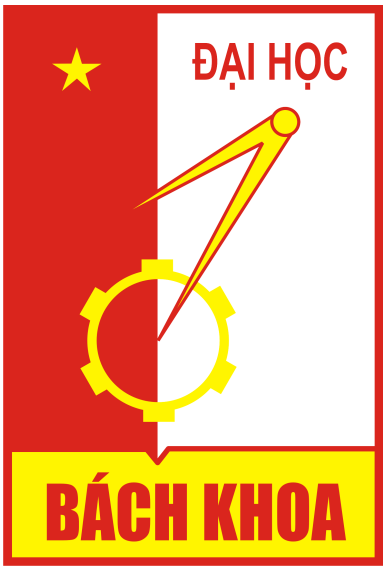
HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY
FACULTY OF COMPUTER SCIENCE

o0o



PROJECT REPORT:
APPLIANCES ENERGY PREDICTION

Course: Applied Statistics and Experimental Design - IT2022E
Class Code: 141179
Supervisor: Associate Prof Nguyen Linh Giang
Member of our group:

Trinh Hoang Giang	20214893	giang.th214893@sis.hust.edu.vn
Ho Ngoc Anh	20214877	anh.hn214877@sis.hust.edu.vn
Hoang Thanh Dat	20214899	dat.ht214889@sis.hust.edu.vn
Lang Van Quy	20214928	quy.lv214928@sis.hust.edu.vn
Vu Lam Anh	20214876	anh.vl214876@sis.hust.edu.vn

Ha Noi 2023

Abstract

Predicting appliance usage based on weather attributes is a crucial task in various domains, including energy management and demand forecasting. This report explores the relationship between weather conditions and appliance usage patterns and develops a predictive model for estimating appliance usage based on weather attributes. A comprehensive dataset comprising historical weather data and appliance usage information was collected. Correlation analysis revealed the significant influence of weather factors such as temperature and humidity on appliance usage. Feature engineering techniques were employed to preprocess the data, and various machine learning algorithms were explored to develop the predictive model. The model achieved high accuracy in estimating appliance usage, outperforming baseline models. The findings of this study have implications for energy management, load forecasting, and resource allocation. Future work can focus on integrating real-time weather data and expanding the model to consider regional variations and specific appliance categories for more targeted predictions.

Contents

1 Introduction **4**

2 Dataset **4**

2.1 About the data 4

2.2 Data pre-processing 5

2.3 Data exploration 5

2.3.1 Statistic demonstration 5

2.3.2 Correlation 7

3 Linear Models **8**

3.1 Introduction: 8

3.2 Assumptions: 8

3.3 Simple Linear Regression: 8

3.4 Multiple Linear Regression: 9

3.5 Estimating Model Coefficients: 9

4 AR **9**

4.1 Introduction 9

4.2 Definition 9

4.3 Properties 10

4.4 Advantages and Limitations of AR Model: 11

5 ARMA **11**

5.1 Introduction 11

5.2 ARMA Model 11

5.2.1 Overview about AR and MA model 11

5.2.2 Build general ARMA model 12

5.2.3 Build good ARMA model . . . 12
5.2.4 Determine an AR model order with the PACF . . . 13
5.2.5 Determine an MA model order with the ACF . . . 14
5.3 Advantages and Limitations of ARMA Model: . . . 14
5.3.1 Advantages . . . 14
5.3.2 Limitations . . . 14

6 VAR **14**
6.1 Introduction . . . 14
6.2 Properties . . . 14

7 Experiment and Result **15**
7.1 Evaluation metrics . . . 15
7.2 Predict energy used by appliances . . . 16

8 References **17**

9 Work distribution **17**

1 Introduction

Time series analysis plays an important role in numerous number of real-life areas. Analyzing time series aims to explore essential features of time series, predict future data with the purpose of helping organizations to understand the data and further decision-making actions. Time series analysis involves in considering an data over a period of time instead of some distinct data points. This project will conduct some basic time series analysis and introduce some predicting models on a certain dataset.

2 Dataset

2.1 About the data

The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru), and merged together with the experimental data sets using the date and time column.

