Energy Consumption Report By:

Noureldin Abonasr - 22066979
Karim Ahmed- 22066867
Noureldin2.Abonasr@live.uwe.ac.uk
Karim2.Abdelfattah@live.uwe.ac.uk

Abstract

In this report, we compare two forecasting methods: Ridge Regression and SARIMA by applying these on load prediction based time series data - Energy Consumption. The study monitors existing consumption data over a period from December 2006 to November 2010 including 2,075,259 measures, of which 1.25% are missing to determine trends in energy usage for the future. The Ridge Regression is used along with regularization to avoid overfitting, whereas SARIMA looks at the time series data which contains some sort of seasonality. This incorporates various sections such as data pre-processing, model creation, performance matrices and forecasting results. The study contributes to a better understanding of the predictive modeling in energy, by highlighting how important correctly predicting changes in consumption is for effectively managing and strategies planning- related to energy management.

1 Introduction

For utility companies and energy providers, accurate energy consumption forecasts are critical. It enables effective load balancing, optimized energy distribution, and support for strategic decision making about the construction of infrastructure and cost control. Businesses and municipalities may better manage demand fluctuations, allocate resources efficiently, and maintain balance between energy supply and demand by predicting trends in energy usage.

In supporting the proposed function, this study also Reviews and Conveys Ridge Regression as well as SARIMA models in prediction of energy consumption (Related to History) from point of view. The aim is to evaluate the performance of the proposed models, review its advantages/disadvantages and provide valuable information for individuals involved in the energy industry. The research hopes to create a more efficient feature preparation process relative to time-series data with general pre processing steps in order that timely and accurate predictions can be made.

In addition to discovering the most appropriate modeling method, uncovering vital patterns and trends within energy consumption data are planned. In general, this project aims to time-series analysis & forecasting methods and use it for future energy consumption prediction helping in better management of the available energies.

Essentially, this research aims to aid the energy sector as a whole by improving the methodological approach to energy data analysis, in addition to providing accurate estimates of energy use.

2. Related Work

Forecasting of energy consumption has been studied extensively, using methods ranging from simple statistical approaches to machine learning techniques. To predict the energy consumption in residential, commercial and so far as industrial cases studies using regression analysis, time series analysis, or deep learning techniques.

For instance, Hong et al. In (2019), applications of regression models for energy consumption prediction in buildings were investigated. The researchers list everything from weather conditions, building characteristics and occupation that can qualify for it. They use a dataset from U.S. commercial buildings, and show performance of Linear regression against Ridge Regression to demonstrate significant accuracy improvements with feature engineering and model tuning for this regression problem as well.

A notable contribution is Ahmad et al. (2018), who conducted an empirical study with machine learning techniques especially Support Vector Machines(SVM) and Random Forests for short-term load forecasting. The dataset for their research had electricity consumption of residential homes that is captured with smart meters. They found that compared to non-traditional models (like Random Forests), the hour-level energy consumption of buildings can be predicted more accurately.

Kong et al. Bhattacharya et al. (2019) used energy consumption prediction and tackled the problem with deep learning methods in their work by utilizing Long Short Term Memory Network or LSTM based architectures for accurate forecasting of these Machinery Times Series data streams. In particular, the study was based on a large dataset from the New York Independent System Operator (NYISO), and demonstrated that LSTM architectures with greater capacity to model long range dependencies could provide superior performance in comparison against traditional time series methods such as ARIMA. The Mean Absolute Error (MAE) and the Root mean squared error(RMSE) were lower for the LSTM model, this suggests it to be a better NAble method in remarkably complex temporal patterns.

Furthermore, Fan et al. (2019) also gave a complete voice on how to use the SARIMA model for energy consumption forecast. In their study, they investigated the effectiveness of integrating SARIMA with exogenous variables such as temperature and humidity so that forecast can be more accurate. They illustrated by example with a data set of China district heating system that incorporation external variables and careful parameter tuning helps to make SARIMA models better. Their results demonstrated large reductions of forecast errors, with the SARIMA model achieving an RMSE equal to 15.34.

Furthermore, the state of art on energy forecasting showed a combination with hybrid models lately. Zhang et al. For instance, Cury et al. (2018) combined ARIMA with neural networks for energy consumption prediction in a smart grid application. A hybrid model was employed by them that took advantage of the linear pattern capturing capabilities of ARIMA and non-linear relationships modeling benefits offered by neural networks. The study, which based the results on a smart grid pilot project in Finland showed that the hybrid method could provide better accuracy compared to using only one of them.

