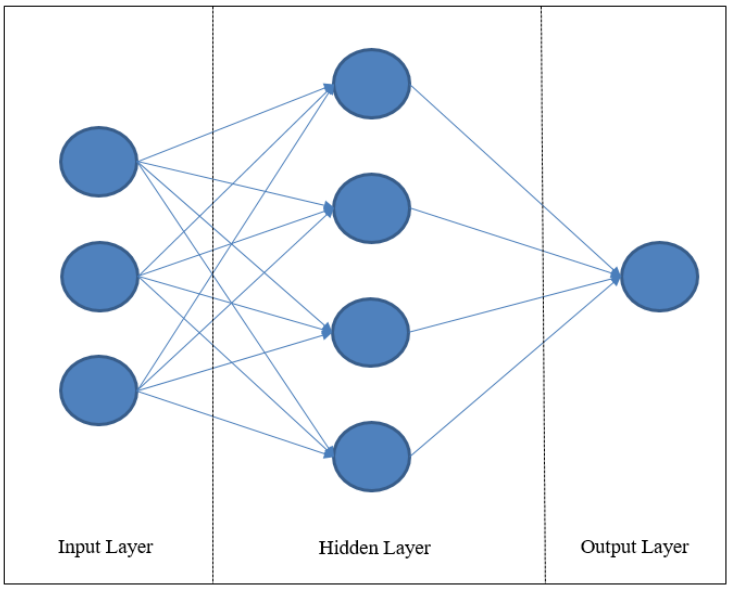

Figure 1. Structure of the artificial neural network (ANN).

Equation (3) describes the relationship of inputs $\boldsymbol{x}_i$ and the j^{th} node in the hidden layer:

$$net_j = \sigma\left(\sum_{i=1}^{n} w_{ij}\boldsymbol{x}_i + b_j\right) \quad (3)$$

where w_{ij} is connection weight, b_j is bias; i is the number of input nodes, and j is the number of hidden nodes. σ is a activation function that transfer the inputs to the hidden layer by way of nonlinear transformation. The widely used activation function are sigmoid function, relu function, and softmax function.

Equation (4) describes the relationship of j^{th} node in the hidden layer and output $\hat{y}_k$:

$$\hat{y}_k = \sigma\left(\sum_j w_{jk} net_j + b_k\right) \tag{4}$$

ANN models can derive the optimal solution for parameters (w, b) by differentiating the loss function, whereby the loss function is expressed as L = loss (y, $\hat{y}$). If the hidden layer comprises more than one sublayer, it may be challenging to derive the optimal solutions for the parameters of the various layers using common differentiation methods. In this instance, the chain rules in calculus can be applied to derive the solutions.

2.3. Ensemble Learning

2.3.1. Random Forest

Random forest is a classic ensemble learning algorithm. Predictions are carried out by combining the results of multiple classifications and regression tree (CART) models. When developing a CART in a random forest, the data and the variables are repeatedly sampled to increase the differences between models and prevent the overfitting problem common to CART models. The form of random forest can be written as:

$$\hat{y}_i = \sum_{m=1}^{M} \sum_{i=1}^{n} \sum_{k=1}^{K} C_k I(\boldsymbol{x}_i \in R_k) \tag{5}$$

where R_k is the kth output space, C_k is average value of R_k, and m is the number of CART in the random forest model.

2.3.2. Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is a popular method used in data analysis competitions recently. It is a strong ensemble learning algorithm improved from gradient boosting decision tree algorithm (GBDT) [10]. In recent years, XGBoost have been actively applied to energy related issues [11–14].

XGBoost combines the results of CART models one by one to establish the prediction model, and uses residual as prediction target. For a given data set with n examples and d features $\mathcal{D} = \left\{(\boldsymbol{x}_i, y_i) \middle| \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1, \ldots, n\right\}$, Equation (6) describes a tree ensemble model using K additive functions to predict the output:

$$\hat{y}_i = \sum_{k=1}^{K} w_{q(\boldsymbol{x}_i)} \tag{6}$$

where $w_{q(\boldsymbol{x}_i)}$ is the CART model. To learn the optimal parameters used in prediction model, Equation (7) describes the regularized objective function Obj:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

where l is a differentiable convex loss function and Ω is the complexity of the model. For a fixed structure $q(x_i)$, the optimal parameter w_j^* and corresponding value Obj^* of output space j can be calculated by

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{8}$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{{G_j}^2}{H_j + \lambda} + \gamma T \tag{9}$$

where G_j and H_j represents the sum of first and second-order gradient statistics in output space j.

2.4. Clustering

Clustering is an unsupervised machine learning method. The purpose of clustering is to analyze the distal relationships of data points and identify underlying data structures, thereby facilitating users in carrying out advanced data analysis. Depending on the nature of the data, clustering approaches can be based on data prototype, class, density, or graphics. Table 1 summarized common clustering algorithm and their applicability from popular machine learning web, scikit-learn (https://scikit-learn.org/stable/modules/clustering.html). This subsection introduces the K-means clustering and gap statistic used in this research.

Table 1. Summary of clustering algorithm.

Method Name	Scalability	Use Case	Geometry
K-means	very large sample, medium clusters	1. general-purpose 2. even cluster size 3. flat geometry 4. not too many clusters	distances between points
Spectral clustering	medium sample, small clusters	1. few clusters 2. even cluster size 3. non-flat geometry	graph distance
Ward hierarchical clustering	large sample, large clusters	1. many clusters, 2. possibly connectivity constraints	distances between points
DBSCAN	very large sample, medium clusters	1. non-flat geometry 2. uneven cluster sizes	distances between nearest points
Birch	large sample large clusters	1. large dataset 2. outlier removal 3. data reduction.	Euclidean distance between points
Mean-shift	not scalable with samples	1. many clusters, 2. uneven cluster size 3. non-flat geometry	distances between points
OPTICS	not scalable	1. non-flat geometry 2. uneven cluster size 3. variable cluster density	distances between points
HDBSCAN	very large sample, medium clusters	1. non-flat geometry 2. uneven cluster sizes	distances between nearest points

2.4.1. K-Means

K-means is a clustering method with relatively simple computational procedures [15]. Although K-means fails to obtain good results in some cases, such as nonspherical, different variance, and different density, it is still popular for its simplicity to implement, known limitations, and excellent

fine-tuning capabilities [16]. Several researchers have proposed different methods to solve different problems [17,18].

K-means can be performed in three steps. First, a k number of cluster centers are randomly established. Then, the Euclidean distance between each sample and the k cluster center is determined, and the sample point is classified into its nearest cluster. Finally, the centers of each cluster are updated using the detailed data until all sample groups reach the shortest distance to the core of their clusters. A detailed formulation is described in Appendix A.

2.4.2. Gap Statistic

The idea of the gap statistic is to compare the total within intracluster variation W_c with its expectation under an appropriate null reference distribution of the data [19]. The estimate of the optimal k is the value for which the total within intracluster variation falls the farthest below this reference curve. Hence, the optimal k is the smallest value k, satisfied in Expression (10):

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1} \tag{10}$$

$\mathrm{Gap}(k)$ and s_{k+1} described in Equations (11) and (12).

$$\mathrm{Gap}(k) = \frac{1}{B}\sum_{b=1}^{B}\log\left(W_{c,b}^{*}\right) - \log(W_c) \tag{11}$$

$$s_k = \sqrt{\frac{1+B}{B}}\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\log\left(W_{c,b}^{*}\right) - \frac{1}{B}\sum_{b=1}^{B}\log\left(W_{c,b}^{*}\right)\right)^2} \tag{12}$$

where B is the number of sampling.

