

Article

Development of a Linear Regression Model Based on the Most Influential Predictors for a Research Office Cooling Load

Ntumba Marc-Alain Mutombo * and Bubele Papy Numbi

Department of Electrical Engineering, Mangosuthu University of Technology, Umlazi 4031, South Africa; numbib@mut.ac.za

* Correspondence: marcntumba@gmail.com; Tel.: +27-63-079-2115

Abstract: Energy consumption in the building sector is a major concern, particularly in this time of worldwide population and energy demand increases. To reduce energy consumption due to HVAC systems in the building sector, different models based on measured data have been developed to estimate the cooling load. The purpose of this work is to develop a linear regression model for cooling load of a research room based on the radiant time series (RTS) components of the cooling load that consider the building material and the environment. Using the forward step method, linear regression models were developed for both all-seasons and seasonal data from three years of cooling load data obtained from the RTS method for a research room at Mangosuthu University of Technology (MUT), South Africa. The male and female occupants, window cooling load, and roof cooling load were found to be the most influential predictors for the cooling load model. The obtained relative errors between the best all-seasons model and seasonal models built with the same predictors for the respective data subsets are almost zero and are given as 0.0073% (autumn), 0.0016% (spring), 0.0168% (summer), and 0.0162% (winter). This leads to the conclusion that the seasonal models can be represented by the all-seasons model. However, further study can be performed to improve the model by incorporating the occupancy behaviours and other components or parameters intervening in the calculation of cooling load using the radiant time series method.

Keywords: radiant time series; cooling load; linear regression



Citation: Mutombo, N.M.-A.; Numbi, B.P. Development of a Linear Regression Model Based on the Most Influential Predictors for a Research Office Cooling Load. *Energies* **2022**, *15*, 5097. <https://doi.org/10.3390/en15145097>

Academic Editor: David Borge-Diez

Received: 30 April 2022

Accepted: 17 June 2022

Published: 12 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Different components contribute to the cooling load of a building, in particular the weather condition, the building material, and the occupancy. The construction of a building is closely related to the climate in which the building is located, and climate change has now become an important issue in the building sector [1]. Solar radiation through the windows is the most contributing component of the room cooling load with about 30% of cooling energy consumption [2,3] allocated to the heat gain through the windows [4]. Hence, the room cooling load profile follows the cycle properties of the solar radiation [5]. The impact of occupancy on the building cooling load is mostly related to the use of appliances. Since the room is designed for a specific number of people and appropriate activities to be performed within it, overheating can sometimes occur when we move away from the designed specifications [6–9], and affect the indoor comfort. The cooling within the building can be improved by the choice of material with excellent thermal inertia [6] that will impact on the radiant time series factors of the cooling load model of the building. Avoiding overcooling and maintaining the best thermal comfort [10] are the main goals of the heating, ventilation, and air conditioning (HVAC) system in the building. This is achieved with a better control strategy of the system based on a good model that considers the contribution of each of the factors contributing to the cooling load in the building.

The building sector is responsible for 40% to 50% of energy consumption. This is because of the high expected occupancy, heavy appliance use, and a relatively high

proportion of window area with respect to walls. For these reasons, the premises have significant HVAC systems installed [11]. Because these systems are scheduled or they can exploit the thermal inertia of the building, their electrical loads can be used for demand-response (DR) applications [12]. For the DR to properly operate, it should be able to predict short-term electricity consumption for individual load by identifying the electricity to be consumed at a specific time and the possible available consumption at the moment [13]. Generally, the load in office buildings is mostly influenced by the weather conditions, the occupancy behaviour, and the usage time [14]. Moreover, the predicted load profile is used as reference to the measured DR signal impacted load profile. This is used to confirm that a DR activity was accomplished, and the market commitments were fulfilled. Cooling load prediction is a possible part of the solution to improve the energy consumption in the building sector. In essence, if the cooling load is forecasted, the HVAC energy consumption can be adjusted based on the cooling load forecasted with an efficient energy management system. This ensures the HVAC energy demand evaluability. The cooling load demand is determined by knowing different predictors. These are variables of the model that allow determining the cooling load based on the developed model. The model is crucial and relevant to analysing and performing energy management plans in the building sector to minimise building energy consumption.

Different methods of prediction have been used before in the energy sector. There are several publications in the literature that focus on predicting power usage in building stock. For the application, there are three types of prediction methods which are the simple averaging, statistic-based models, and the artificial-intelligence-based models. The two last are mostly used due, particularly, to the large amount of data that can be collected and the evolution in computer systems and software. The statistic-based models are mathematical models using sets of statistical assumptions on the data to develop the appropriate models, and the widely used ones in this category are multiple regressions and time series forecasting. Linear regression [15] has been used to develop the relationship between the independent variable and its predictors, especially if there is a linear behaviour between these variables. Additionally, the reliability and the re-applicability of this method have made it popular [16], whereas artificial-intelligence-based methods use machine learning algorithms to develop models, and the artificial neural network (ANN) is the most used in this category. Most of the research on load forecasting is focused on classical forecasting methods such as autoregressive inference moving average (ARIMA) and linear regressions [17]. For instance, different classification methods were used to forecast the electricity consumption of the United States in the short term, and it has been found that the autoregressive integrated moving average (ARIMA) time series model performs best compared to all other methods [18]. In addition, similar error rates are obtained when ARIMA models are used to predict short-term load [19,20]. The time series model performs better when predicting the load in the near future. However, the weakness of this method is the lack of understanding of what produces consumption variance. Moreover, the generated models are often difficult to interpret [13]. More accuracy is obtained by using multiple regression, ANN, and regression tree algorithms to predict weekly residential load. This has been demonstrated for Hong Kong data [21]. In addition, a multiple linear regression model was used to predict hourly system demand for United States utilities, using time and outdoor temperature as independent variables. The model was obtained by gradually adding predictors and potential cross effects [22]. More load regression models and ANN studies can be found in [23,24]. A study on the use of ANN to forecast short-term load was also performed, and the algorithm generates detailed and complicated models that are prone to overfitting; there is no systematic validation of the results of prediction [25]. Other load prediction studies that focus especially on office buildings can be found in [18,26,27].

The models in previous works are developed from measured data related to the Northern Hemisphere, with the climate parameters and the material used in building different from the Southern Hemisphere. In contrast to the previous models, the developed models in this work are based on the determination of the cooling load components by the

radiant time series (RTS) method, considering the building material and the environmental parameters for a time resolution of 1 h. This allows the evaluation of the cooling load at the early stage of the building design without measurement. To achieve optimal energy efficiency and thermal comfort, the dynamics of building physics [10], which are introduced in the time series method by the RTS factors for the present case, and different inner heat gain sources are considered in the model. The primary goal of this paper is to propose a data analysis methodology for identifying and producing models based on the RTS cooling load components that can predict a cooling load for a research room in an office building at Mangosuthu University of Technology (MUT), South Africa. The cooling load consists of three years of data for autumn, spring, summer, and winter seasons. The model related to all data together is developed with a minimum number of predictors, and the predictors from the all-season model are used as variables for the models related to each season. Following that, the best model is used to evaluate the models' performance based on the acquired Akaike information criterion (AIC), Bayesian information criterion (BIC), performance score, and root mean squared error (RMSE).

2. Materials and Methods

The data obtained from the simulation based on the RTS method are processed in R so that exploratory data analysis (EDA) can be performed. The EDA produces a processed and delineated dataset as the main output. Here, the data are merged and placed into a suitable format for further processing in R, then the data are analysed to understand how they are distributed and clustered. The analysis also aims to determine irrelevant data if possible. This is accomplished using scatterplots (illustrate data distributions against seasons for different years, and plots the variables paired against each other to recognise nonlinear correlations and dependencies) and boxplots (show data distributions against temporal attributes) [28]. The processed data are then divided into two groups, one for training and the other for testing. The training dataset is used for statistical analysis, while the test dataset is used to evaluate model performance. In the studied case, no variable data transformation is performed. To build a statistical model between the predictors and the result variable, a forward linear regression model selection method is used. The linear regression models are statistically evaluated [28] by fitting the training and testing data subsets for the room cooling load. The predictor's coefficients, along with the p-values, the coefficient of correlation (R), the coefficient of determination (R^2), and the root mean squared error (RMSE) [5], are recorded for all the models. The common attribute to the models is their performance and ability to interact with different variables, defining the sources of heat gain that contribute to the cooling load, and to satisfy the needs and comfort within the building [29].

The developed model is based on multiple linear regression method due to the high number of predictors. Multiple linear regression is a relation between a response or dependent variable and two or more independent variables or predictors [30]. The predictors are linearly independent to avoid the collinearity problem between them. By assigning a linear equation to the available data, its related linear model is given by [31]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon, \quad (1)$$

where Y is the dependent variable; X_i are the independent variables; β_i are the coefficients of regression with $i = 0, 1, 2, \dots, k$; and ε is a model error.

Essentially, the goal of this strategy is to find the optimal model using only a subset of the predictors. The best model is obtained by analysing the statistical parameters and removing and interchanging different predictors based on statistical significance. From the first analysis of the data, number of males (NoM), number of females (NoF), number of occupants (NoOcc), occupancy cooling load (OccupCL), lighting cooling load (LightingCL), information and technology cooling load (ITCL), outdoor temperature (OutTemp), indoor temperature (InTemp), wall sol-air temperature (WSATemp), total surface irradiance

(SurfIrrad), wall cooling load (WallCL), window cooling load (WindowCL), roof sol-air temperature (RSATemp), and roof cooling load (RoofCL) are potential predictors.