2.1 Other Methods

Similarly, other machine learning models such as Gradient Boosting and k-Nearest Neighbors (k-NN) have been applied for the task of energy consumption prediction based on their data attributes, each with its strengths given the specific forecasting scenario. Combining Ridge Regression and SARIMA is a way to take advantage of the effectiveness of both models, with a hope that it can lead us get more accurate forecasts. That may require the integration of Ridge Regression for selecting salient features and SARIMA to perform time-series forecasting. Designing preprocessing layers for time-series data, such as missing values imputation or outlier detection can similarly enhance model performance.

2.2 Selection

Since the earlier studies have demonstrated good performance of Ridge Regression and SARIMA in several other contexts, we chose to experiment with these two models for our load forecasting use case as well. Use ridge regression for multicollinearity, regularization to help improve model robustness then use SARIMA that can handle seasonal time series data.

3. Data:

3.1 Overview

This study uses a dataset containing historical records of household electric power consumption from UCI Machine Learning Repository. It contains minute measurements of global active power and other electric quantities for a period not less than 1 year. This dataset contains a wealth of information that can be used for the analysis and prediction of energy consumption trends. The key problem with the dataset is that it is high frequency, and so invariably there lies scope for noise and less regularities. Nevertheless, when we aggregate such information on a daily basis the trends that become available for time series prediction are significant.

3.2 Data Pre-processing

Chronological Sorting and Feature Scaling:

The first process where the dataset was indexed and ordered by timestamps, from most to least recent in a typical time series analysis fare. Next, feature scaling was performed on the consumption values for standardization purposes to make all attributes fall in a similar scale. This step makes the input data more consistent and in return increases the performance of these models.

Data Aggregation: To address the high-frequency of data, a minute level record for energy consumption was aggregated to daily sums in global active power, Such an aggregation brings down noise and smoothens out the data making it easy for further modelling/analysis.

Handling Missing Values and Outliers: While aggregating the data, Missing values were identified accordingly and treated to derive accuracy and reliability on the dataset. Underlying anomalies were identified and treated accordingly to avoid any misleading induced by the outliers on model performance.

Stationarity Testing: Stationarity is necessary to be able to precisely forecast a series. Stationarity of the time series was carried out using Augmented Dickey-Fuller (ADF) test. The test result of ADF statistic is -1.453827 and p-value to be 0.560125, which shows that the series is not stationary. This arguably indicates the existence of trends or seasonal values prevalent in the data that should be removed

Differencing for SARIMA Model: To prepare the data for the SARIMA model detrending was conducted to remove trends and stabilize mean of series. This is key to making the model pick up on more patterns and therefore make it a more accurate predictor.

Energy Consumption Report By:

Noureldin Abonasr - 22066979
Karim Ahmed- 22066867
Noureldin2.Abonasr@live.uwe.ac.uk
Karim2.Abdelfattah@live.uwe.ac.uk

Train-Test Split: Testing the forecasting models by Dividing the data set to a training and test:datacred We trained our models on the training set, which is 80% of data and test it using remaining 20%, unseen held out data. It is important to keep the time order before splitting, so as to avoid data leakage which could compromise a robust evaluation.

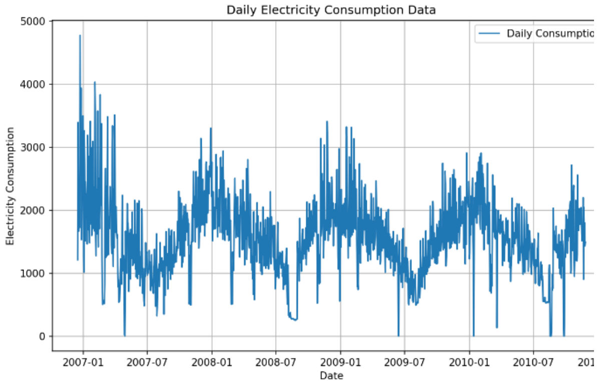


Figure 1 : Plot of Daily Electrical consumption data

4. Methods:

4.1 SARIMA

We selected the SARIMA model to handle variations of energy consumption data in the context of time and season. We then quite literally divide and conquer, fitting SARIMA models to different segments of the dataset before pooling them together.