3. Methodology

3.1. Data Description and Statistic

The data examined in this study were from a chiller monitoring system in an undisclosed research center. The system was a 230RT air-conditioning chiller equipped with a variable-frequency, centrifugal compressor. The operating data between April 2018 and May 2019 were collected. Each data point represents one minute. After excluding the idle and maintenance times, a total of 316,749 data points and 28 variables were retained. Using a ratio of 8:2, 253,399 data points were used for training, and 63,350 data points were used for testing. The target of research was the COP of chiller system. The descriptive statistics of the training data and the key variables are illustrated in Figure 2 and tabulated in Table 2.

Table 2. Descriptive statistics of the key variables of chiller system.

Variables	Mean	Standard Deviation	Maximum	Minimum
COP	4.411	0.786	12.329	0.424
Power (kW)	38.139	6.045	127.224	30.001
Load rate (%)	20.76	4.853	62.358	1.825
Flow (GPM)	558.238	89.15	738.068	87.912

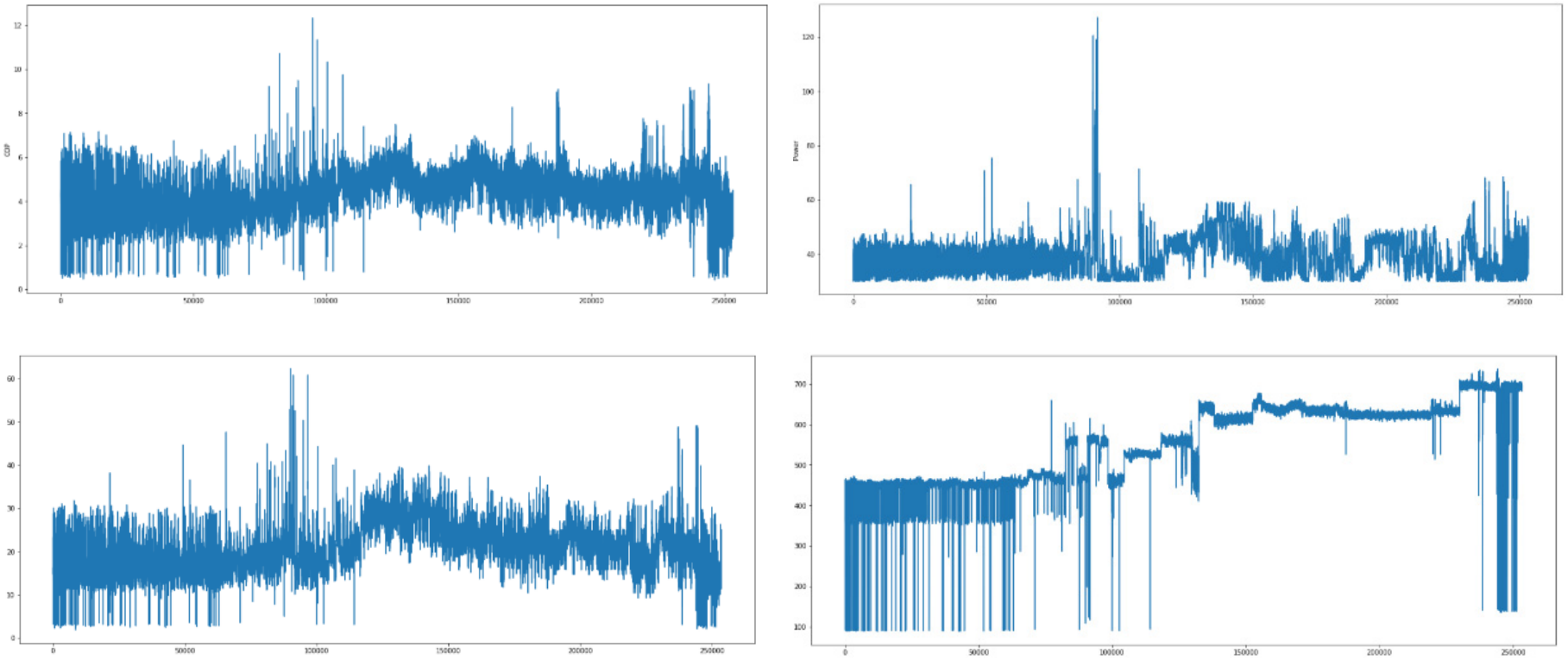

Figure 2. Trend chart of the key variables of chiller system.

The top left image in Figure 2 is a trend chart for COP. The chart shows that the COP values were predominantly distributed between 2 and 6. The top right figure is a trend chart for power consumption. The chart shows that power consumption was significantly higher in specific periods. The full distance of the data approximated 100. The lower left figure is a trend chart for load rate. The distribution was similar to COP. The lower right figure is a trend chart for chilled water flow.

Then, calculate the maximal information coefficient (MIC) for COP. MIC provides a measure of the strength of the linear or nonlinear association between two variables [20]. To ensure a fair comparison, MIC normalized the values and obtained modified values between zero and one. Table 3 tabulated some variables with higher correlation coefficient.

Table 3. Maximal information coefficient of performance (COP).

Variables	kW/RT	Exhaust Temperature (°C)	Load	Supply Cooling Water Temperature Different (°C)	Inhale Temperature (°C)	Chilled Water Flow (GPM)
Maximal information coefficient	0.9634	0.5342	0.4907	0.3255	0.3130	0.2714

Subsequently, a scatter diagram was plotted to observe the distribution relationships between each variable. Scatter relationships of interest are plotted in Figures 3–5. Figure 3 is a scatter diagram of COP and kW/RT. COP and kW/RT presented a reciprocal relationship. The anticipated results were a curve presenting a convex to origin. Instead, five curves and numerous sporadic scatter points were plotted in Figure 3. Therefore, we speculated that other variables influencing the scattering of COP and kW/RT were present.

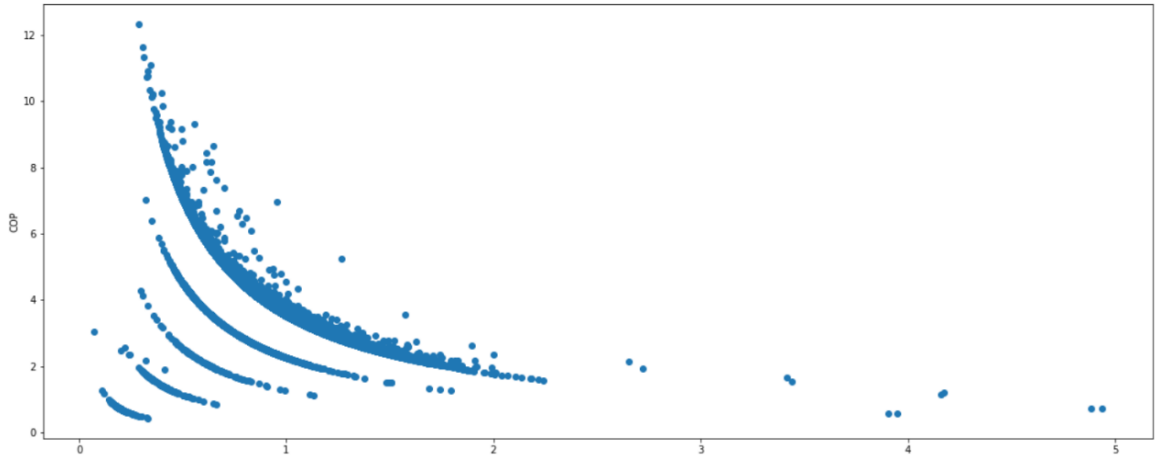

Figure 3. Scatter diagram of the kW/RT and COP.

Figure 4 is a scatter diagram of the condenser flow trend and COP. Figure 4 shows that the data were distributed into six distinct clusters in an apparent manner. Most of the condenser flow trend values ranged between 175 and 200, and the degree of COP dispersion increased concurrently with the condenser flow trend. Figure 5 is a scatter diagram of the chilled water flow and COP. The degree of COP dispersion increased concurrently with the chilled water flow. A block distribution of data points could be vaguely observed.