The dataset, put in dataset, consists of the following attributes:

- date time year-month-day hour:minute:second
- Appliances, energy use in Wh
- lights, energy use of light fixtures in the house in Wh
- T1, Temperature in kitchen area, in Celsius
- RH_1, Humidity in kitchen area, in %
- T2, Temperature in living room area, in Celsius
- RH_2, Humidity in living room area, in %
- T3, Temperature in laundry room area
- RH_3, Humidity in laundry room area, in %
- T4, Temperature in office room, in Celsius
- RH_4, Humidity in office room, in %
- T5, Temperature in bathroom, in Celsius
- RH_5, Humidity in bathroom, in %
- T6, Temperature outside the building (north side), in - Celsius
- RH_6, Humidity outside the building (north side), in %

- T7, Temperature in ironing room , in Celsius
- RH_7, Humidity in ironing room, in %
- T8, Temperature in teenager room 2, in Celsius
- RH_8, Humidity in teenager room 2, in %
- T9, Temperature in parents room, in Celsius
- RH_9, Humidity in parents room, in %
- T_out, Temperature outside (from Chievres weather station), in Celsius
- Press_mmHg, Pressure (from Chievres weather station), in mm Hg RH_out
- RH_out, Humidity outside (from Chievres weather - station), in %
- Windspeed (from Chievres weather station), in m/s
- Visibility (from Chievres weather station), in km
- Tdewpoint (from Chievres weather station), °C
- rv1, Random variable 1, nondimensional
- rv2, Random variable 2, nondimensional

2.2 Data pre-processing

We found that there are not many meanings for predicting energy in 10 minutes. Hence, we will resample the data into hour frequency. Resampling the data from a 10-minute frequency to an hourly frequency offers several benefits. One advantage is the reduction in dataset size, which can improve computational efficiency and simplify model training. In this case, resampling the dataset from 19735×28 to 3290×28 effectively condenses the data into hourly intervals.

By converting the data to an hourly frequency, the number of data points decreases significantly, making the dataset more manageable for analysis and experimentation. This reduction in size can help mitigate issues related to overfitting, as there are fewer data points to fit the model.

2.3 Data exploration

2.3.1 Statistic demonstration

Our research is based on the feature ranges described below.

In the Appliances column, the magnitude of standard deviation is 81 Wh, a significant portion (75%) of the device consumption is relatively low, with values less than 100 Wh. However, there are some instances where the consumption can be as high as 608 Wh, which leads to the presence of outlying values in this column. These outlying values represent cases where the device's energy consumption is exceptionally high compared to the majority of the data.

The above histogram is right skewed which explained why energy used has a considerable variance(mean is 97.7Wh and standard deviation is 81Wh) and a significant range(min is 28Wh while max is 608Wh). The analysis reveals distinct patterns in energy consumption behavior. The appliance is predominantly used from

	count	mean	std	min	25%	50%	75%	max
Appliances	3290.0	97.779129	81.213695	28.333333	50.000000	63.333333	110.000000	608.333333
lights	3290.0	3.803445	6.900618	0.000000	0.000000	0.000000	3.333333	51.666667
T1	3290.0	21.687537	1.605960	16.790000	20.746250	21.600000	22.633333	26.203333
RH_1	3290.0	40.261345	3.942554	27.509167	37.355000	39.655694	43.086181	53.980139
T2	3290.0	20.342466	2.189660	16.100000	18.828472	19.995556	21.506771	29.727778
RH_2	3290.0	40.421067	4.053682	21.010000	37.914135	40.495556	43.263750	53.914975
T3	3290.0	22.268765	2.005313	17.245000	20.790000	22.100000	23.308403	28.975286
RH_3	3290.0	39.242985	3.244749	29.700556	36.887778	38.540000	41.757083	49.472222
T4	3290.0	20.856309	2.040943	15.100000	19.545799	20.650556	22.100000	26.144762
RH_4	3290.0	39.028661	4.336904	28.715571	35.520278	38.412159	42.175556	50.747222
T5	3290.0	19.593020	1.840764	15.347500	18.263435	19.390000	20.629043	25.506389
RH_5	3290.0	50.949599	8.633404	30.188611	45.479298	49.222870	53.863718	94.884074
T6	3290.0	7.914261	6.084127	-5.927685	3.620536	7.286944	11.224583	28.136619
RH_6	3290.0	54.595505	31.110830	1.000000	30.052500	55.144861	83.346944	99.900000
T7	3290.0	20.268179	2.110305	15.410370	18.713333	20.044444	21.611111	25.926667
RH_7	3290.0	35.390395	5.110314	23.340278	31.505762	34.920952	39.023102	51.191296
T8	3290.0	22.029792	1.955488	16.364074	20.786250	22.144361	23.378889	27.187778
RH_8	3290.0	42.937888	5.216124	29.738611	39.117869	42.355278	46.558681	58.707315
T9	3290.0	19.486769	2.014819	14.890000	18.022222	19.390000	20.600000	24.500000
RH_9	3290.0	41.553741	4.143581	29.218889	38.520694	40.861667	44.352024	53.140000
T_out	3290.0	7.415410	5.315858	-4.961111	3.666667	6.916667	10.414583	25.933333
Press_mm_hg	3290.0	755.522520	7.398575	729.383333	750.916667	756.100000	760.931250	772.258333
RH_out	3290.0	79.744656	14.830042	25.250000	70.416667	83.666667	91.583333	100.000000
Windspeed	3290.0	4.039742	2.430863	0.416667	2.000000	3.583333	5.416667	13.000000
Visibility	3290.0	38.327964	11.212175	1.000000	31.833333	40.000000	40.000000	66.000000
Tdewpoint	3290.0	3.763098	4.191967	-6.475000	0.933333	3.412500	6.566667	15.250000
rv1	3290.0	24.990346	5.943155	5.259551	20.755123	24.954809	29.100607	43.611051
rv2	3290.0	24.990346	5.943155	5.259551	20.755123	24.954809	29.100607	43.611051