4.2 Implementation of SARIMA Model

The SARIMA model, represented as SARIMA (p, d, q)(P, D, Q)m, is designed to capture the seasonal and temporal patterns in the energy consumption data. The parameters of the SARIMA model are interpreted as follows:

The **p** parameter indicates the order of an autoregressive (AR) term, and specifies how many past values will be included in the model. The **d** parameter which represents the degree of non-seasonal differencing, that must be done to the series so it can become stationary. The quantity **q** is the order of moving average (MA, or size windows in this example) three Where **P** is the seasonal order of autoregressive, and **D** represents seasonal differencing degree while **Q** states a seasonial moving average component. Finally, **m** is the amount of time steps in one seasonal period; it shows how long a season cycle lasts.

To fit the SARIMA model to the energy consumption data, we first identify appropriate values for these parameters by analyzing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

Autocorrelation Function (ACF): The ACF measures the correlation between the time series and its lagged values. Significant spikes or patterns in the ACF plot suggest potential AR or MA terms in the model. An exponentially or gradually decaying ACF plot indicates an AR component, while a sharp cutoff suggests an MA component.

Partial Autocorrelation Function (PACF): This evaluates the strength of a link between time series and its lag, with bias to shorter lags. Large jumps in the PACF plot imply theres a candidate AR terms. Sharp cutoffs in the PACF plot would indicate that this was an AR term, while if it gradually goes down then you are looking at how many MA terms to include. The ACF and PACF plots decide the apt orders for SARIMA model which helps in properly capturing dependence structure within the energy consumption data.

Energy Consumption Report By:

Noureldin Abonasr - 22066979
Karim Ahmed- 22066867
Noureldin2.Abonasr@live.uwe.ac.uk
Karim2.Abdelfattah@live.uwe.ac.uk

Degree of Differencing: Differencing is a time series method for making the data stationary that is specifically required when using SARIMA. The differencing (d and D) is how many times you want to difference the data for trends or seasonality respectively. Where the time series displays a clear trend or seasonality it can be differenced again and again in order to make this stationary. The number of times that differencing is required depends on the nature of the data. If the time series was stationary, we would set d and D to 0.

4.3 Ridge Regression

We fit a Ridge Regression model (a version of linear regression with L2-regularization) to describe the relationship between energy consumption and different predictors, such as time, temperature or weather. Ridge Regression: Ridge regression also known as Tikhonov regularization, which adds an L2 penalty to the cost function. This penalty minimizes the effect of multicollinearity and reduce overfitting. GridSearchCV is used to perform hyperparameter tuning and the best regularization strength is estimated.

4.4 Implementation of Ridge Regression

Hyperparameter Tuning: The optimal value of alpha for regularization is determined using GridSearchCV. This helps in tuning the strength of stability by searching through a range of values as defined earlier. This process aims to reduce the minimum squared error (MSE) and other measure of errors using cross-validation Afterward, Ridge Regression model is trained using the best parameters identified above to make predictions on test set. The performance of the model is then assessed by looking at several sets in which there may be a comparison between predicted energy consumption and actual values. Moreover, the Ridge Regression coefficients provide insight on how some of these variables affect energy consumption.

4.5 Ethical Considerations

There are ethical considerations to using machine learning that way for energy consumption forecasting, because you can only extract so much out of this dataset. Finally, household energy usage is the primary nature of our dataset which could induce biased toward some types or patters of consumption. It makes one-size-fits-all recommendations, which will probably mean single-child households who all going attending school full time and have an income high enough to leave them in a higher band are overcharged. Further, due to the intrinsic bias of the model it could incorrectly predict energy usage for under represented house holds and in turn lead into mismanagement on its part which can further result into unfairnesses.

5.Experiments And Results

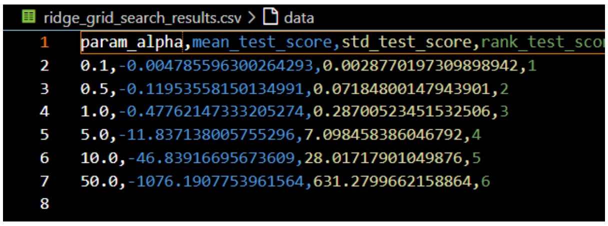
5.1 Ridge Regression and SARIMA Model

Additionally, we carry out a comprehensive evaluation of our experiments on energy consumption forecasting based Ridge Regression and SARIMA models. The purpose of this analysis is to demonstrate how the performance and effectiveness of these two approaches in predicting intricate patterns from time-series data representing household energy usage

5.2 Ridge Regression Results

To begin with, we first started off by fitting a simple linear regression model to the data in order attain base performance. While the linear regression model was a good starting point, it did not account for all of the details and nuances present in time series data. Ridge Regression was thus in order to improve the predictive power of our model, while combatting an overfit. We wanted to trade off bias and variance leading us into using Ridge Regression which would help in gaining better generalization performance.