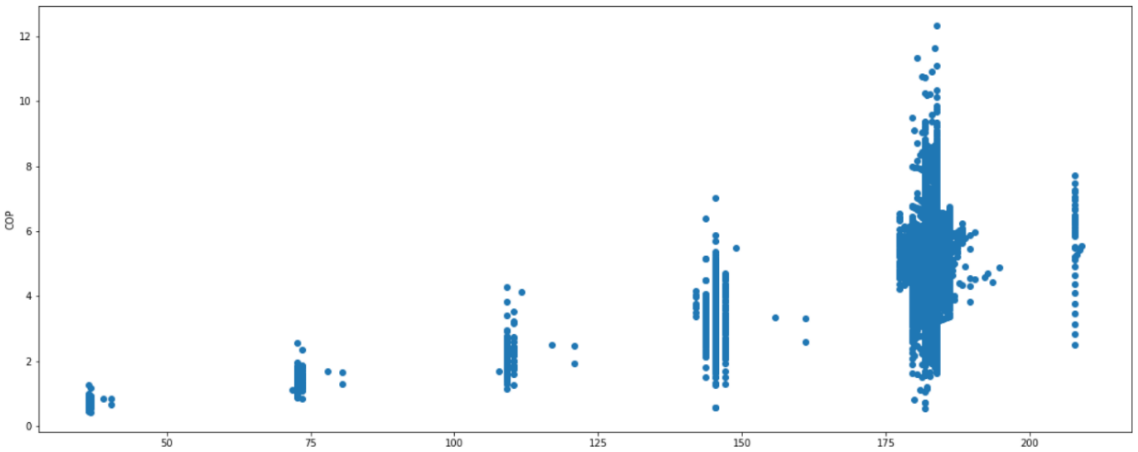

Figure 4. Scatter diagram of the condenser flow trend and COP.

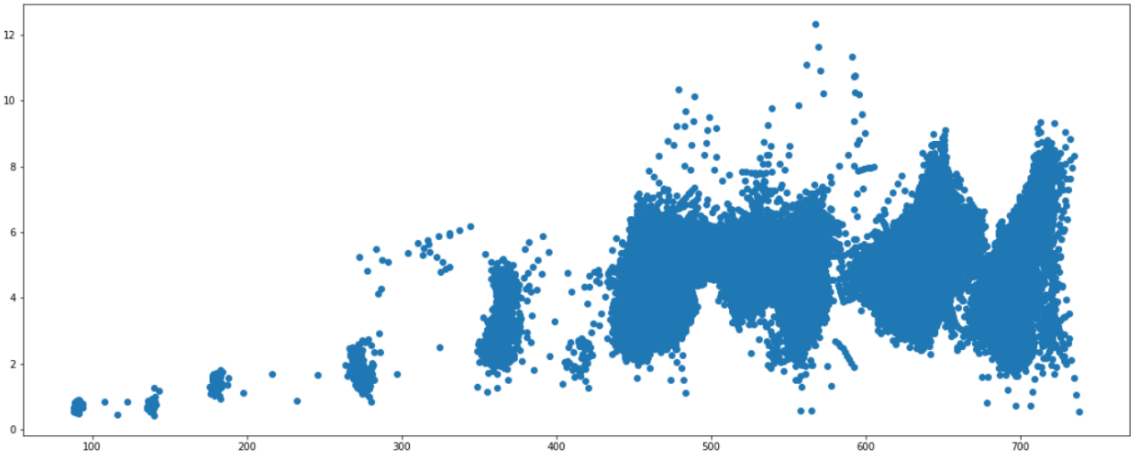

Figure 5. Scatter diagram of the chilled water flow and COP.

3.2. Model

This subsection describes the integration of clustering and machine learning. First, a suitable clustering approach was selected based on the data characteristics. In order to obtain a robust clustering effect, we also recommend using other clustering methods as validation. The necessity of estimating the optimal clustering value k was determined based on the approach. Estimation methods primarily included the elbow method, silhouette coefficient, and gap statistic. Third, the clustering method was employed to cluster the trained data, and the necessity of adjusting the clustering value k was determined by observing the clustering trends. Fourth, the clusters were then incorporated into a chiller prediction model to optimize the parameters and derive the final prediction model. Finally, the results of the different prediction models were compared based on the test data and the model assessment standards. Figure 6 summary the flow chart of establishing prediction model.

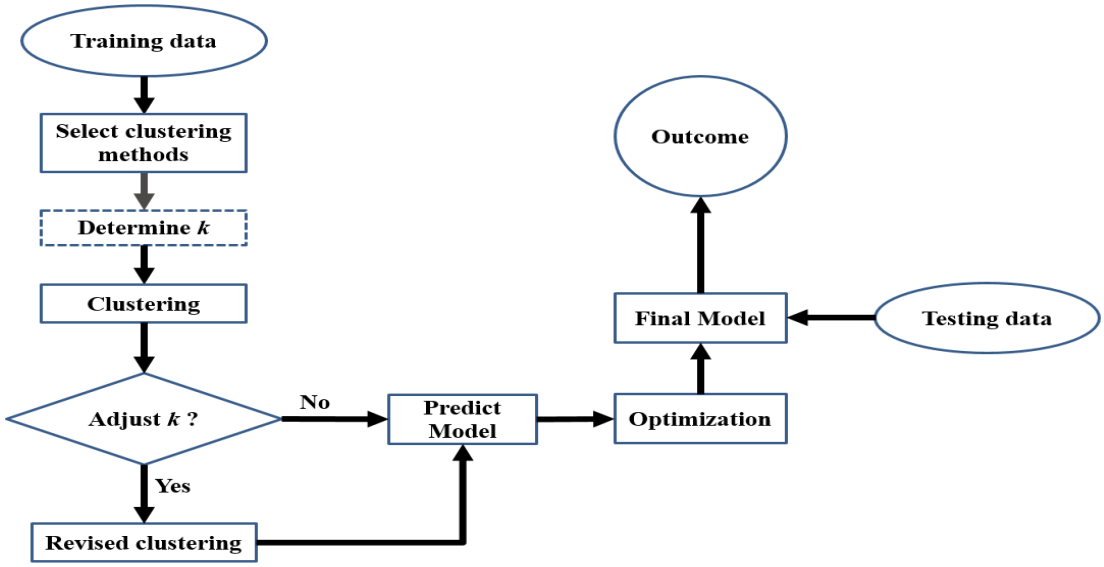

Figure 6. Flowchart of establishing prediction model. The procedure first selected a suitable clustering method according to the training data. Then, we determined the best number of cluster k, if the clustering method was needed. After obtaining the result of clustering, we drew and observed the scatterplot of clustering to determine whether to adjust k or not. Finally, we used the machine learning algorithm to train each cluster to obtain the prediction models, and optimized these models to obtain the final model.

3.3. Evaluation Metrics

To evaluate the performance of the prediction models, three different metrics were used: The MSE (mean square error; Equation (13)), the CVRMSE (coefficient of variation of root-mean squared error; Equation (14)) and the MAPE (mean absolute percentage error; Equation (15)).

$$\text{MSE} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N} \tag{13}$$

$$\text{CVRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}}{\frac{1}{N}\sum_{i=1}^{N} y_i} * 100 \tag{14}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \tag{15}$$

where $\hat{y}_i$ is the predicted value, y_i is the actual value, and N is the total number of data.

MSE intuitively represent the error of predicted value and actual values. CVRMSE gives an indication of the model's ability to predict the overall load shape that is reflected in the data. MAPE provides an overall assessment of the general percent error [21]. In addition to these three metrics, we also took computation speed into account.