The correlation chart is drawn to determine the correlations between the variables. The best model is obtained by elimination of less statistically significant predictors. The model with all the predictors is built, the significance of the predictors determined, and the best predictors selected based on their p -values. The selected model is not only the one that has the highest coefficient of determination (R^2) among all possible models, but it must also not have multicollinearity between the predictors.

R^2 is not an appropriate indicator in this case because the model is built to minimise the error rate. A high R^2 means a low error rate on the training data, and as more predictors are added to the model, R^2 increases [32]. There is no assurance that the model will perform accurately when exposed to a new set of data. To overcome this situation, the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) measures can be used to evaluate the selected model. These two criteria create a trade-off between the complexity of the model due to many predictors and the obtained explanatory factor. Both metrics impose a penalty to the model with the difference that the penalty imposed by BIC is higher compared to the penalty imposed by the AIC for each predictor added to the model [13].

The procedure concludes with the fitting of regression models between the predictors found in the previous statistical investigation. The training dataset's observations are utilised to fit the regression models. A regression methodology was chosen because it produced acceptable prediction results in the previous investigations and the models are interpretable, allowing conclusions to be formed about the interpretations [33]. A diagram depicting the used methodologies of the paper is given in Figure 1.

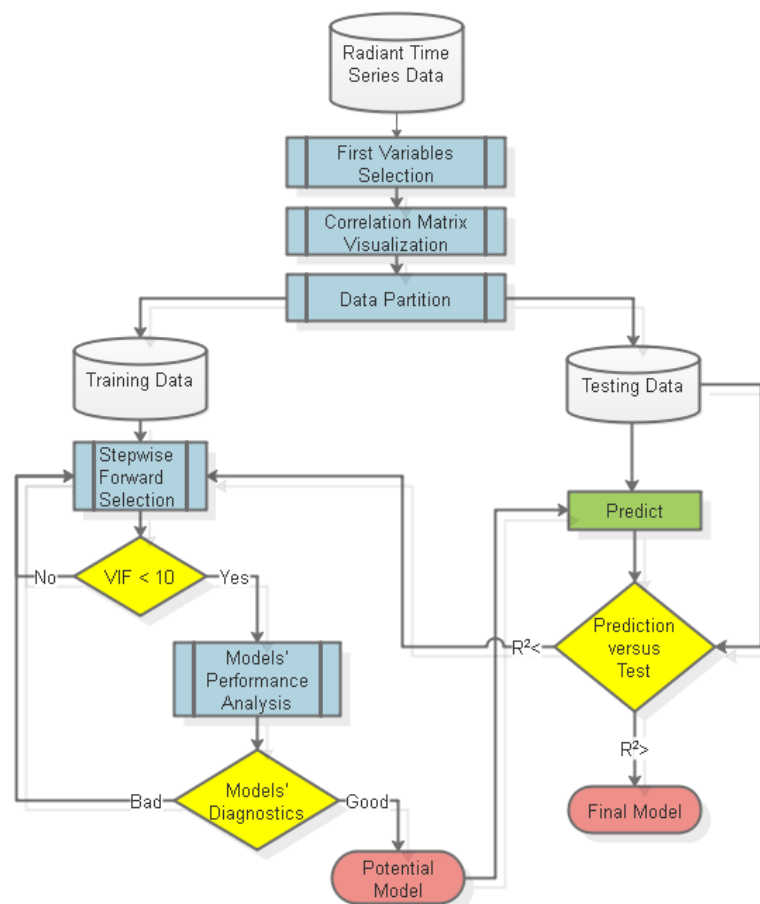


Figure 1. A diagram depicting the used methodologies of the paper.

3. RTS Data Collection and Processing

To develop a linear model for the research room cooling load, the data are calculated first. For this article, the data are generated for the Industrial Energy Efficiency Training and Resource (IEETR) Centre Laboratory room located in the New Engineering Building at Mangosuthu University of Technology (MUT) at Umlazi in KwaZulu-Natal province, South of Durban. These data are obtained by calculation using the RTS method due to the unavailability of the measurement. The room is used for research in renewable energy systems. The work performed in this laboratory is mostly simulation-based. The room is located at the second floor of the building with two opposite inner walls shared with other rooms, and the front wall is shared with the interior door that leads to the hallway. The exterior wall exposed to the rising sun has three folding windows, and single-layered glass occupying 60% of the wall, while the remaining 40% of the exterior wall is made of bricks. The room is equipped with ten desktop computers, one printer to be installed, and fluorescent tubes for lighting. The room is designed to accommodate 10 people working from Monday to Thursday for 8 h a day and Friday for 5 h, with the starting work hour being from 8:00 a.m.

Data for the laboratory are generated with a MATLAB code over 3 years, from 1 January 2015 to 31 December 2017. The MATLAB code was designed based on the RTS method for cooling load calculation. The method RTS was developed by the American Society of Heat and Refrigeration Engineering Association (ASHREA) and is one of the most powerful and most accurate methods in the calculation of the cooling load for buildings. In the RTS method, radiant coefficients are considered for the building material and room light, occupancy, and equipment for 24 h of the day. The data relating to the clear sky optical depth were obtained by calculation from the solar insolation measurement performed onsite by the Suran Lab. The hourly outdoor temperature data were obtained from daily min–max temperatures from the NASA Meteorology website. The temperature was obtained from POWER Data Access Viewer which is a NASA Prediction of Worldwide Energy Resource. Daily maximum, minimum, and average temperatures are given. To obtain the hourly temperature, the area of the triangle built from the max–min values of the temperature is divided into segments and the hourly values of the temperatures are randomly generated within these segments using a MATLAB algorithm. To perform the calculation, all variables in the dataset are set to a global time resolution of 1 h for the processed data frame. Further, to apply the data processing to the occupancy, light, and IT instruments, the data related to these variables are determined for each hour using dummy variables (i.e., 0—no occupancy, light and IT systems off, 1—occupancy, light and IT systems on). This dummy variable is also related to hours of work per day.

4. Exploratory Data Analysis

Secondly, the exploratory data analysis (EDA) is performed on the data obtained from the RTS calculation. The outcomes of the EDA are reported in this section. This means that the data are divided into a training set and a testing set. In addition, the essential variable modifications that can lead to further analysis and modelling are explored.

4.1. Temporal Attributes of the Data: Clustering of Data

The seasonal scatterplot of the room cooling load (RoomCL) for different years and the subplots depicting variations in the hourly, daily, monthly, and seasonal room cooling load in boxplots are obtained in Figures 2 and 3, respectively. The hourly room cooling load from 1 January 2015 to 31 December 2017 is presented in Figure 2. High values of cooling load are recorded during spring and summer while relatively low values are recorded in autumn and winter. Most of the values of the room cooling load are below 1800 W.