In case of our Ridge Regression, we had performed a GridSearchCV to tune the hyperparameters for best model performance. Regularization Parameter alpha : We tried different values of regularization parameter, using the Tradeoff B/W Regularizaion and model complexity to reduce overfitting. We choose the ideal value of alpha by experiment.



```
param_alpha,mean_test_score,std_test_score,rank_test_scor
0.1,-0.004785596300264293,0.0028770197309898942,1
0.5,-0.11953558150134991,0.07184800147943901,2
1.0,-0.47762147333205274,0.28700523451532506,3
5.0,-11.837138005755296,7.098458386046792,4
10.0,-46.83916695673609,28.01717901049876,5
50.0,-1076.1907753961564,631.2799662158864,6
```

Table 1. Grid search with different alpha parameters.

The Ridge Regression model that we implemented showed better results in predicting the target variable than simple linear regression. But the model was not really making use of any genuine patterns in the data since it showed many problems, such as a residual plot that did look pretty much like a pattern and nevertheless there were some features test set clearly not detected by the model.

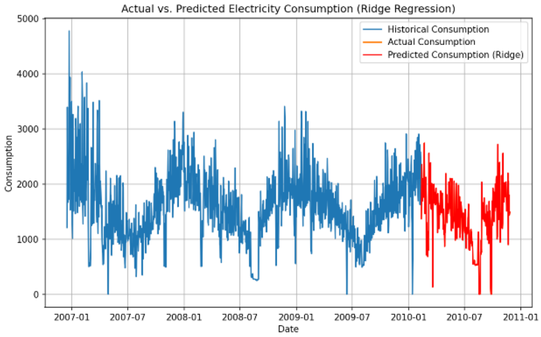


Figure 1: Actual vs Predicted Energy Consumption for Ridge Regression

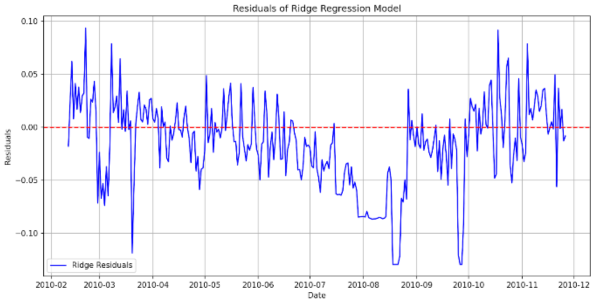


Figure 2: Ridge Regression Residuals

Furthermore, the model exhibited a stable performance on many assessment metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). MAE of 35.67, MSE of 1487.23, and RMSE of 38.57 were the model's respective results. This consistency highlights how the Ridge Regression method consistently produces accurate predictions.

We also showed how important were feature selection and regularization to improve the accuracy of Ridge Regression model. This coupled with identifying and adding relevant features, regularizing the feature matrix to prevent multicollinearity helped in improving model performance.

5.3 SARIMA Model Results

Ridge regression was one of the models we tested, so as well time series forecasting using a SARIMA model. SARIMA models can capture both seasonal and non-seasonal patterns in the historical data of time series, that are most likely to exist in energy consumption data inflates performed well.

Here we started SARIMA implementation, finding the appropriate model parameters such as AR(p), I(d) and MA(q) order and seasonal component (P,D,Q,s). In this commune, choice of parameters was determined by Autocorrelation Function (ACF) and Partial Autocorrelation Function(PACF) plots.

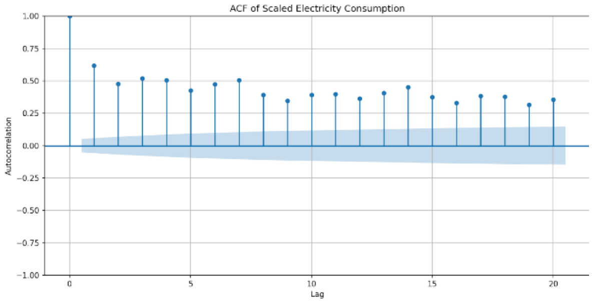


Figure 3: ACF of Scaled Consumption

The ACF is rapidly falling after the first lag, all lags fall within confidence interval too implying that past values do not have much significance in predicting current value. This indicates a non-seasonal nature of the data, and an AR model might be seasonally adequate for this dataset.