4. Discussion

In this chapter, we elucidate whether integrating clustering and machine learning improved the model's predictive accuracy of energy baselines. The aforementioned machine learning model and chiller data were used to train and validate the prediction model. The target of validation was chiller COP, and the variables used in this research were the variables with high MIC values. The simulation environment was Anaconda, the popular data science platform, and the machine learning models were

package from scikit-learn (https://scikit-learn.org/stable/preface.html). The assessment results of the test data are tabulated in Table 4.

Table 4. Evaluation metrics of predict model for COP.

Model	MSE	CVRMSE	MAPE	Time (s)
Linear regression	0.0341	0.0456	0.0351	0.24
Lee simplified model	0.5057	0.1758	0.1608	0.44
Multivariate polynomial regression	0.4263	0.1615	0.1467	0.29
ANN	0.013	0.0282	0.0205	38.6
Random forest	0.003347	0.0143	0.0069	54.6
XGBoost	0.003326	0.0143	0.0075	19.2

In Table 4, the four evaluation metrics, MSE, CVRMSE, MAPE, and Time(s), are calculated. The results indicate that ensemble learning model, random forest, and XGBoost had the better prediction error. The three-error metric of the XGBoost model and random forest model were relatively similar, and the computation speed of XGBoost model was faster than random forest model. Although the evaluation metrics of the three regression models were acceptable, they were less favorable in terms of performance compared to the ensemble learning model, only outperforming the ensemble learning model in computation time. The performance of ANN model was between the regression models and ensemble learning. Then, we assessed whether integrating clustering and machine learning improved the accuracy of the prediction models.

According to the Figure 4, the data were distributed into six distinct clusters in an apparent manner. Although the data seemed a bit uneven, they were well separated from each other. So, we tried to use K-means as the clustering method. Gap statistic is the ideal method for calculating the clustering value *k*. To validate the choice, we also tried to run and compare different clustering methods. The outcome is presented in Appendix B. From the results, K-means was a great choice in this research.

K-means clustering and gap statistic were performed on the 28 variables of the chiller data. The clustering results were then consolidated onto a graph. Based on the calculation results, the condenser flow trend was the most suitable variable of the 28 variables for clustering. Figure 7 is a scatter diagram of the condenser flow trend and COP after clustering. The diagram shows that K-means distributed the data into ten clusters.

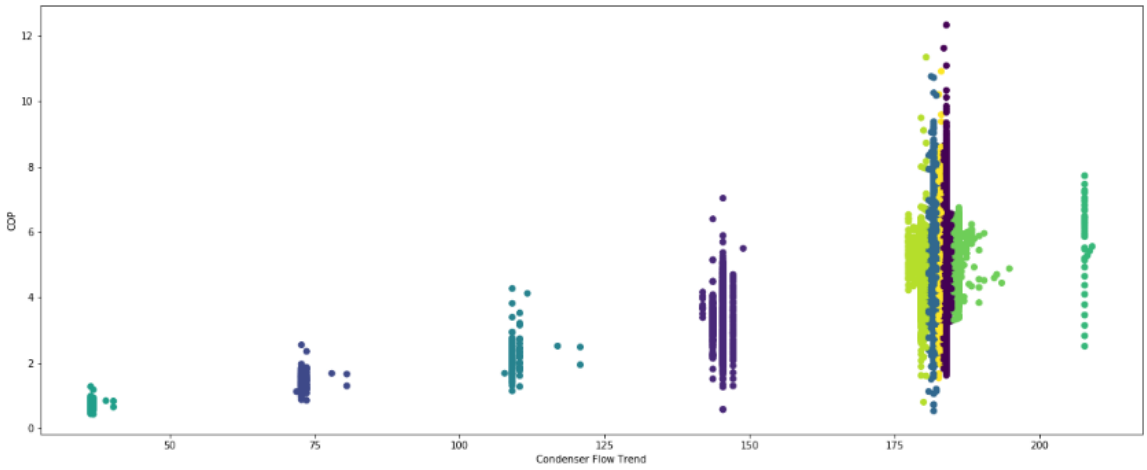

Figure 7. Scatter diagram of the condenser flow trend and COP after clustering.

Figure 8 is a scatter diagram of kW/RT and COP after clustering using the condenser flow trend. The diagram shows that besides a small number of scatter data, the data points of each cluster presented a convex to origin. The data distribution mode was more precise than that plotted chart in Figure 3.

Based on the aforementioned two points, we validated that the condenser flow trend was a suitable variable for clustering chiller COP data.

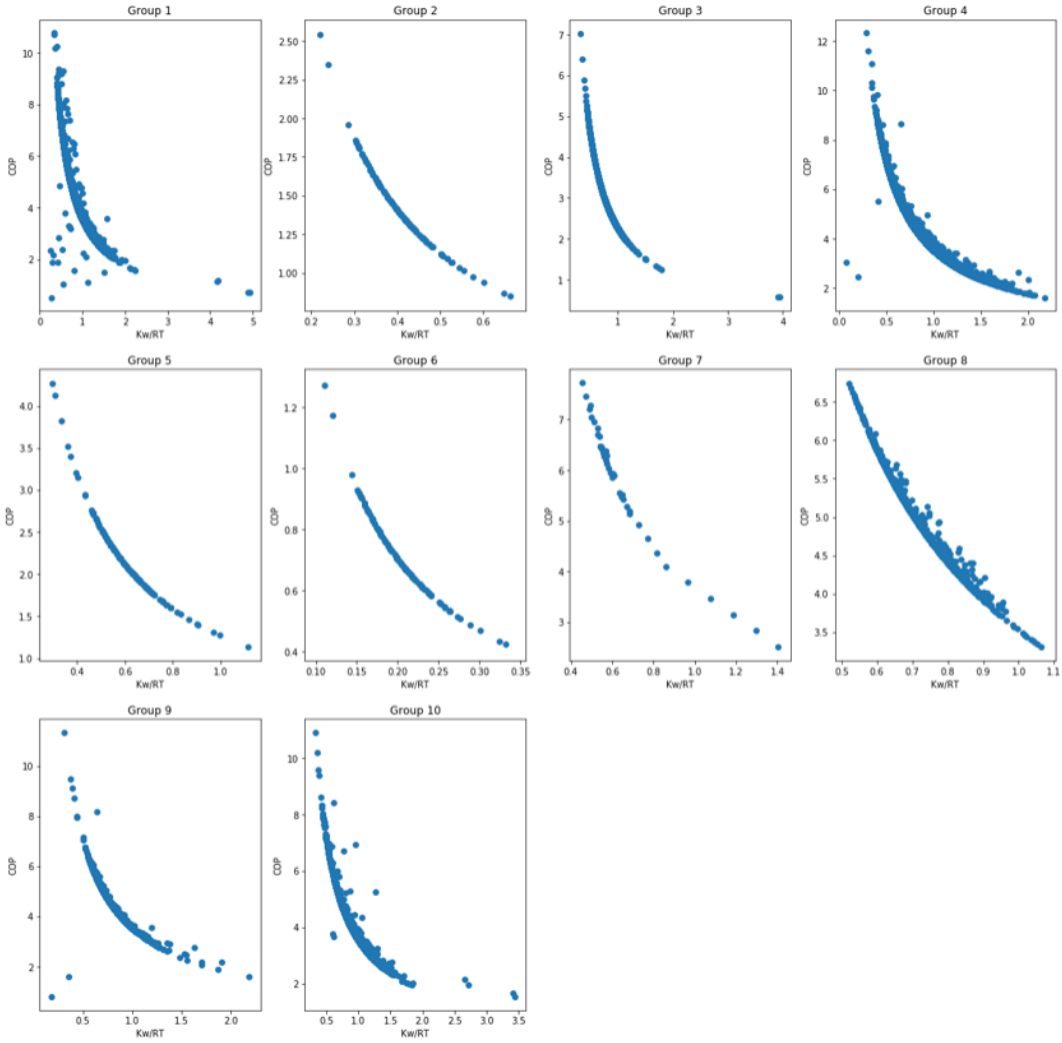

Figure 8. Scatter diagram of the kW/RT and COP after clustering.