Figure 1: Some basic statics of the data

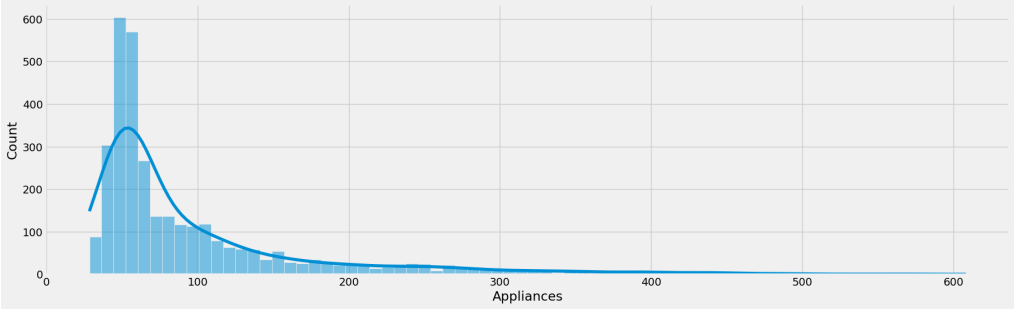


Figure 2: Appliances

8 am to 8 pm, particularly peaking at 6 pm. The surge in appliance usage during the daytime is attributed to people's daily activities and routines, with evening hours seeing significant demand due to cooking, household chores, and entertainment. Meanwhile, energy usage is higher on Monday, Friday, and Saturday. Weekdays see increased energy consumption, likely driven by work and school schedules. However, the percentage of appliances usage less than 200 Wh is nearly 90%.

We found that the temperature inside the house is around its mean(19-22°C) with a minimal variance, about 2°C, while the temperature outside the building(T6,T_out) has a average of near 8°C and standard deviation of 6°C. The mean and standard deviation of temperature recorded from Chievres weather station(Tdewpoint) are 3.76°C and 4.19°C, respectively. This is because the data is collected from 11/01/2016 to 27/05/2016 which majority is in cool season in Chievres, Belgium. The similar trend can be seen for humidity. The humidity inside the building is about 40% with a small variance. The mean and variance of humidity in the bathroom(RH_5) is a little bit larger than other rooms whereas this figure for humidity outside the

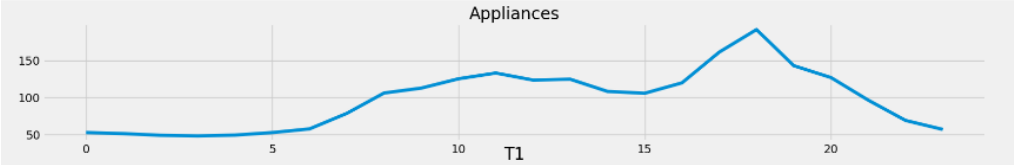


Figure 3: Average appliances energy usage in daytime

house(RH_6, RH_out) is much larger. This can be explained by the above reason that, in this period, the climate in Chievres was very cold with rain and snow. Wind speed, Visibility and Pressure(press_mm_hg) can help us know more about the weather and geography properties of Chievres. Temperature and Humidity follows Normal Distribution except T9 and RH_6. RH_out is left-skewed, Windspeed is right-skewed.