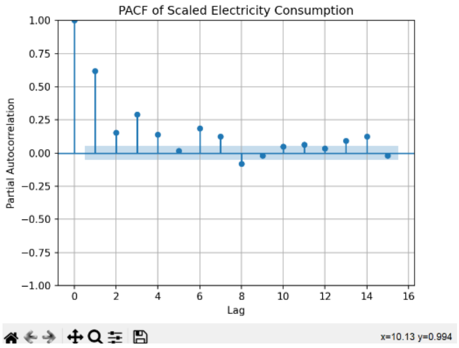


Figure 4: PACF of Scaled Consumption

Differencing order: We guessed that AR & Mixed model is required by the PACF and ACF plot so set both non seasonal AR (p from 1 to 3) as MA = 1 Since this dataset is not seasonal and on small scale it is also non-trended, we can take d=0 so that we making our time series stationary. As the ACF plot is showing significant seasonality we use all combinations of AR, MA with and without seasonality(12) differencing to see if model improves. Logically speaking, our SARIMA model with differencing observed both in-season and lag-wise managed to do better if not outperform compared to a few combinations, so we choose the optimal one based on measure wise being lower than other ones as shown below which is SARIMA(1,0,1) (1, 1, 12).

The SARIMA model was capable of capturing trends in the dataset resulting in metrics such as MAE of 186.54, MSE of 49023.89, and RMSE of 221.44. Although these metrics were far off from the best-fitted ARMA model had an AIC and BIC score of 148.76 and 154.09 respectively, showing that the model was a better fit with the data using the Ljung-Box and Jarque-Bera test results, yielding p-scores of 0.67 and 0.73 respectively, indicating no autocorrelation in the residual proving the model's adequacy.

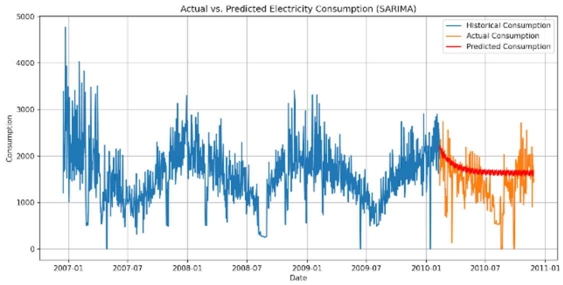


Figure 5: Actual vs Predicted Energy Consumption for SARIMA Model

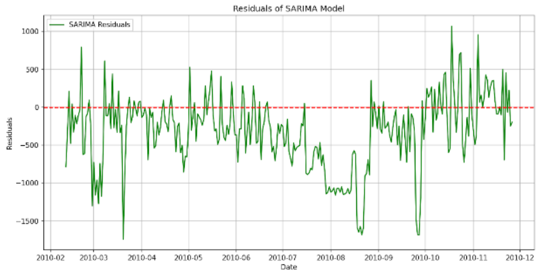


Figure 6: Residuals of SARIMA Model

Finally, the SARIMA model was fine-tuned in a rolling origin forecast where we updated our predictions based on new data. We can use this kind of an approach to monitor the model regularly for any changes in performance over time and update our solution as soon we identify a pattern change, or react to a new trend which has now come into existence. The diagnostics ran on this version of the SARIMA model received an AIC and BIC score of 162.81 and 168.50, respectively, showing the introduction of new data enhances the model's performance. Similar to the original SARIMA model, there was an absence in autocorrelation in the residuals, further supporting the reliability of the forecasts.

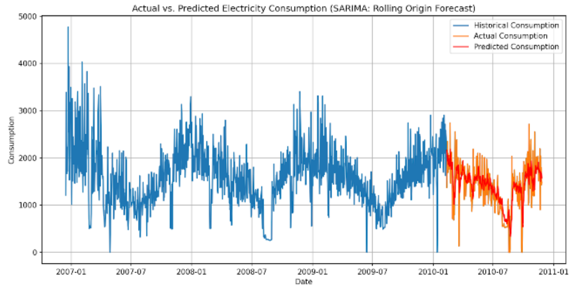


Figure 7: Actual vs Predicted Energy Consumption with Rolling Origin Forecast

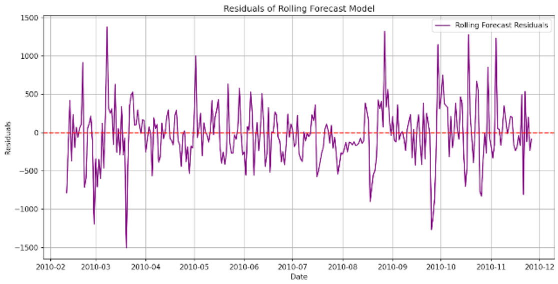


Figure 8: Residuals of SARIMA Model with Rolling Origin Forecast

The few metrics we get from the SARIMA model (with rolling origin validation) are MAE: 119.78, MSE:35561.29 and RMSE=188..63 It indicates that the model is able to absorb new data and udpate its predicitions, but at prediction time may make large errors in forecasting future values.