Figure 9 shows the outcome of gap statistic. The x-coordinate is the number of cluster k, and the y-coordinate is the gap value Gap(k). The optimal value for clustering is the smallest value k satisfied Expression (10). Here, the optimal value for clustering was $k = 10$. Subsequently, the clustered data was incorporated into the prediction models, and the individual test error and overall test error of 10 clusters were calculated. The results were presented as sum of squares (SSE) and MSE, where SSE was the value of MSE without average. The ideal results and post-integration performance of the different prediction models are tabulated in Tables 5 and 6.

Solely examining the overall error of the models, the performance of the models was similar for the clustered data and the unclustered data. A closer observation of the performance of individual clusters revealed that the models performed better in 7 of the 10 clusters compared to the unclustered data, suggesting that poor model performance was a direct result of a few individual clusters. We performed an in-depth review into the clustering results to explain this phenomenon and found that Clusters 1, 4, 8, 9, and 10 were the aforementioned larger data clusters with values ranging between 175 and 200. The cluster boundaries of these clusters were less prominent compared to the other clusters. We speculate that the clustering approach adopted in this study was less capable of processing the

data volume, resulting in the clustering results not fully reflecting the data modes. In response, we attempted to calibrate the clusters to resolve this issue.

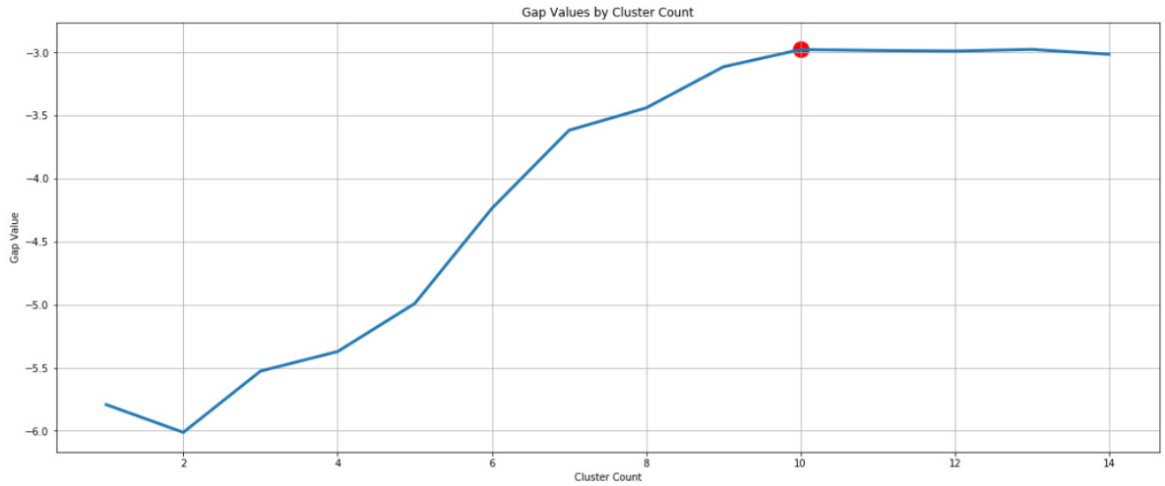

Figure 9. Output of gap statistic. The optical value for clustering was $k = 10$.

Table 5. Evaluation metrics of predict model for COP after clustering of each group.

Group	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**
SSE	19.0886	0.17391	0.0608	140.049	0.13358
MSE	0.00201	0.00295	0.00011	0.00305	0.0014
Group	**Group 6**	**Group 7**	**Group 8**	**Group 9**	**Group 10**
SSE	0.0521	1.609	293.0232	0.1844	14.4399
MSE	0.00084	0.04597	0.0673	0.00121	0.00536

Table 6. Evaluation metrics of predict model for COP after clustering.

Evaluation Metrics	**Total MSE**	**Total CVRMSE**	**Total MAPE**
Value	0.00707	0.01906	0.00904

Two calibration methods were adopted. The first method involved independently clustering the five sets of data to eliminate the effects of the other data. The second method was grouping the data in the five clusters without clear boundaries into one cluster for analysis. The assessment results of the two calibration methods are tabulated in Table 7.

Table 7. Evaluation metrics of predict model for COP calibrated.

	Method 1	**Method 2**
Total MSE	0.003386	0.002616
Total CVRMSE	0.014329	0.012591
Total MAPE	0.006657	0.006075

The table shows that the calibrated results produced using the first method were similar to the initial clustering results. In contrast, the calibrated results produced using the second method were better than the original clustering results, suggesting that integrating clustering and machine learning can improve model predictions after appropriate calibration.

The percentages of improvement between the results of this study and those of the original prediction models are tabulated in Table 8. The target of comparison was the XGB model, which

had the best performance among the original prediction models. The results show that although computation time increased by 80% after clustering and calibration, the MSE, CVRMSE, and MAPE of the proposed method reduced by 21.35%, 11.96%, and 19%, respectively, suggesting a significant improvement in prediction accuracy.

Table 8. Lift percentage of proposed model.

Model	MSE	CVRMSE	MAPE	Time (s)
Proposed model	21.35%	11.96%	19%	−80%

The results confirm that clustering can effectively enhance the quality of chiller data and increase the efficiency of incorporating machine learning in the prediction of chiller data if the limitations were satisfied: (1) If the data could be clustered well or (2) if the clustering method failed to get good results, the revised approach must work.

5. Conclusions

In this study, we first simulated the common prediction models for chiller system. The best results were produced by the random forest and XGBoost models. Then, we employed statistical analysis methods, K-means clustering, and gap statistic to identify the ideal clustering variables and clustering value *k*. We successfully identified the key variables suitable for clustering and enhanced data quality and usability for prediction. We adopted MSE, CVRMSE, MAPE, and times as the assessment standards. After simulation and suitable calibration, MSE, CVRMSE, and MAPE improved by 21.35%, 11.96%, and 19%, respectively, without drastically increasing computation time. Therefore, we successfully improved the prediction accuracy of the model.

The findings of this study may serve as a reference for third parties responsible for assessing energy efficiency in the future. Applying the procedures outlined in this study for establishing a prediction model can effectively improve the accuracy of energy efficiency verification, reduce prediction error, and enhance the reliability of the improvement method.

In this research, the situations in which clustering methods may fail to get good results were not fully listed. In the future, the flowchart of establishing prediction model can be expanded for application in general contexts.

Author Contributions: Methodology, C.-C.L.; Project administration, C.-Y.L.; Writing—original draft, C.-C.L.; Writing—review & editing, C.-W.C.; Supervision, C.-W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

COP	Coefficient of performance
T_{ci}	Cooling water inlet temperature
T_{wi}	Chilled water outlet temperature
Q_e	Cooling capacity
XGBoost	Extreme gradient boosting
ANN	Artificial neural networks
MSE	Mean-square error
CVRMSE	Coefficient of variation of root-mean squared error
MAPE	Mean absolute percentage error
MIC	Maximal information coefficient

Appendix A. Review of Detailed Machine Learning Algorithm

Appendix A.1. Lee Simplified Model

Equation (A1) describe the prediction model of coefficient of performance (COP):

$$\frac{1}{cop} = -1 + \frac{T_{ci}}{T_{wi}} + \frac{1}{Q_e}\left[-A_0 + A_1 T_{ci} - A_2 \frac{T_{ci}}{T_{wi}}\right] \tag{A1}$$

where A_0, A_1 and A_2 are coefficients of model and can be derived by regression analysis. Let $\alpha = \left(\frac{1}{COP} + 1 - \frac{T_{ci}}{T_{wi}}\right) * Q_e$, Equation (A1) becomes:

$$\alpha = -A_0 + A_1 T_{ci} - A_2 \frac{T_{ci}}{T_{wi}} \tag{A2}$$

then set $\beta = \alpha + A_2 \frac{T_{ci}}{T_{wi}}$, Equation (A2) becomes:

$$\beta = A_1 T_{ci} - A_0 \tag{A3}$$

The coefficient A_2 can be calculated by regressing α on $\frac{T_{ci}}{T_{wi}}$, and the coefficients A_0 and A_1 can be calculated by regressing β on T_{ci}.