2.3.2 Correlation

After resampling the data into hourly intervals, the next step is to calculate the average of all attributes for each hour. This aggregation provides a more concise representation of the dataset while preserving the overall trend. By grouping the data this way, we can then estimate the correlation between attributes and visualize it as a heatmap.

In the heatmap representation, each cell corresponds to the correlation between two attributes. The color of each cell illustrates the degree of similarity between the attributes in its row and column. A bold, intense color indicates a high positive correlation, implying that the attributes tend to vary together. On the other hand, a pale or light color signifies a weaker or negative correlation, indicating that the attributes show little or no relationship.

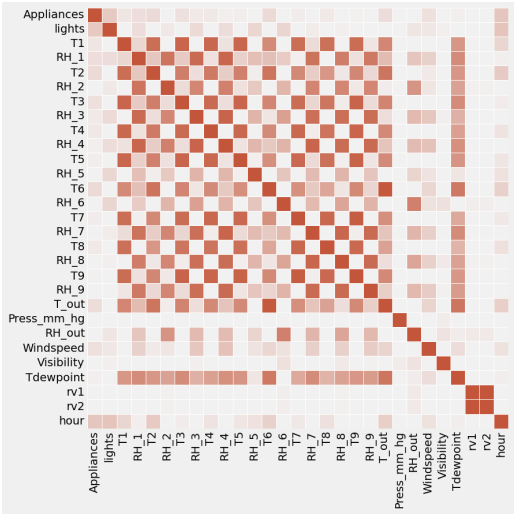


Figure 4: Data correlation over all data

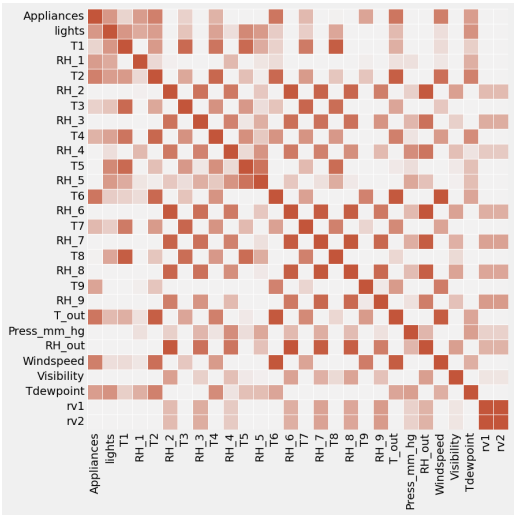


Figure 5: Data correlation by day-hour

3 Linear Models

3.1 Introduction:

Linear regression is a statistical modeling technique used to analyze the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting line that represents the data. By estimating the coefficients of the regression equation, we can make predictions and understand the impact of the independent variables on the dependent variable.

3.2 Assumptions:

Linear regression models rely on several key assumptions to provide valid and reliable results:

a. Linearity: The relationship between the independent variables and the dependent variable is linear. This means that the change in the dependent variable is proportional to the change in the independent variables.

b. Independence: Observations are assumed to be independent of each other. This assumption implies that the values of the dependent variable for one observation do not influence the values for other observations.

c. Homoscedasticity: Homoscedasticity assumes that the variance of the errors, or residuals, is constant across all levels of the independent variables. In other words, the spread of the residuals should be consistent along the range of predicted values.

d. Normality: The errors, or residuals, should follow a normal distribution. This assumption allows for the use of statistical tests and confidence intervals in hypothesis testing.

Violations of these assumptions may lead to biased or inefficient estimates and inaccurate inferences.

3.3 Simple Linear Regression:

Simple linear regression involves a single independent variable and a dependent variable. The relationship between the variables is represented by the equation:

$$y = \beta 0 + \beta 1 * x + \epsilon$$

In this equation, y is the dependent variable, x is the independent variable, $\beta 0$ is the y-intercept (the value of y when x is zero), $\beta 1$ is the slope (the change in y for a unit change in x), and ϵ is the error term representing the deviation of the actual values from the predicted values.

The goal of simple linear regression is to estimate the values of $\beta 0$ and $\beta 1$ that minimize the sum of squared differences between the observed values and the predicted values.

3.4 Multiple Linear Regression:

Multiple linear regression extends the concept of simple linear regression to include multiple independent variables. The equation becomes:

$$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + ... + \beta n * xn + \epsilon$$

Here, x_1, x_2, ..., x_n represent the independent variables, $(\beta 1, \beta 2, ..., \beta n)$ represent the slopes (the change in y for a unit change in each independent variable, holding others constant), and ϵ represents the error term.