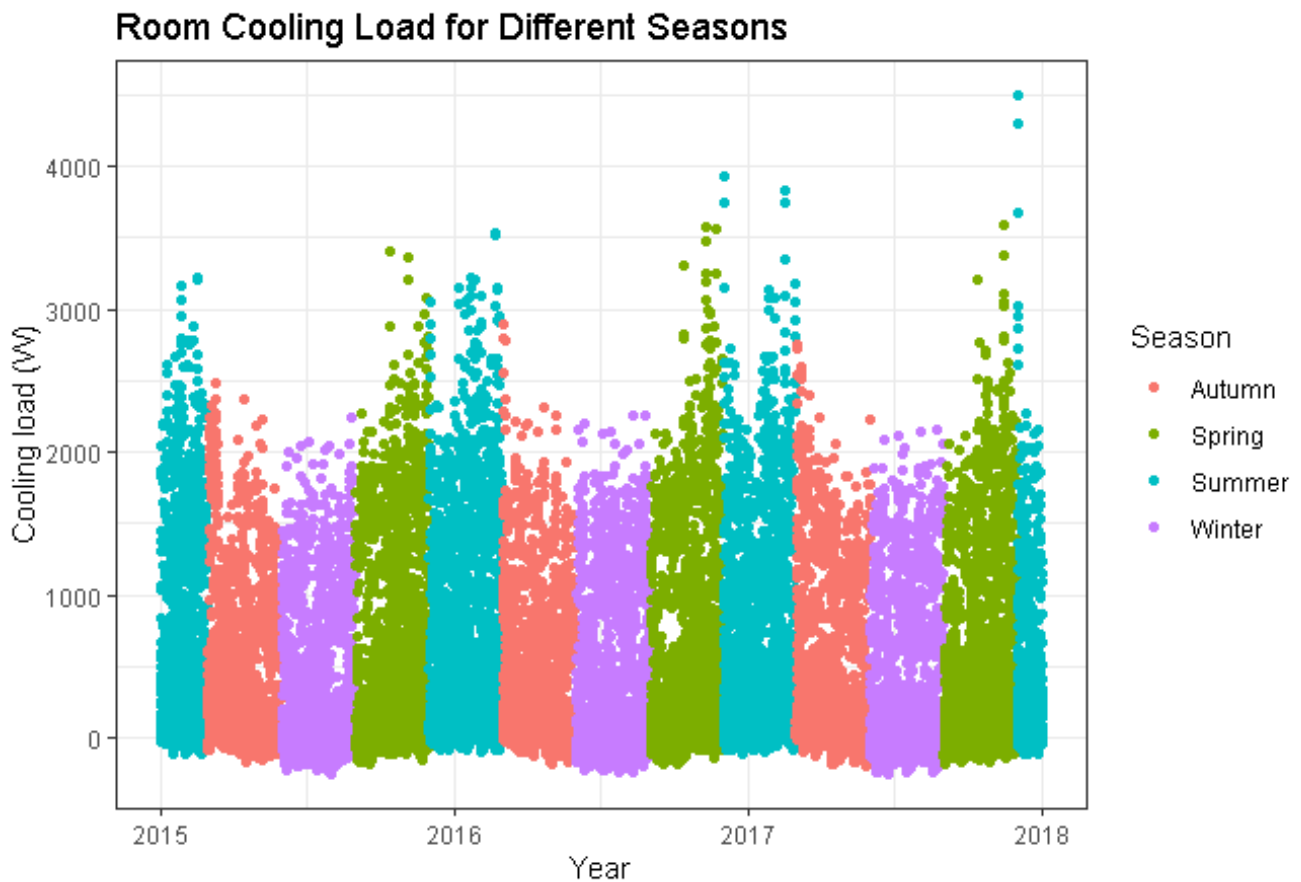


Figure 2. Room cooling load related to different hours from 1 January 2015 to 31 December 2017.

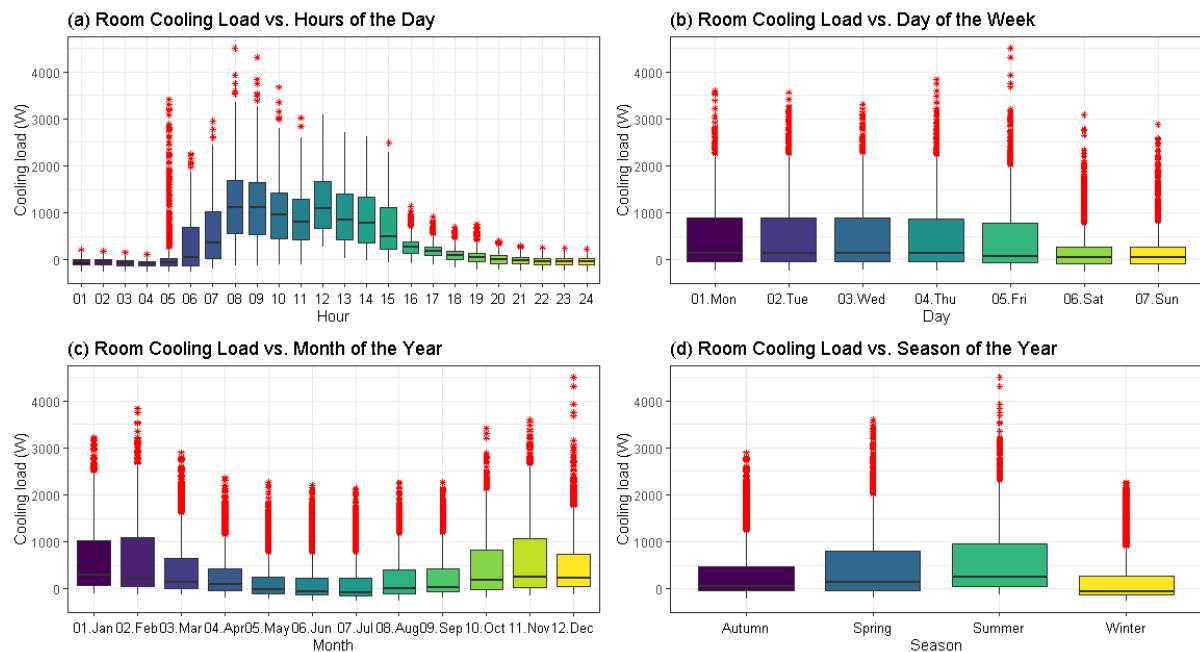


Figure 3. Room cooling load for different times of the year: (a) Room cooling load depicting variations in the hourly boxplots; (b) room cooling load depicting variations in the day boxplots; (c) room cooling load depicting variations in the monthly boxplots; (d) room cooling load depicting variations in the seasonal boxplots.

Four time horizons related to the variation in the load variable are analysed. Those are hours of the day, days of the week, months of the year, and seasons of the year.

The presence of probable outliers is represented by red stars. These points are later checked to determine if they are actually outliers. As observed in subplot (a), high average hourly room cooling loads (>1000 W) are obtained between 8 a.m. and 3 p.m. The RoomCL is practically null from 6 p.m. to 5 a.m. The hourly increase of the RoomCL is significant between 6 a.m. to 3 p.m., following a parabolic shape. This is justified by the sunrise, the working time, the sunset contributing to the room cooling load with solar insolation heat gain, occupancy latent heat, and loss of both after 6 p.m. The room cooling load has a stable base value with a low variance during days of the week, as seen in subplot (b). The daily values of the room cooling loads are calculated for each day of the week, from Monday to Sunday. The room cooling load for working days is higher compared to weekends. This is justified by the inactivity in the room during weekends, the only room cooling load contributor being the solar insolation. The monthly room cooling load in Figure 3c presents a convex curve shape with a minimum value in July. Significant variations can be distinguished monthly. As seen, the room cooling load is high from September to November, and December to February. These months correspond to spring and summer, respectively. During these two seasons, the solar insolation and the ambient temperature are relatively high for spring and considerably high for summer. The rest of the months correspond to the autumn and winter seasons.

The solar insolation and the ambient temperature are relatively low in autumn and considerably low in winter, leading to a low contribution of the solar insolation and ambient temperature to the room cooling load. Although the month of December falls during the summer season and therefore is a hot month, the room cooling load is remarkably lower. This can be justified by the vacations where there is a period of inactivity in the room for almost half of the month. Based on subplot (d), the room cooling load is high in spring and summer, while low values are obtained in autumn and winter, as previously justified in subplot (c). Most of the high values of the cooling load are obtained in summer and autumn, as seen in Figure 3d. Due to the difference of RoomCL for different seasons, the analysis is performed on the overall data, and each data subset is related to a specific season, respectively.

4.2. Pairwise Scatterplots and Variable Transformation Checking

The next phase in EDA is to find patterns and correlations between variables. The focus is on two groups of data: the all-seasons data, containing all the four seasons' data together, and the seasonal data, containing the data for each season separately. In Figure 4, a correlation plot for the all-seasons dataset is shown. The name of all variables and their distribution plots are given in diagonally. The upper triangle displays the Spearman rank correlation coefficient [28] and the statistical significance of the correlation between variables. The degree of significance between variables is indicated by the red asterisks with higher degree represents by three asterisks. In a nutshell, the p -value expresses the probability that no true relationship exists between the variables. We can notice that the RoomCL, which is the dependent variable, correlates with all the predictors except the WorkDay and the InTemp. The lower triangle of the subplots displays the pairwise scatterplots of the variables along with an average curve to reveal the nonlinear dependencies. The effect of the InTemp is inexistent for all the variables. By comparing the correlations between the independent variables, the ones with high correlations with the RoomCL are selected over the ones with low correlation for two correlated independent variables. We can notice that the independent variables are correlated. By defining high correlation between two independent variables being ≥ 0.50 , we can consider RoofCL with RoofIrrad, SATemp, WindowCL, WallCL, WallTemp, and OutTemp; RoofIrrad with SATemp, WindowCL, SurfIrrad, and WallTemp; SATemp with WindowCL, SurfIrrad, WallTemp, and OutTemp; WindowCL with SurfIrrad, WallTemp, and OutTemp; WallCL with WallTemp and OutTemp; SurfIrrad with WallTemp; WallTemp with OutTemp; ITCL

with LightCL, OccupCL, WorkDay, NoOcc, and NoM; LightCL with OccupCL, WorkDay, NoOcc, and NoM; OccupCL with WorkDay, NoOcc, and NoM; NoOcc with NoF and NoM; and NoF with NoM, as seen in Figure 4. The related key patterns for linearly connected variables and nonlinearly connected variables are correlated with each other. This leads to the collinearity problems.

In summary, the following results are obtained from the exploratory data analysis:

The RoomCL variation is detected between 6 a.m. and 8 p.m. on a daily basis and during working hours.

The high record of the RoomCL load is on workdays (Monday to Friday).

For monthly based, the high records of the RoomCL are from October to March; this corresponds to summer and spring seasons.

These results lead to the upcoming statistical analysis.