Appendix A.2. Random Forest

Let $(\boldsymbol{x}_n, y_n)$ represent a data set with n instances, the form of CART can be written as:

$$\sum_{i=1}^{n} \sum_{k=1}^{K} C_k I(\boldsymbol{x}_i \in R_k) \tag{A4}$$

where R_k is the k^{th} output space and C_k is average value of R_k. Output space are split by calculating feature j and node s satisfied Expression (A5):

$$\min_{j,s}[\min_{C_1} \sum_{\boldsymbol{x}_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{C_1} \sum_{\boldsymbol{x}_i \in R_2(j,s)} (y_i - c_2)^2] \tag{A5}$$

Combine Expression (A4) and Expression (A5), the form of random forest can be written as:

$$\hat{y}_i = \sum_{m=1}^{M} \sum_{i=1}^{n} \sum_{k=1}^{K} C_k I(\boldsymbol{x}_i \in R_k) \tag{A6}$$

where m is the number of CART in random forest model.

Appendix A.3. Extreme Gradient Boosting

For a given data set with n examples and d features $\mathcal{D} = \left\{(\boldsymbol{x}_i, y_i) \middle| \boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1, \ldots, n\right\}$, Equation (A7) describes a tree ensemble model using K additive functions to predict the output:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{A7}$$

where f_k is the k^{th} CART model. The CART model can be expressed as Equation (A8):

$$f_k(\boldsymbol{x}_i) = w_{q(\boldsymbol{x}_i)} \tag{A8}$$

where q is the structure of CART that maps the inputs x_i to the corresponding output space, w is the weights of output space. To learn the optimal parameters used in prediction model, Equation (A9) describes the regularized objective function Obj:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{A9}$$

where l is a differentiable convex loss function and Ω is the complexity of the model. Here, the loss function is least squares method $(y_i - \hat{y}_i)^2$, and Ω is defined as Equation (A10):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{A10}$$

where T is the number of output space, γ and λ are hyper parameters.

Because tree ensemble model is an additive function, the objective function should satisfy $Obj^{(t)} < Obj^{(t-1)}$. Let $\hat{y}_i{}^{(t)}$ be the prediction of the i^{th} instance at the t^{th} iteration, Equation (A11) becomes:

$$\hat{y}_i{}^{(t)} = \sum_{k=1}^{t} f_k(\boldsymbol{x}_i) = \hat{y}_i{}^{(t-1)} + f_t(\boldsymbol{x}_i) \tag{A11}$$

and Equation (A9) becomes:

$$Obj^{(t)} = \sum_{i=1}^{n} l\Big(y_i, \hat{y}_i{}^{(t-1)} + f_t(\boldsymbol{x}_i)\Big) + \Omega(f_t) + constant \tag{A12}$$

Here, the term $\sum_{k=1}^{t} \Omega(f_k)$ can be expanded to $\Omega(f_t) + \sum_{k=1}^{t-1} \Omega(f_k)$, and $\sum_{k=1}^{t-1} \Omega(f_k)$ can be regarded as a constant.

To minimize the objective function, Equation (A12) can be expanded and rewritten as following.

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} \Big[y_i - \Big(\hat{y}_i{}^{(t-1)} + f_t(\boldsymbol{x}_i)\Big)\Big]^2 + \Omega(f_t) + constant \\ &= \sum_{i=1}^{n} \Big[\Big(y_i - \hat{y}_i{}^{(t-1)}\Big) - f_t(\boldsymbol{x}_i)\Big]^2 + \Omega(f_t) + constant \\ &= \sum_{i=1}^{n} \Big[l\Big(y_i, \hat{y}_i{}^{(t-1)}\Big)^2 - 2l\Big(y_i, \hat{y}_i{}^{(t-1)}\Big) f_t(\boldsymbol{x}_i) + f_t^2(\boldsymbol{x}_i)\Big] + \Omega(f_t) + constant \\ &= \sum_{i=1}^{n} \Big[l\Big(y_i, \hat{y}_i{}^{(t-1)}\Big)^2 + g_i f_t(\boldsymbol{x}_i) + \tfrac{1}{2} h_i f_t^2(\boldsymbol{x}_i)\Big] + \Omega(f_t) + constant \end{aligned} \tag{A13}$$

where $g_i = \frac{\partial l\left(y_i, \hat{y}_i{}^{(t-1)}\right)}{\partial \hat{y}_i{}^{(t-1)}}$ and $h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i{}^{(t-1)}\right)}{\partial \hat{y}_i{}^{(t-1)}}$ are first- and second-order gradient statistics on the loss function. Then, remove the constant term, and the objective function becomes Equation (A14):

$$Obj^{(t)} = \sum_{i=1}^{n} \Big[g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i)\Big] + \Omega(f_t) \tag{A14}$$

Finally, $f_t(x_i)$ and $\Omega(f_t)$ are substituted by Equation (A8) and (A10):

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} \Big[g_i f_t(\boldsymbol{x}_i) + \tfrac{1}{2} h_i f_t^2(\boldsymbol{x}_i)\Big] + \Omega(f_t) \\ &= \sum_{i=1}^{n} \Big[g_i w_{q(\boldsymbol{x}_i)} + \tfrac{1}{2} h_i w_q^2(\boldsymbol{x}_i)\Big] + \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \\ &= \sum_{j=1}^{T} \Big[G_j w_j + \tfrac{1}{2}\Big(H_j + \lambda\Big) w_j^2\Big] + \gamma T \end{aligned} \tag{A15}$$

where $G_j = \left(\sum_{i \in I_j} g_i\right)$ and $H_j = \left(\sum_{i \in I_j} h_i\right)$ represents the sum of first- and second-order gradient statistics in output space j.

For a fixed structure $q(\boldsymbol{x}_i)$, the optimal parameter w_j^* and corresponding value Obj^* of output space j can be calculated by

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{A16}$$

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{{G_j}^2}{H_j+\lambda} + \gamma T \tag{A17}$$

Appendix A.4. K-Means

Let $\left\{\boldsymbol{x}_i \middle| \boldsymbol{x}_i \in \mathbb{R}^d, i = 1,\ldots,n\right\}$ be the set of d-dimensional points to be clustered into a set of k clusters, $\left\{\boldsymbol{\mu}_c^{(t)} \middle| \boldsymbol{\mu}_c^{(t)} \in \mathbb{R}^d, c = i,\ldots,k\right\}$ be the cluster centers. Equation (A18) calculates the Euclidean distance of each sample and classified into its nearest cluster $S_c^{(t)}$ at the t^{th} iteration:

$$S_c^{(t)} = \left\{\boldsymbol{x}_i : \|\boldsymbol{x}_i - \boldsymbol{u}_c^{(t)}\|^2 < \|\boldsymbol{x}_i - \boldsymbol{u}_{c'}^{(t)}\|^2, \forall i = 1,\ldots,n\right\} \tag{A18}$$

Equation (A19) describes how to update $\boldsymbol{\mu}_c^{(t)}$:

$$\boldsymbol{u}_c^{(t+1)} = \frac{1}{n_c}\sum_{\boldsymbol{x}_i \in S_c^{(t)}} \boldsymbol{x}_i \tag{A19}$$

where n_c is the number of points in c^{th} cluster. K-means repeats formula (A18) and (A19) until $S_c^{(t+1)} = S_c^{(t)}$.