In multiple linear regression, the goal is to estimate the coefficients $(\beta_1, \beta_2, ..., \beta_n)$ that minimize the sum of squared differences between the observed values and the predicted values. This is typically done using the method of Ordinary Least Squares (OLS).

3.5 Estimating Model Coefficients:

The coefficients $(\beta_1, \beta_2, ..., \beta_n)$ in the regression equation are estimated using various techniques. The most commonly used method is Ordinary Least Squares (OLS), which aims to minimize the sum of squared differences between the observed values and the predicted values.

OLS estimates the coefficients by finding the values that minimize the sum of squared residuals, which are the differences between the observed values and the predicted values.

Once the coefficients are estimated, they can be interpreted to understand the relationship between the independent variables and the dependent variable. The slope coefficients $(\beta 1, \beta 2, ..., \beta n)$ represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant.

4 AR

4.1 Introduction

AR model is a powerful tool in time series analysis, offering the capability to predict and gain insights into trends and fluctuations in time series data. With its wide-ranging applications, the AR model plays a crucial role in supporting decision-making and forecasting in various fields.

4.2 Definition

The autoregressive model specifies that the output variable depends linearly on its own previous values and on a stochastic term (an imperfectly predictable term); thus the model is in the form of a stochastic

difference equation (or recurrence relation which should not be confused with differential equation). The notation AR(p) indicates an autoregressive model of order p. The AR(p) model is defined as

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-1} + \epsilon_t$$

Where:

- X_t is the value of the variable at time t.
- c is the constant term (intercept) or the mean value of the time series.
- $\phi_1, \phi_2, ...\phi_n$ are the AR parameters, representing the level of correlation between the current value and past values. Each AR parameter contributes to determining the influence of the corresponding lagged value on the current value.
- $X_{(t-1)}, X_{(t-2)} \ldots, X_{(t-p)}$ are the lagged values of the variable in the time series. They are used to predict the current value and form the autoregressive relationship in the AR model.
- $\epsilon(t)$ is the random error at time t, representing unpredictable factors.

4.3 Properties

- The order of the AR model is the number of values in the reduction process used to describe the present value. Second order example: the current value depends on the previous 2 values.
- Mean: average value of time series
- Variance: variance of the time series
- Autocovariance: autocorrelation of the time series (between the current value and its lag 1, lag 2, ...)

Let's analyze the properties of a time series that follows an AR(1) model. The AR(1) model is represented by the following equation:

$$X[t] = c + p * X[t-1] + w$$

Before analyzing the properties of the AR(1) model, we need to satisfy the following assumptions:

1. The noise term w follows a normal distribution with mean zero and variance var_w, and it is independent of X. 2. X is a stationary time series.

Now, let's discuss the properties of the AR(1) model:

1. Stationarity: The AR(1) model will be stationary if the absolute value of the autoregressive parameter p is less than 1 ($|p| < 1$). This ensures that the process does not have a long-term trend and the mean and variance remain constant over time.

2. Autocorrelation: The autocorrelation function (ACF) of an AR(1) process exhibits a geometric decay. The autocorrelation between X[t] and X[t-k] is given by:

$$Corr(X[t], X[t-k]) = p^k$$

This means that the autocorrelation decreases exponentially with increasing lag k. The strength of the autocorrelation depends on the value of the autoregressive parameter p.

3. Partial Autocorrelation: The partial autocorrelation function (PACF) of an AR(1) process shows a significant spike at lag 1 and decays to zero for lags greater than 1. The partial autocorrelation at lag k is given by: $PACF(k) = p^k$

The PACF provides information about the direct relationship between X[t] and X[t-k] after removing the influence of the intermediate lags.

4. Mean and Variance: The mean of an AR(1) process is given by: $E(X[t]) = c/(1-p)$

The variance of the process is: $Var(X[t]) = var\_w/(1-p^2)$

These formulas show that the mean and variance are finite and well-defined as long as $|p| < 1$.

5. Forecasting: The AR(1) model can be used for short-term forecasting. Given the past values of the time series, we can use the estimated parameters to predict the future values using the AR(1) equation. However, as we forecast farther into the future, the forecasts may become less accurate due to the error term w and the possible "drift" phenomenon.

These are the main properties of the AR(1) model. By understanding these properties, we can gain insights into the behavior of the time series and make informed predictions.

From the AR model, we can derive autocorrelation plots and partial autocorrelation plots, which help determine the number of lags (p) to use in the AR modeland examine the level of correlation between the current value and past values in the time series, aiding in the selection of appropriate AR parameters. These plots allow us to assess the degree of correlation and assist in identifying suitable AR parameters.