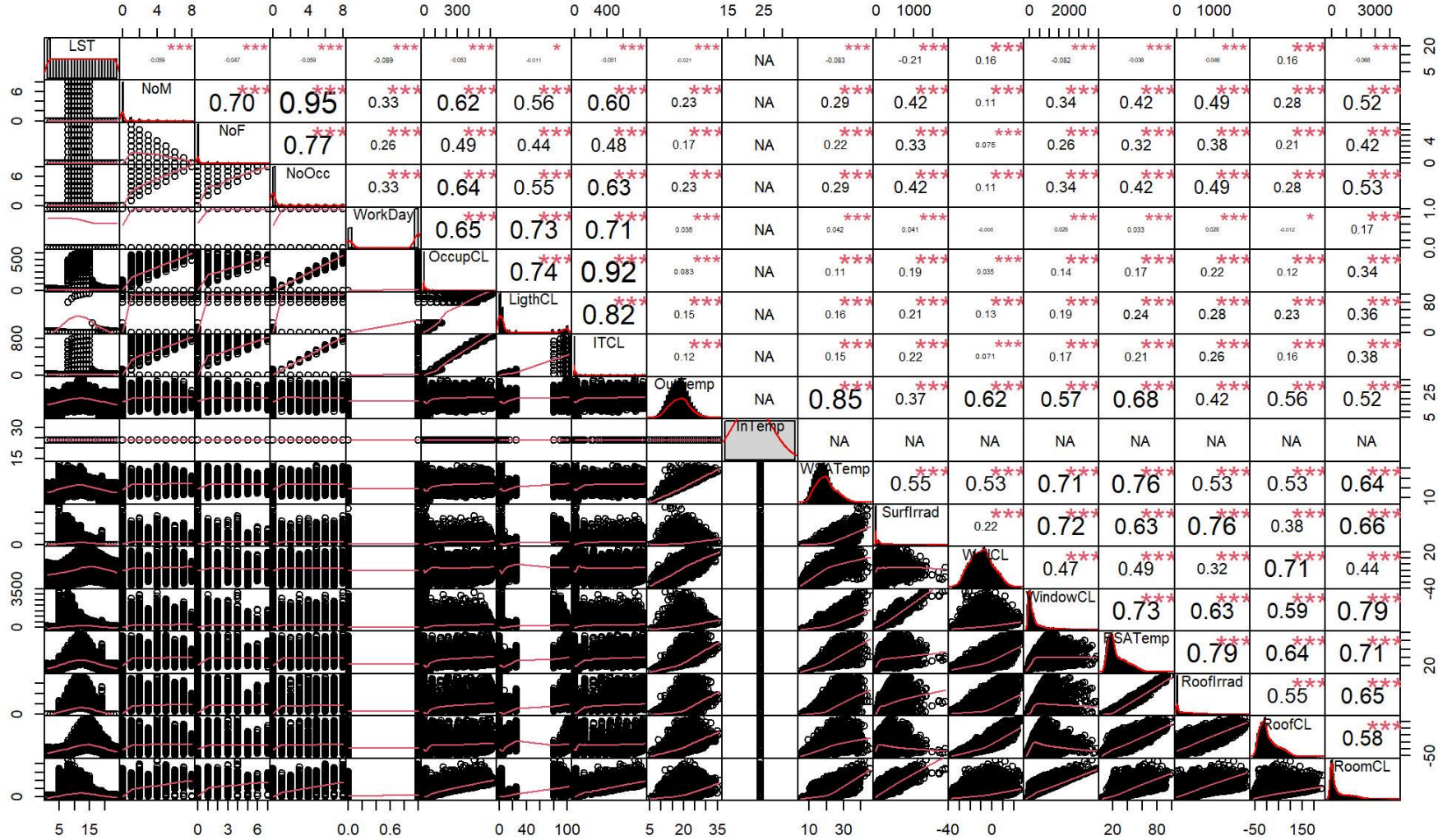


Figure 4. Pairwise correlation diagram for all-seasons dataset from 1 January 2015 to 31 December 2017.

5. Statistical Analysis and Regression Model from All-Seasons Dataset

Statistical models are developed in this section by inputting data into a stepwise function. Following that, the stepwise function fits regression models to training data. Finally, the constructed regression models' reliability are checked using the testing dataset.

5.1. Stepwise Selection: Fitting Linear Models

The data time, time, day, month, and season variables are removed from the dataset for being factor variables and not numerically contributing to the RoomCL, and the all-seasons data subsets are split into two data subsets. A total of 70% of the data are for training and the rest are for testing. InTemp value is constant and removed for not contributing or correlating with any variable in the dataset, as specified in Section 4.2. The model containing all the remaining variables is considered. Then, the forward elimination of predictors not contributing to the model is applied to the training subset. The best model is chosen by performing a number of statistics and diagnostic tests drawn to evaluate the linear regression model beyond EDA. The models with a high adjusted coefficient of determination are chosen. After performing the forward stepwise selection and considering predictors correlated to the dependent variable, the variance inflation factors (VIF) are determined for different predictors in the obtained models and are given in Table 1. This allows to eliminate correlated predictors and verify their independency. In each model, only the predictors with $VIF < 10$ are considered. Furthermore, the linearity, homogeneity of variance, influential observations, and the normality of residuals are tested. From the selected models, the optimal model is chosen by considering (i) the number of predictors, (ii) AIC, BIC, and RMSE values, and (iii) the adjusted R^2 values given in Table 2. By examining Table 1, it is observed that NoM, NoF, WindowCL, and RoofCL are predictors in all the five models. Because of the number of predictors, the collinearity problems should be checked.

Table 1. Variance inflation factors (VIF) for selected model predictors.

Models	No. Pred. *	Predictor's VIFs							
		LST	NoM	NoF	WSA Temp	WallCL	WinCL	RSA Temp	RoofCL
1	7	1.245	1.441	1.276	-	4.100	1.929	6.662	8.517
2	7	1.285	1.382	1.266	5.248	5.530	2.590	-	4.071
3	6	1.156	1.352	1.245	-	3.965	1.301	-	4.071
4	5	1.126	1.308	1.229	-	-	1.245	-	1.323
5	4	-	1.301	1.227	-	-	1.172	-	1.290

* Number of predictors.

Table 2. Statistic parameters for all-seasons selected models.

Models	No. Pred.	AIC	BIC	R^2	R^2 (Adj.)	RMSE	p -Value	Perf. Score * (%)
1	7	215,137	215,208	0.9828	0.9828	83.17	$<2.2 \times 10^{-16}$	100
2	7	215,169	215,239	0.9827	0.9827	83.24	$<2.2 \times 10^{-16}$	42.2
3	6	215,222	215,284	0.9827	0.9827	83.36	$<2.2 \times 10^{-16}$	0.18
4	5	215,220	215,275	0.9827	0.9827	83.36	$<2.2 \times 10^{-16}$	0.34
5	4	215,218	215,265	0.9827	0.9827	83.36	$<2.2 \times 10^{-16}$	0.60

* Performance score.

From Table 1, only model 3, model 4, and model 5 present $VIF < 5$ for their predictors, making them the potential models in term of variables' independence. We can also notice from Table 2 that model 5 has a lower number of predictors. In addition, summary statistics for AIC, BIC, R^2 , R^2 (Adj.), RMSE estimates, p -values, and performance score for all the all-seasons selected models obtained by the forward step method are given in Table 2. Both potential models present almost the same AIC, BIC, R^2 , and R^2 (Adj.). The AIC and BIC of these models are higher and the RMSE are lower than the values obtained for model 1 and model 2. In contrast, model 1 and model 2 have the lower AIC and BIC values, and their performance scores are much higher. The R^2 and R^2 (Adj.) are almost equal for the five models.

Model 5 is selected as the optimal model and is in the group of models with higher AIC and BIC with a performance score of 0.60% which is below the 10% recommended for linear regression models. It has a minimum number of predictors for the same R^2 and R^2 (Adj.) of 0.983 compared to other models.

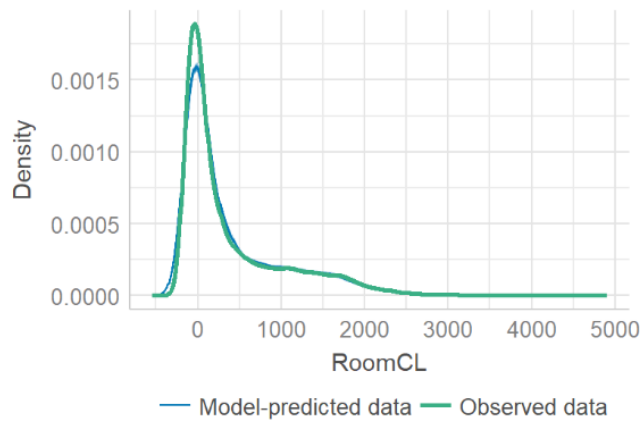
The visual check of model's assumptions for the all-seasons dataset is shown in Figure 5. This figure presents the diagnostic plots of residuals and the correlations between predictors for the all-seasons model. The linear relationship assumptions of the model are checked through residuals vs. fitted values. We observe a horizontal line without distinct patterns, which is an indication of a linear relationship. The homogeneity of variance of the residuals, also called homoscedasticity, is checked through the standardised residuals square vs. fitted residuals. A good indication of homoscedasticity is when a horizontal line with equally spread points is obtained; this is the case for the chosen model. The collinearity of variables is checked through the VIF vs. the predictors. The VIF of a predictor is low when it is <5 , and the VIF of a predictor is moderate when it is <10 . In both cases, the assumptions of the model are good because the VIFs are <5 , which is good for the model. In addition, the influential cases of extreme values that might influence the regression results when included or excluded from the analysis are checked through the standardised residuals vs. leverage plots. In this case, the Cook's distance of each point is calculated, and the point must be inside the contour lines defined by the green dashed lines.