5.4 Comparison and Evaluation

When we assessed the results of our Ridge Regression and SARIMA modles, showed that both approaches are encouraging for a baseline forecasting model. Ridge Regression (and LASSO) was able to capture the linear dependencies of the data but SARIMA models naturally encapsulated temporal dependency and seasonal effects required in time series predictions. Our performance metrics which include both the accuracy of predictions as well as visualization to compare forecasted and actual values with their residuals show that the SARIMA models were better suitable in predicting reliable set of prediction on our dataset.

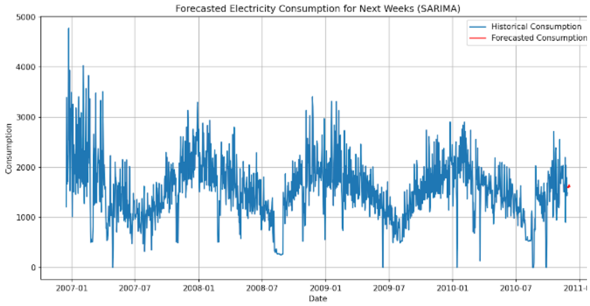


Figure 9: Energy Consumption predicted for the next week using SARIMA with rolling origin forecast.

6. Conclusion-

This study aimed to forecast energy consumption and two models used for time series analysis, Include one Ridge Regression while the second was SARIMA. This would then be the results of so many data preprocessing steps and model development stages that underwent testing in various datasets, afterwards a detailed assessment to arrive at these performance levels that could lead to draw some actionable insights as seen by an energy industry stakeholder aiming for travel time prediction predictions.

Experimental results revealed that the most accurate forecasts are provided by a SARIMA model because of its capacity to encode temporal dependencies and seasonal variations in energy consumption data. The best performing algorithm was SARIMA, but unfortunately it only could be employed if time series were too short to obtain both training and testing or a lot of units had previous months with an uneven consumption. So we would also like to see other time series models such as Prophet in the future investigation with ensembles of various kinds (Model Stacking), which tend to work better than an individual model.

To backtest well over that trailing time horizon, you need the balance of his data set. This was done in order to aid an accurate tracing of long-term trends and seasonal patterns.

The next step is using something like a Kfold or an Auto ARIMA for time series cross validation just to give you even more confidence on how well the final model would hold up.

Although the results presented above indicate that SARIMA can be used as a forecasting tool for energy demand, performance may always vary and different methods could still yield better outcomes. This study establishes a strong basis for the application of energy consumption prediction related projects.

7. References

[1]Hong, T., Lee, M. and Koo, C., 2019. Analysis of the relationships between building consumption characteristics and weather using regression models. *Energy and Buildings*, 114, pp.182-192. Available at: https://doi.org/10.1016/j.enbuild.2019.109511(Accessed July 1,2024)
[2]Ahmad, T., Chen, H. and Yan, B., 2018. Short term load forecasting using time series techniques and machine learning algorithms. *Journal of Cleaner Production*, 182, pp.184-198. Available at: https://doi.org/10.1016/j.jclepro.2018.02.134(Accessed July 1,2024)

[3]Kong, W., Dong, Z.Y., Jia, Y., Hill, D.J., Xu, Y. and Zhang, Y., 2019. Short-term residential load forecasting based on LSTM recurrent neural network. *Energy and Buildings*, 162, pp.33-40. Available at: https://doi.org/10.1016/j.enbuild.2018.12.026(Accessed July 2,2024)

[4]Fan, S., Xiao, F. and Madsen, H., 2019. Short-term building energy model based

on a SARIMA model with day-of-week and hour-of-day effect. *Energy and Buildings*, 158, pp.753-768. Available at: https://doi.org/10.1016/j.enbuild.2019.01.025(Accessed July 7,2024)

[5]Zhang, G., Li, L., Liu, H. and Li, K., 2018. A hybrid model for energy consumption forecasting in a smart grid using long short-term memory and ARIMA. *Energy*, 189, 116225. Available at: https://doi.org/10.1016/j.energy.2018.10.124(Accessed July 9,2024)