Appendix A.5. Gap Statistics

Using data set defined in Appendix A.4, let D_c be the sum of the pairwise distances for all points in cluster S_c and W_c be the pooled within-cluster sum of squares around the cluster means. Equation (A20) and (A21) describe the formula of D_c and W_c:

$$D_c = \sum_{\boldsymbol{x}_i \in S_c}\sum_{\boldsymbol{x}_j \in S_c}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = 2n_c * \sum_{\boldsymbol{x}_i \in S_c}\|\boldsymbol{x}_i - \boldsymbol{u}_c\|^2 \tag{A20}$$

$$W_c = \sum_{c=1}^{k}\frac{1}{2n_c}D_c = \sum_{c=1}^{k}\sum_{\boldsymbol{x}_i \in S_c}\|\boldsymbol{x}_i - \boldsymbol{u}_c\|^2 \tag{A21}$$

The idea of gap statistic is to standardize the graph of $\log(W_c)$ by comparing it with its expectation under an appropriate null reference distribution of the data [19]. The estimate of the optimal k is the value for which $\log(W_c)$ falls the farthest below this reference curve. Hence, the optimal k is the smallest value k satisfied expression (A22):

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1} \tag{A22}$$

$\mathrm{Gap}(k)$ and s_{k+1} are described in Equation (A23) and (A24).

$$\mathrm{Gap}(k) = \frac{1}{B}\sum_{b=1}^{B}\log\left(W_{c,b}^*\right) - \log(W_c) \tag{A23}$$

$$s_k = \sqrt{\frac{1+B}{B}}\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\log\left(W_{c,b}^*\right) - \frac{1}{B}\sum_{b=1}^{B}\log\left(W_{c,b}^*\right)\right)^2} \tag{A24}$$

where B is the number of sampling.

Appendix B. Compare of Different Clustering Methods

In this appendix, we ran and compared different clustering methods to validate whether K-means is a good choice or not. In total, we ran four clustering methods to compare with K-means. The four clustering methods are Mean-shift, OPTICS, Birch, and HDBSCAN. We summarized a detailed information of each clustering methods. Table A1 describes the detailed information.

Table A1. Detailed information of each clustering methods.

Methods	Parameters	Number of Clustering	Time (s)
Mean-shift	bandwidth	12	617
OPTICS	epsilon MinPts	40	2432
Birch	Not necessary	3	3
HDBSCAN	Not necessary	18	40
K-means	number of clustering	10	2

From Table A1, the computation speed of Birch, HDBSCAN and K-means are better than Mean-shift and OPTICS. Then, we plotted the scatter diagram of each clustering methods in Figure A1. From Figure 1, none of these five methods could perfectly separate the data, and a calibration method was necessary for the next research. Observing the scatter diagram, K-means seems to be a better method. It well separated data from each other without noises except data, which values ranging between 175 and 200. The calibration of K-means appeared easier than others. Hence, we selected K-means as the clustering method used in this research.