This is fulfilled by model 5. The normal distribution of residuals is examined through the normal Q-Q plots. It is good if the residuals points follow the straight green line. In the present study, most of the points are on the green line, which is a good indication of the normality of distribution of residuals even if the distribution curve presents a certain skewness. Furthermore, the data not following the requirements, meaning the points out of or away from the reference lines, were checked to ensure that they will not affect the model's performance. These points are not outliers or due to RTS method errors. On the contrary, they represent the rare cases where the maximum number of occupancies is reached for specific sunny days, especially in summer. Furthermore, from the posterior predictive check, the model-predicted lines resemble the observed data line; this is confirmed in Figure 6.

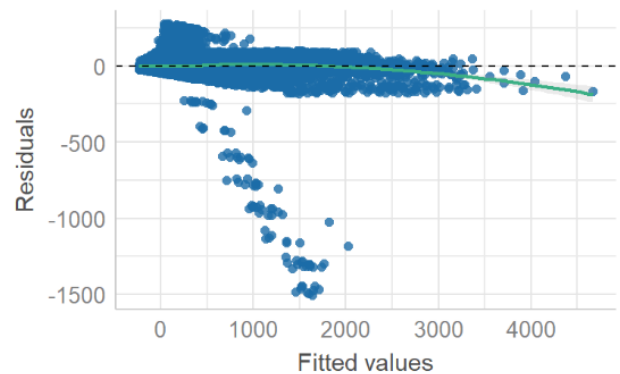
The prediction obtained from the selected all-seasons model compared to the test data in Figure 6 presents a linear relation with an adjusted R^2 of 0.98 (close to 1) with a positive slope. This shows how good the selected model is.

(a) Posterior Predictive Check

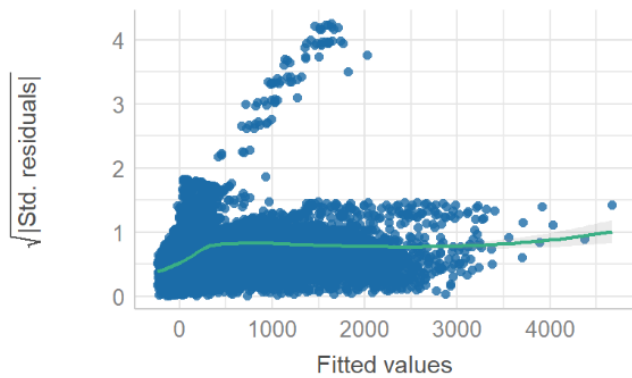
Model-predicted lines should resemble observed data line

**(b) Linearity**

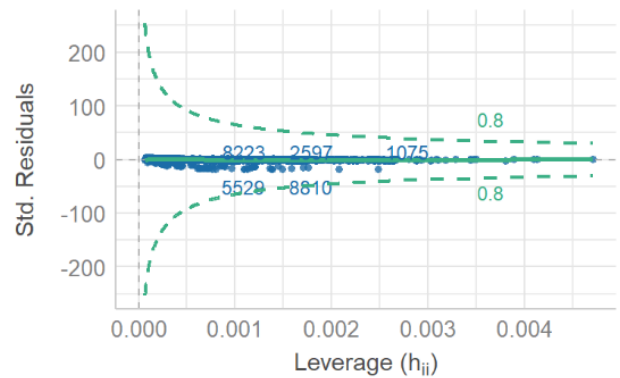
Reference line should be flat and horizontal

**(c) Homogeneity of Variance**

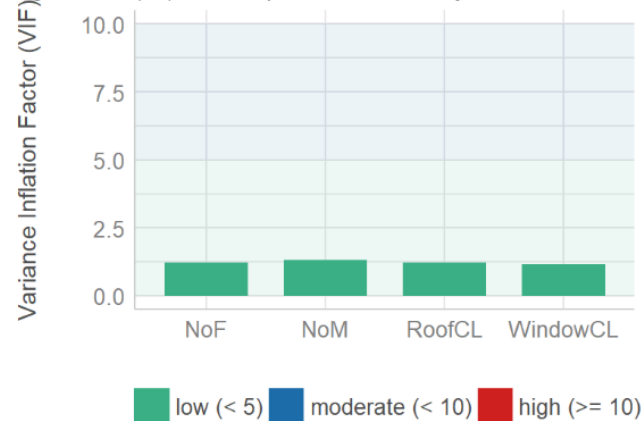
Reference line should be flat and horizontal

**(d) Influential Observations**

Points should be inside the contour lines

**(e) Collinearity**

Higher bars (>5) indicate potential collinearity issues

**(f) Normality of Residuals**

Dots should fall along the line

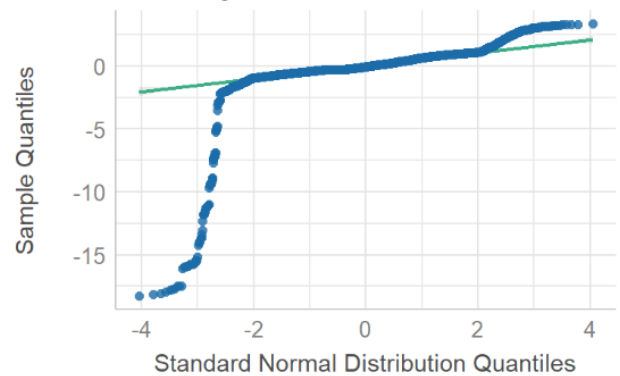


Figure 5. Visual check of model's assumptions for all-seasons dataset: (a) posterior predictive check; (b) linearity; (c) homogeneity of variance; (d) influential observations; (e) collinearity; (f) normality of residuals.

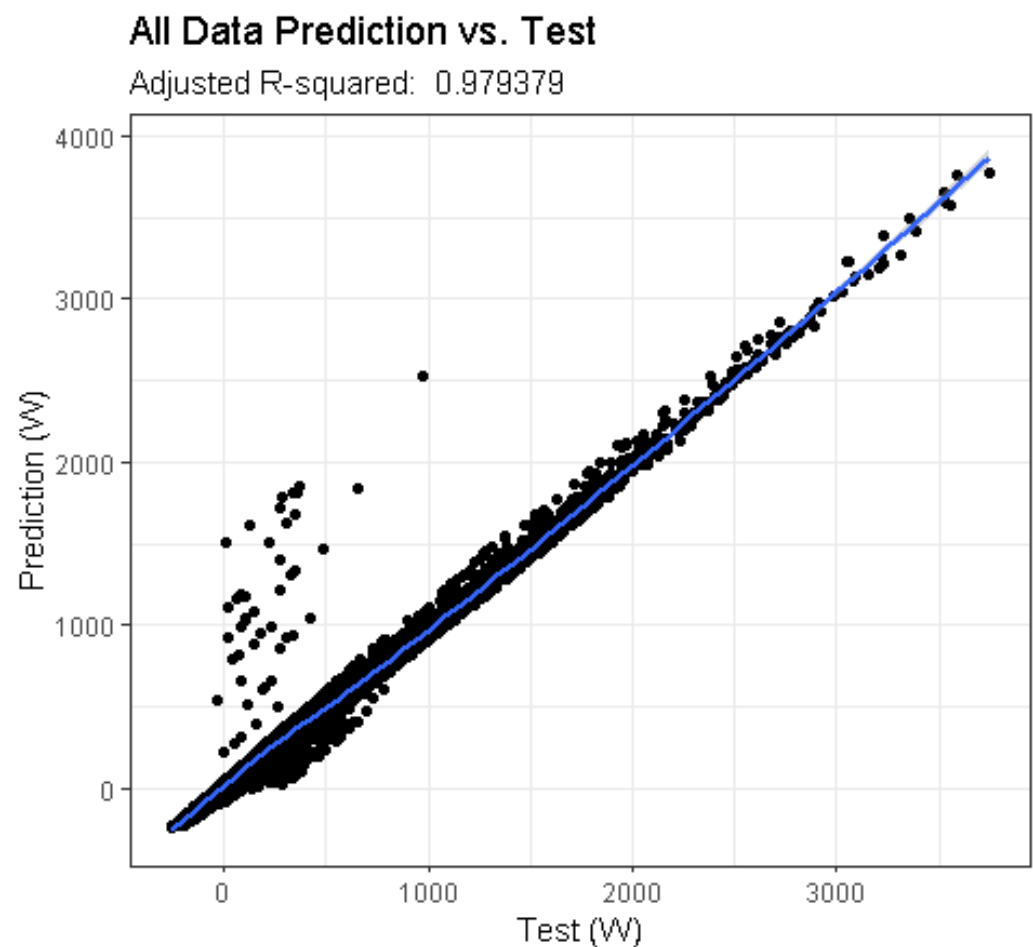


Figure 6. Comparison of the prediction and the test results for a selected model of all seasons.

5.2. Compilation of Optimal Regression Models

The procedure used to determine the all-seasons model was also used to determine the four seasonal models. Thus, five models are produced: one model corresponding to all datasets and four models corresponding to seasonal datasets for the selected predictors. These models are calibrated with data from the training dataset. The predictors are selected from statistical analysis of different model predictors' parameters based on the model diagnostics and performance scores from training data, and the results are obtained from testing datasets. The occupancy depends on the day of the week and hours of the day. It is multiplied by the values 0 (no work hours) and 1 (work hours). This allows determining NoM and NoF predictors.

The next step is to test and compare different regression models obtained from all-seasons and different-season datasets related to the RoomCL. The choice of one-hour time resolution is due to the radiant factors used to determine 24 h cooling load in the RTS method. The visual check of the four seasonal models assumptions and the comparison of the prediction and the test results are presented in Figures 7 and 8, respectively. The observation is that the results obtained in Figures 5 and 6 are also obtained in Figures 7 and 8, respectively.

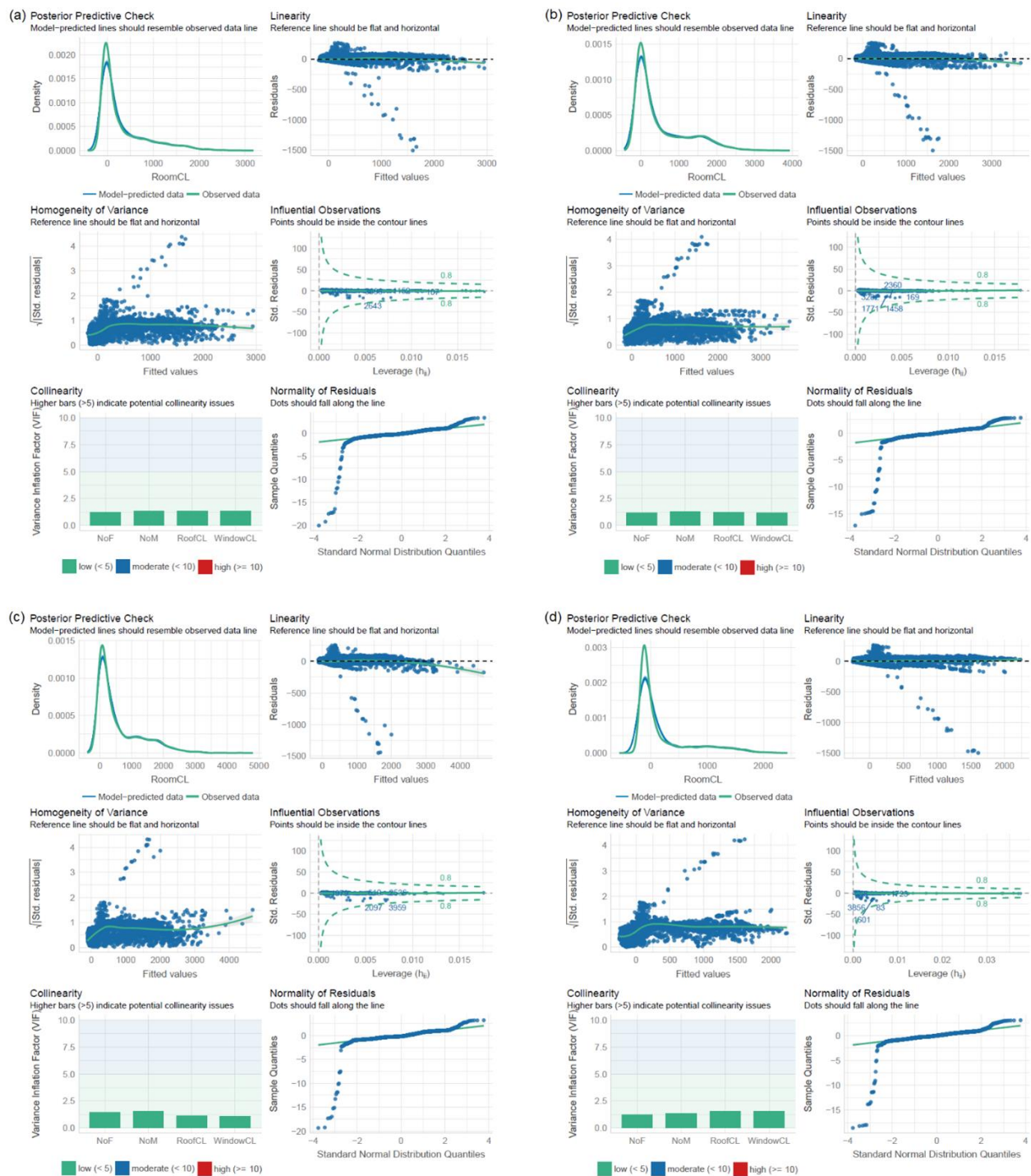


Figure 7. Visual check of models' assumptions for the four seasons of the year: (a) autumn season; (b) spring season; (c) summer season; (d) winter season.

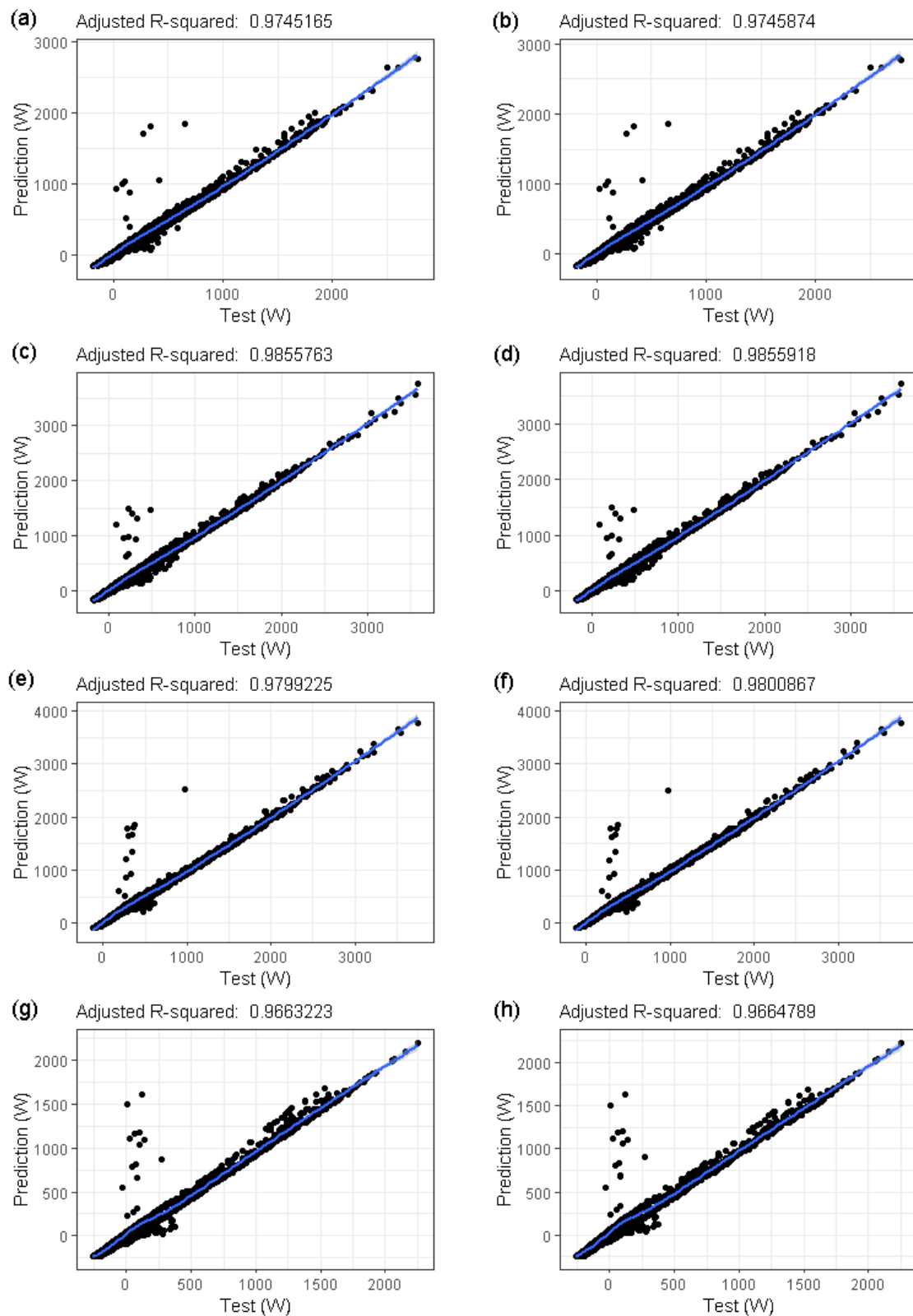


Figure 8. Comparison of the prediction and the test results for selected models of different seasons: (a) autumn—prediction vs. test for all data; (b) autumn—prediction vs. test for autumn data; (c) spring—prediction vs. test for all data; (d) spring—prediction vs. test for spring data; (e) summer—prediction vs. test for all data; (f) summer—prediction vs. test for summer data; (g) winter—prediction vs. test for all data; (h) winter—prediction vs. test for winter data.

The predictors selected for the all-seasons dataset best model are used to determine the RoomCL model of different seasons, and the results are given in Table 3.

Table 3. Predictor's coefficients for different selected models.