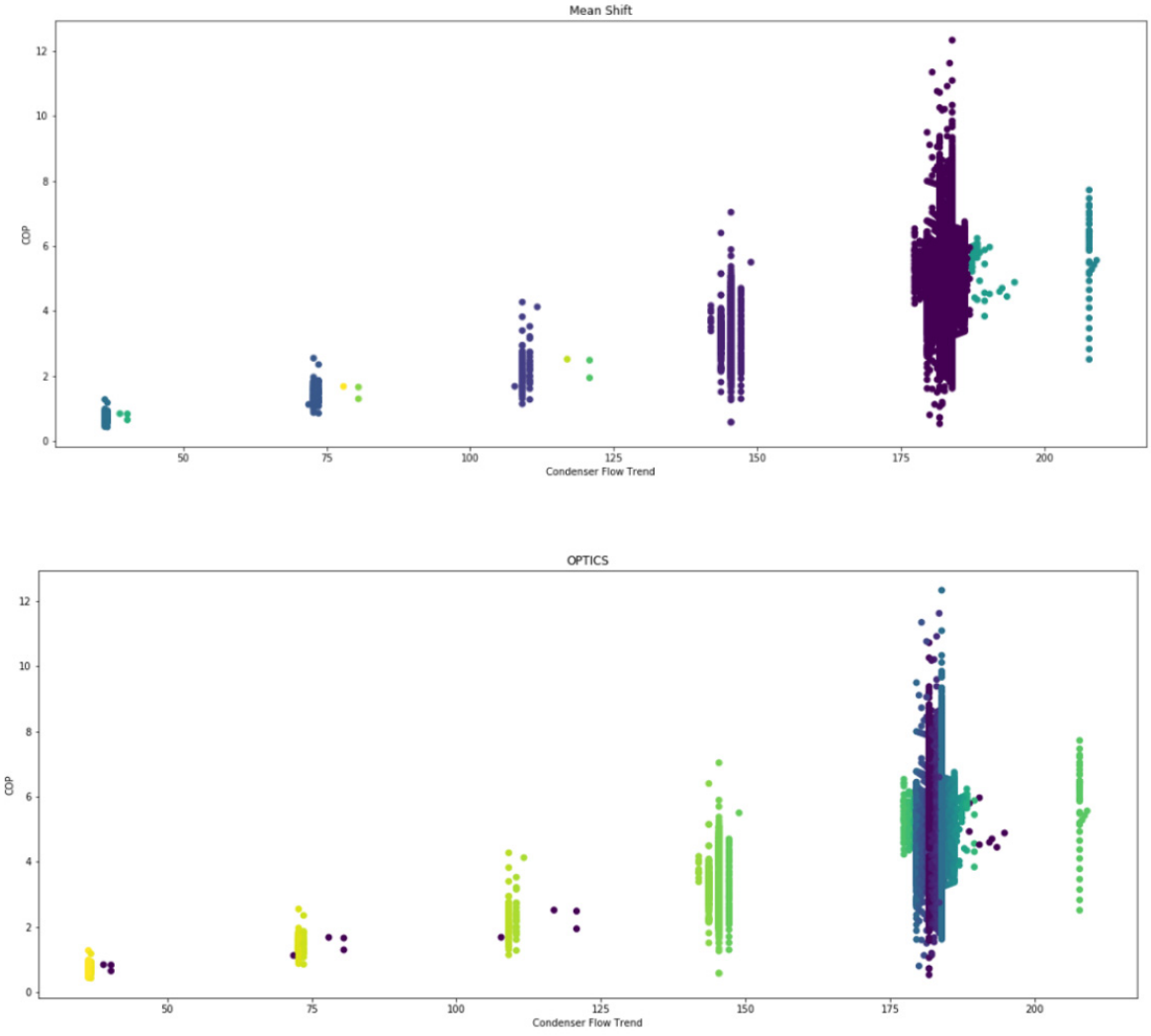

Figure A1. *Cont.*

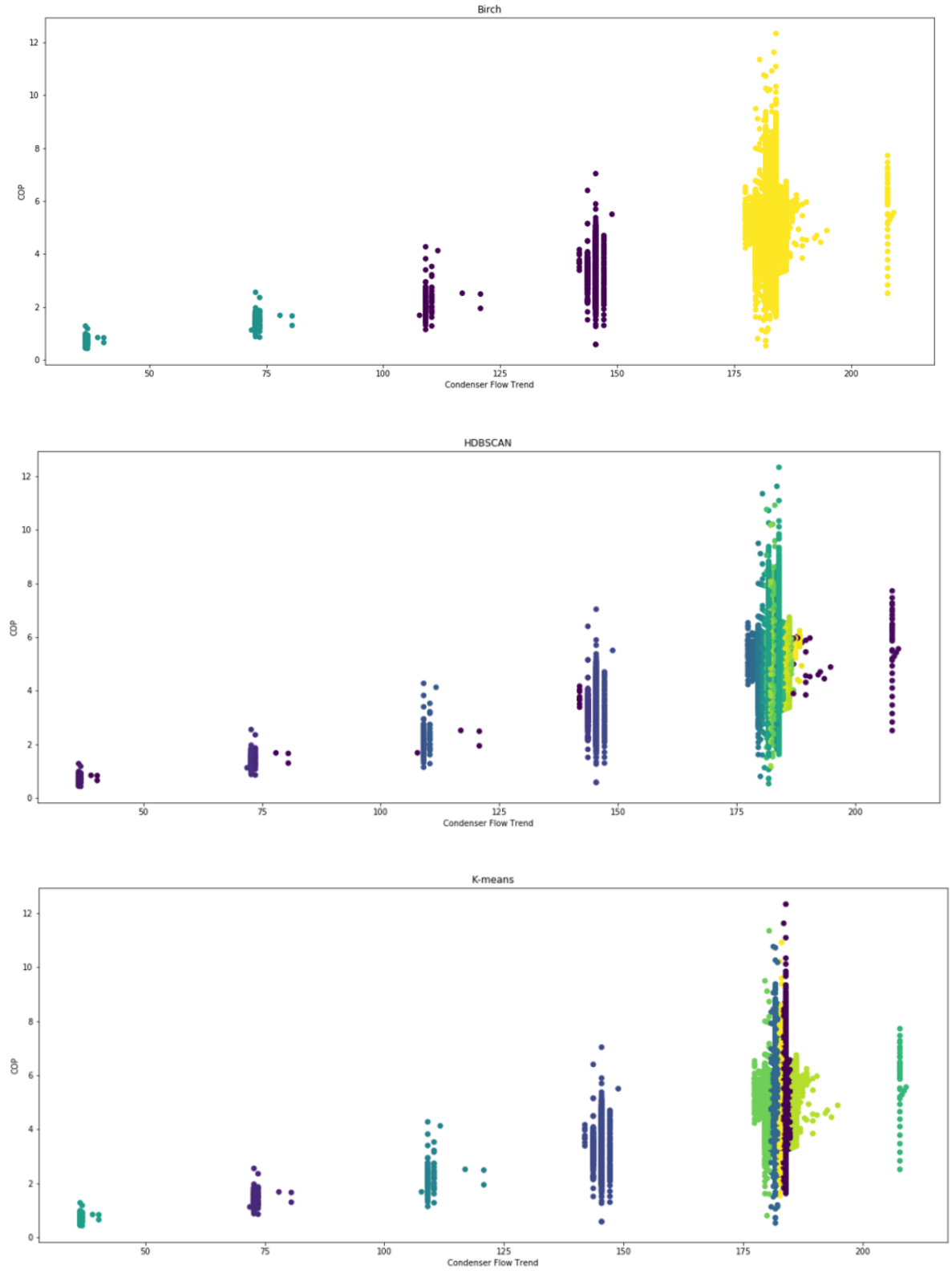

Figure A1. Scatter diagram of the condenser flow trend and COP in different clustering methods. There were five methods validated in this research. The scatter diagram are Mean-shift, OPTICS, Birch, HDBSCAN, and K-means from top to bottom.

References

1. Lee, T.S. Thermodynamic Modeling and Experimental Validation of Screw Liquid Chillers. *Ashrae Trans.* **2004**, *110*, 206–216.
2. Reddy, T.A.; Andersen, K.K. An Evaluation of Classical Steady-State Off-Line Linear Parameter Estimation Methods Applied to Chiller Performance Data. *HVAC&R Res.* **2002**, *8*, 101–124. [CrossRef]
3. Adnan, W.N.W.M.; Dahlan, N.Y.; Musirin, I. Modeling baseline electrical energy use of chiller system by artificial neural network. In Proceedings of the 2016 IEEE International Conference on Power and Energy (PECon), Melaka, Malaysia, 28–29 November 2016; pp. 500–505.
4. Kim, J.-H.; Seong, N.C.; Choi, W. Modeling and Optimizing a Chiller System Using a Machine Learning Algorithm. *Energies* **2019**, *12*, 2860. [CrossRef]
5. Yu, F.; Ho, W.; Chan, K.; Sit, R. Probabilistic and electricity saving analyses of Mist Coolers for Chiller System in a Hotel. *Energy Procedia* **2017**, *143*, 154–160. [CrossRef]
6. Yu, F.W.; Ho, W.; Chan, K.; Sit, R. Critique of operating variables importance on chiller energy performance using random forest. *Energy Build.* **2017**, *139*, 653–664. [CrossRef]
7. Malinao, J.; Judex, F.; Selke, T.; Zucker, G.; Caro, J.; Kropatsch, W. Pattern mining and fault detection via COP_therm-based profiling with correlation analysis of circuit variables in chiller systems. *Comput. Sci. Res. Dev.* **2016**, *31*, 79–87. [CrossRef]
8. Habib, U.; Zucker, G.; Blochle, M.; Judex, F.; Haase, J. Outliers detection method using clustering in buildings data. In Proceedings of the IECON 2015—41st Annual Conference of the IEEE Industrial Electronics Society, Yokohama, Japan, 9–12 November 2015; pp. 694–700.
9. Habib, U.; Hayat, K.; Zucker, G. Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering. *Complex Adapt. Syst. Model.* **2016**, *4*, 1762. [CrossRef]
10. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
11. Chakraborty, D.; Elzarka, H. Advanced machine learning techniques for building performance simulation: A comparative analysis. *J. Build. Perform. Simul.* **2019**, *12*, 193–207. [CrossRef]
12. Park, S.; Moon, J.; Jung, S.; Rho, S.; Baik, S.W.; Hwang, E. A Two-Stage Industrial Load Forecasting Scheme for Day-Ahead Combined Cooling, Heating and Power Scheduling. *Energies* **2020**, *13*, 443. [CrossRef]
13. Wang, Z.; Hong, T.; Piette, M.A. Building thermal load prediction through shallow machine learning and deep learning. *Appl. Energy* **2020**, *263*, 114683. [CrossRef]
14. Zheng, H.; Yuan, J.; Chen, L. Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies* **2017**, *10*, 1168. [CrossRef]
15. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
16. Fränti, P.; Sieranoja, S. How much can k-means be improved by using better initialization and repeats? *Pattern Recognit.* **2019**, *93*, 95–112. [CrossRef]
17. Liang, J.; Bai, L.; Dang, C.; Cao, F. The K-Means-Type Algorithms Versus Imbalanced Data Distributions. *IEEE Trans. Fuzzy Syst.* **2012**, *20*, 728–745. [CrossRef]
18. Melnykov, I.; Melnykov, V. On K-means algorithm with the use of Mahalanobis distances. *Stat. Probab. Lett.* **2014**, *84*, 88–95. [CrossRef]
19. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2001**, *63*, 411–423. [CrossRef]
20. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting Novel Associations in Large Data Sets. *Science* **2011**, *334*, 1518–1524. [CrossRef] [PubMed]
21. Granderson, J.; Touzani, S.; Custodio, C.; Sohn, M.; Fernandes, S.; Jump, D. *Assessment of Automated Measurement and Verification (M&V) Methods*; Lawrence Berkeley National Laboratory: Berkeley, CA, USA, 2015.