Models	Intercept	NoM	NoF	WindowCL	RoofCL
All seasons	7.149×10^{-1}	4.950×10^{-1}	6.211×10^{-1}	1.743×10^{-3}	1.668×10^{-2}
Autumn	13.47	1.034	1.232	5.601×10^{-3}	3.878×10^{-2}
Spring	1.606	9.421×10^{-1}	1.124	3.681×10^{-3}	3.312×10^{-2}
Summer	1.619	1.136	1.577	2.308×10^{-3}	2.956×10^{-2}
Winter	1.842	9.868×10^{-1}	1.224	8.978×10^{-3}	6.324×10^{-2}

By comparing the coefficients of the obtained models for different seasons to the obtained model with all data, we can notice that the coefficients are close to each other, except for the coefficients of the winter season. The models' performance analysis results considering AIC, BIC, R^2 , R^2 (Adj.), RMSE, and sigma are given in Table 4.

Table 4. Results of the performance analysis for selected regression models.

Model	AIC	BIC	R^2	R^2 (Adj.)	RMSE	Sigma
All seasons	2.152×10^5	2.153×10^5	0.983	0.983	83.349	83.360
Autumn	53,471.736	53,510.365	0.978	0.978	78.787	78.830
Spring	54,163.473	54,202.044	0.983	0.983	89.834	89.883
Summer	52,472.254	52,510.778	0.988	0.988	78.149	78.192
Winter	54,807.873	54,846.583	0.973	0.973	84.063	84.108

For the obtained models, it is observed that low RMSEs are obtained in summer and autumn seasons, while high RMSEs are obtained in spring and winter seasons, referring to all data models. For each obtained model, the RoomCL prediction and RoomCL test results are compared and the corresponding R^2 (Adj.) determined. We can notice a strong correlation between prediction and test in both cases, and high R^2 (Adj.) are obtained.

The relative errors between R^2 (Adj.) coefficients from the test results comparisons are given in Table 5. They are almost zero for all models related to different seasons.

Table 5. Comparison of actual and expected R^2 (Adj.) for the selected RoomCL models.

Model	R^2 (Adj.)		
	Actual (All Seasons)	Expected (Actual Season)	Relative Error (%)
Autumn	0.9745165	0.9745874	0.0073
Spring	0.9855763	0.9855918	0.0016
Summer	0.9799225	0.9800867	0.0168
Winter	0.9663223	0.9664789	0.0162

6. Discussion

The most influential predictors used for the all-seasons data selected model are also used in the models of different seasons. The use of the most influential predictors to build the model leads to a model with lower coefficient of correlation and higher residual standard error. These results are due to the elimination of certain predictors that have a significant correlation with the dependent variable, which in this case is the RoomCL, due to certain statistics rules and hypotheses. Nevertheless, the difference between the

models with high number of predictors satisfying statistic requirements and models with low number of influential predictors is small. This indicates the use of simplified prediction models with few predictors included, as these exhibit an acceptable correlation coefficient. We can conclude that the variables NoM, NoF, WindowCL, and RoofCL are the highest ranked predictors for the linear regression model of the RoomCL for the research room of the new engineering building of MUT since they give the rates of smallest errors in the prediction evaluation. Here, NoM and NoF constituted the occupancy of the room and are closely related to the hours, the WorkDay, the LightCL, and the ITCL. This demonstrates how individual presence is required [34] for heat production and its impact on the cooling load, and it is an important aspect to include in such an analysis. The RoomCL presents a nonlinear relationship with the outdoor temperature, as highlighted by many studies [22]. The contribution of solar irradiance through the wall is much smaller compared to the window; at the same time, the solar light can be used to light the room. The windows-to-wall ratio should be optimised to minimise both heating and cooling loads, as well as the total energy consumed by lighting while maintaining a suitable lighting level for occupants [35]. The work hour variable is also an important predictor because it deals with daily work schedules, but it is substantially connected with NoM and NoF. The error decreases if more data are employed to build the predictions. For example, the error is reduced by 19% for the model built with seven predictors compared to the model built with four predictors for the RoomCL. These results confirm that if the influential predictors are left out of the model, this can have a significant impact on the performance of the model, as the model with lower predictors generally has a higher error rate and induces inaccurate prediction. This demonstrates the difficulty in developing a model with an accurate RoomCL prediction with small number of predictors from low number of data. Moreover, the obtained residual errors are 0.83% for all seasons, 0.79% for the autumn season, 0.90% for the spring season, 0.78% for the summer season, and 0.84% for the winter season. From the results obtained in Table 5, we can conclude that for a large-sized dataset, the all-seasons model can represent different seasons of the year to analyse and predict the RoomCL.

Incorporating other components or related parameters intervening in the calculation of cooling load using RTS method, e.g., building radiant factors, building georeferenced coordinates, etc., can increase the model prediction performance. In addition, the NoM and NoF can be replaced by occupancy, and other influential predictors can be added instead to enlarge the nature and the impact of predictors in the model. However, this will impose new constraints on how the prediction model can be validated. Moreover, as all the parameters that could affect the construction and accuracy of the prediction model are not considered, such as the behavioural attributes of the occupants, the model obtained will influence the prediction results [36]. These data could inform how occupants interact with computer systems and the light in the room. However, it is important to not include these factors in the models because the factors such as occupant interactions with computer systems and light are inaccurate to predict, the forecasts will therefore be more uncertain, and the error rates could be higher. Furthermore, an increase in the number of predictors will increase the demand for input data that may be difficult to obtain or may even not be available. Building a model with small dataset and high number of predictors leads to a model with low performance, and the correlation between predictors could increase, leading to high values of VIF for predictors. In addition, the Köppen climate classification [37] can be incorporated in the determination of mutation or special conditions due to the climate that may affect the cooling load, and lead to uncertainty or unpredictable field operation.

Thereby, it is possible to predict the RoomCL with a limited number of predictors with respect to NoM, NoF, WindowCL, and RoofCL using the coefficients obtained from all data. Because WindowCL and RoofCL can be determined by calculation, the RoomCL can be determined by anticipating the NoM and the NoF in the room for a specific day. The model can be used for prediction of the RoomCL and could contribute significantly to the building of an accurate algorithm for the optimal control of the HVDC system for

the rooms and buildings by measuring the values of WinCL and RoofCL using sensors. It should be noted that the model was created for a research lab in the new engineering building at MUT in South Africa. Obviously, this raises concerns about the results' general applicability. The methodological approach presented in this article is supposed to be applicable to all rooms and buildings if they present similarities. If similar data for a specific room or even a building can be collected, the present analysis steps can be applied to this data to derive correlation models and regression models used for the prediction of the cooling load of these rooms and buildings. The present correlation model can be used for buildings situated in similar climatic conditions and usage features to the new MUT engineering building room. However, to be sure about such statements, additional validation on datasets from other similar buildings is required. This is accomplished by considering the building materials and usage, which determine the RTS factors. As a result, the model coefficients must be recalculated based on the data from these new research rooms and buildings.

7. Conclusions

In this paper, the linear multiregression model for the prediction of a room cooling load based on the radiant time series method components for a renewable energy research room at Mangosuthu University of Technology (MUT) was presented. The model for the room was developed with male and female occupants, window cooling load, and roof cooling load as the most influential predictors. To obtain the adequate model:

The exploratory data analysis was performed to pool the data, then statistical regression models with the most influential predictors in a stepwise function were obtained from all-seasons data and the seasonal data. The obtained seasonal models were compared to the performance of all-seasons models using the testing dataset. The exploratory analysis led us to split the data into autumn, spring, summer, and winter seasons when differences in temperature and insolation were remarkable.

Further, a non-influence of the indoor temperature for both comfort and efficiency of the research room was noticed. Male and female occupants, window cooling load, and roof cooling load were the most significant predictors for the room cooling load.

However, AIC, BIC, and RMSE measurements designated models with more predictors as optimal; this is often the case. The predictions with models with fewer predictors have been shown to result in higher error rates. This was due to the elimination of certain predictors correlated to RoomCL.

The obtained relative errors between the best all-seasons model and each seasonal model built with the same predictors for the respective data subsets are almost zero and are given by 0.0073% (autumn), 0.0016% (spring), 0.0168% (summer), and 0.0162% (winter).

The conclusion extracted from these results is that the all-seasons model can be used for any season to determine the room cooling load for any room in the specific conditions with MUT room for a large dataset. In addition, the obtained model allows to determine or evaluate the cooling load from a low number of predictors at the early stage of building without measurement.

Author Contributions: Methodology, N.M.-A.M.; software, N.M.-A.M.; writing—original draft, N.M.-A.M.; review, B.P.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data related to this study can be obtained by email to mutombo.marc-alain@mut.ac.za or numbib@mut.ac.za.

Acknowledgments: The authors would like to thank the Mangosuthu University of Technology (MUT) for financial support through their postdoctoral program. We would also like to thank the anonymous reviewers for their valuable comments.

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

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
ANN	Artificial neural network
ARIMA	Autoregressive inference moving average
ASHREA	American Society of Heat and Refrigeration Engineering Association
BIC	Bayesian information criterion
DR	Demand–response
EDA	Exploratory data analysis
HVAC	Heating, ventilation, and air conditioning
IEETR	Industrial Energy Efficiency Training and Resource
InTemp	Indoor temperature
IT	Information and technology
ITCL	Information and technology cooling load
LightingCL	Lighting cooling load
MUT	Mangosuthu University of Technology
NoF	Number of females
NoM	Number of males
NoOcc	Number of occupants
OccupCL	Occupancy cooling load
OutTemp	Outdoor temperature
RMSE:	Root mean squared error
RoofCL	Roof cooling load
RoofIrrad	Roof irradiance
RoomCL	Room cooling load
RSATemp	Roof sol-air temperature
RTS	Radiant time series
SA	South Africa
SATemp	Surface area temperature
SurfIrrad	Total surface irradiance
VIF	Variance inflation factors
WallCL	Wall cooling load
WallTemp	Wall temperature
WindowCL	Window cooling load
WorkDay	Working days
WSATemp	Wall sol-air temperature

References

1. Congedo, P.M.; Baglivo, C.; Seyhan, A.K.; Marchetti, R. Worldwide dynamic predictive analysis of building performance under long-term climate change conditions. *J. Build. Eng.* **2021**, *42*, 103057. [\[CrossRef\]](#)
2. Lee, E.S.; Fernandes, L.L.; Goudey, C.H.; Jonsson, C.J.; Curcija, D.C.; Pang, X.; DiBartolomeo, D.; Hoffmann, S. *A Pilot Demonstration of Electrochromic and Thermochromic Windows in the Denver Federal Center, Building 41, Denver, Colorado*; Energy Technologies Area Lawrence Berkeley National Laboratory, University of California: Berkeley, CA, USA, 2013.
3. King, J.; Perry, C. *Smart Buildings: Using Smart Technology to Save Energy in Existing Buildings*; American Council for an Energy-Efficient Economy: Washington, DC, USA, 2017.
4. Yang, S.; Wan, M.P.; Ng, B.F.; Dubey, S.; Henze, G.P.; Chen, W.; Baskaran, K. Model predictive control for integrated control of air-conditioning and mechanical ventilation, lighting and shading systems. *Appl. Energy* **2021**, *297*, 117112. [\[CrossRef\]](#)
5. Sandels, C.; Brodén, D.; Widén, J.; Nordström, L.; Andersson, E. Modeling office building consumer load with a combined physical and behavioral approach: Simulation and validation. *Appl. Energy* **2016**, *162*, 472–485. [\[CrossRef\]](#)
6. Han, F.; Liu, B.; Wang, Y.; Dermentzis, G.; Cao, X.; Zhao, L.; Pfluger, R.; Feist, W. Verifying of the feasibility and energy efficiency of the largest certified passive house office building in China: A three-year performance monitoring study. *J. Build. Eng.* **2021**, *46*, 103703. [\[CrossRef\]](#)
7. Wang, Y.; Kuckelkorn, J.M.; Zhao, F.-Y.; Liu, D.; Kirschbaum, A.; Zhang, J.-L. Evaluation on classroom thermal comfort and energy performance of passive school building by optimizing HVAC control systems. *Build. Environ.* **2015**, *89*, 86–106. [\[CrossRef\]](#)

8. Kuhn, T.E.; Bühler, C.; Platzer, W.J. Evaluation of overheating protection with sun-shading systems. *Sol. Energy* **2001**, *69*, 59–74. [\[CrossRef\]](#)
9. Haase, M.; Amato, A. An investigation of the potential for natural ventilation and building orientation to achieve thermal comfort in warm and humid climates. *Sol. Energy* **2009**, *83*, 389–399. [\[CrossRef\]](#)
10. Yang, S.; Wan, M.P.; Chen, W.; Ng, B.F.; Dubey, S. Experiment study of machine-learning-based approximate model predictive control for energy-efficient building control. *Appl. Energy* **2021**, *288*, 116648. [\[CrossRef\]](#)
11. Hellberg, C. Förbättrad energistatistik för lokaler—Stegvis STIL—Rapport för år 1—Inventering av kontor och förvaltningsbyggnader. *Statens Energimyndighet Rapp. ER* **2007**, *34*, 1–92.
12. Oldewurtel, F.; Sturzenegger, D.; Morari, M. Importance of occupancy information for building climate control. *Appl. Energy* **2013**, *101*, 521–532. [\[CrossRef\]](#)
13. Sandels, C.; Widén, J.; Nordström, L.; Andersson, E. Day-Ahead Predictions of Electricity Consumption in a Swedish Office Building from Weather, Occupancy, and Temporal data. *Energy Build.* **2015**, *108*, 279–290. [\[CrossRef\]](#)
14. Zhao, J.; Lasternas, B.; Lam, K.P.; Yun, R.; Loftness, V. Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. *Energy Build.* **2014**, *82*, 341–355. [\[CrossRef\]](#)
15. Kottagoda, N.T.; Rosso, R. *Applied Statistics for Civil and Environmental Engineers*; Wiley-Blackwell: Oxford, UK, 2008.
16. Menapace, A.; Santopietro, S.; Gargano, R.; Righetti, M. Stochastic Generation of District Heat Load. *Energies* **2021**, *17*, 5344. [\[CrossRef\]](#)
17. Hora, C.; Dan, F.C.; Bendea, G.; Secui, C. Residential Short-Term Load Forecasting during Atypical Consumption Behavior. *Energies* **2022**, *15*, 291. [\[CrossRef\]](#)
18. Aman, S.; Frincu, M.; Charalampos, C.; Noor, U.; Simmhan, Y.; Prasanna, V. *Empirical Comparison of Prediction Methods for Electricity Consumption Forecasting*; Tech. Rep.; University of Southern California: Los Angeles, CA, USA, 2014; pp. 14–942.
19. Hagan, M.T.; Behr, S.M. The Time Series Approach to Short-Term Load Forecasting. *IEEE Power Eng. Rev.* **1987**, *PER-7*, 56–57. [\[CrossRef\]](#)
20. Amjady, N. Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *IEEE Trans. Power Syst.* **2001**, *16*, 498–505. [\[CrossRef\]](#)
21. Tso, G.K.F.; Yau, K.K.W. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* **2007**, *32*, 1761–1768. [\[CrossRef\]](#)
22. Hong, T.; Gui, M.; Baran, M.E.; Willis, H.L. Modeling and forecasting hourly electric load by multiple linear regression with interactions. In Proceedings of the IEEE PES General Meeting, Minneapolis, MN, USA, 25–29 July 2010; pp. 1–8.
23. Fan, S.; Hyndman, R. Short-Term Load Forecasting Based on a Semi-Parametric Additive Model. *IEEE Trans. Power Syst.* **2011**, *27*, 134–141. [\[CrossRef\]](#)
24. Aman, S.; Simmhan, Y.; Prasanna, V.K. Improving energy use forecast for campus micro-grids using indirect indicators. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 11 December 2011; pp. 389–397.
25. Hippert, H.S.; Pedreira, C.E.; Souza, R.C. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Trans. Power Syst.* **2001**, *16*, 44–55. [\[CrossRef\]](#)
26. Djuric, N.; Novakovic, V. Identifying important variables of energy use in low energy office building by using multivariate analysis. *Energy Build.* **2012**, *45*, 91–98. [\[CrossRef\]](#)
27. Lam, J.C.; Wan, K.K.W.; Cheung, K.L.; Yang, L. Principal component analysis of electricity use in office buildings. *Energy Build.* **2008**, *40*, 828–836. [\[CrossRef\]](#)
28. Warner, R.M. *Applied Statistics: From Bivariate Through Multivariate Techniques*; Sage Publications: Southend Oaks, CA, USA, 2007.
29. Sandels, C.; Widén, J.; Nordström, L. Forecasting household consumer electricity load profiles with a combined physical and behavioral approach. *Appl. Energy* **2014**, *131*, 267–278. [\[CrossRef\]](#)
30. Onaolapo, A.K.; Carpanen, R.P.; Dorrell, D.G.; Ojo, E.E. A Comparative Assessment of Conventional and Artificial Neural Networks Methods for Electricity Outage Forecasting. *Energies* **2022**, *15*, 511. [\[CrossRef\]](#)
31. Wu, L.; You, S.; Dong, J.; Liu, Y.; Bilke, T. Multiple Linear Regression Based Disturbance Magnitude Estimations for Bulk Power Systems. In Proceedings of the 2018 IEEE Power & Energy Society General Meeting (PESGM), Portland, OR, USA, 5–10 August 2018; pp. 1–5.
32. Witten, D. *An Introduction to Statistical Learning: With Applications in R*; Springer: Berlin/Heidelberg, Germany, 2013.
33. Gelman, A.; Su, Y.-S. *Data Analysis Using Regression and Multilevel/Hierarchical Models*; Cambridge University Press: Cambridge, UK, 2006.
34. Page, J.; Robinson, D.; Morel, N.; Scartezzini, J.-L. A generalised stochastic model for the simulation of occupant presence. *Energy Build.* **2007**, *40*, 83–98. [\[CrossRef\]](#)
35. Albatayneh, A.; Juaidi, A.; Abdallah, R.; Manzano-Agugliaro, F. Influence of the Advancement in the LED Lighting Technologies on the Optimum Windows-to-Wall Ratio of Jordanians Residential Buildings. *Energies* **2021**, *14*, 5446. [\[CrossRef\]](#)
36. Widén, J.; Lundh, M.; Vassileva, I.; Dahlquist, E.; Ellegård, K.; Wäckelgård, E. Constructing load profiles for household electricity and hot water from time-use data—Modelling approach and validation. *Energy Build.* **2009**, *41*, 753–768. [\[CrossRef\]](#)
37. Chen, D.; Chen, H.W. Using the Köppen classification to quantify climate variation and change: An example for 1901–2010. *Environ. Dev.* **2013**, *6*, 69–79. [\[CrossRef\]](#)