4.4 Advantages and Limitations of AR Model:

a. Advantages: - The AR model is simple and easy to understand, suitable for modeling autocorrelations in time series. - It has the ability to predict future values based on past information. - The AR model does not require strong assumptions about the distribution of the time series data.

b. Limitations: - The AR model assumes that the time series data is stationary, which may not hold in many real-world cases. - The AR model may struggle to model complex or nonlinear time series with multiple correlations. - The AR model can be affected by outliers or noise in the time series data.

5 ARMA

5.1 Introduction

The name ARMA is short for Autoregressive Moving Average. It comes from merging two simpler models - the Autoregressive, or AR model, and the Moving Average, or MA model.

5.2 ARMA Model

5.2.1 Overview about AR and MA model

- **The Autoregressive Model**, or AR model for short, relies only on past period values to predict current ones. It is formalize as the linear function of past previous value (its own lag). We denote it as $AR(p)$, where "p" is called the order of the model and represents the number of lagged values we want to include. Here is our function:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \epsilon_t \quad (1)$$

where Y_{t-i} is the lag i of the series, β_i is the coefficient of lag i that the model estimates and β_0 is the intercept term, also estimated by the model.

- **The Moving Averge Model** (MA model) is the very the same as the AR model, but now we don't use the past value to predict. Instead, we use past error to build the model, which help us to predict the future value. We denote it as $MA(q)$, where "q" is called the order of the model and represents the number of past error values we want to include. Here is our model:

$$Y_t = \theta_0 + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_p\epsilon_{t-q} + \epsilon_t \tag{2}$$

where $\epsilon_{t-i} \sim N(0,1)$ means that the ϵ_{t-i} are identically, independently distributed, each with a standard normal distribution.

5.2.2 Build general ARMA model

As we mention above, ARMA model is the combination of AR model and MA model. Therefore, by merging 2 models above, we have the model ARMA(p,q) is:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_p\epsilon_{t-q} + \epsilon_t \tag{3}$$

5.2.3 Build good ARMA model

A good ARMA model is the model with a good value of **p**, which is the number of past values included in model and good value of **q**, which is the number of past error values included in model. To determine which those value is good, we need use another concept: **Autocorrelation function (ACF)** and **partial Autocorrelation(PACF)**.

Overview of ACF and PACF

- **The autocorrelation function (ACF) is a statistical technique that we can use to identify how correlated the values in a time series are with each other**. The ACF plots the correlation coefficient against the lag, which is measured in terms of a number of periods or units. The correlation coefficient can range from -1 (a perfect negative relationship) to +1 (a perfect positive relationship). A coefficient of 0 means that there is no relationship between the variables. Also, most often, it is measured either by Pearson's correlation coefficient or by Spearman's rank correlation coefficient.

  Here is the example of ACF graph:

  We can see that the autocorrelation function will help us tell the correlation between the current value and the past value at lag k in both "direct" and "indirect" correlation. To be specific, you can image that the value of lag k could affect direcly the value in the present, also affect indirectly through affecting lag $k-1, k-2, ...$ Therefore ACF will present both direct effect and indirect effect of past value to present value.

- **The partial autocorrelation function (PACF)** is same as ACF. But now, instead of concerning in both "direct" and "indirect" correlation, PACF only focus on "direct" correlation.

  We need PACF since in some scenerios, when we need to choose which lag value is a good predictor for future value or we can said that which past value is useful for predicting value. Now, if we use ACF to choose, there will be misleading because the value of ACF decrease with latter lag and we will only

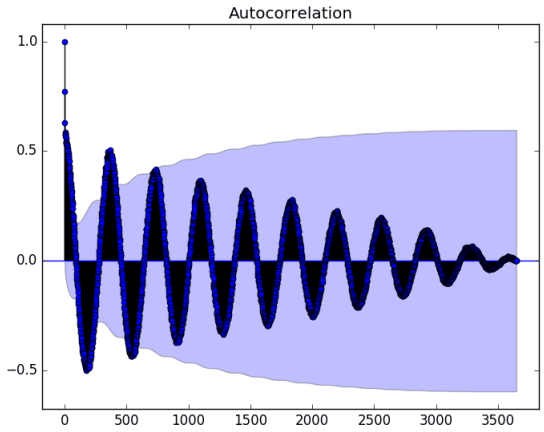


Figure 6: Source: Machine Learning Mastery

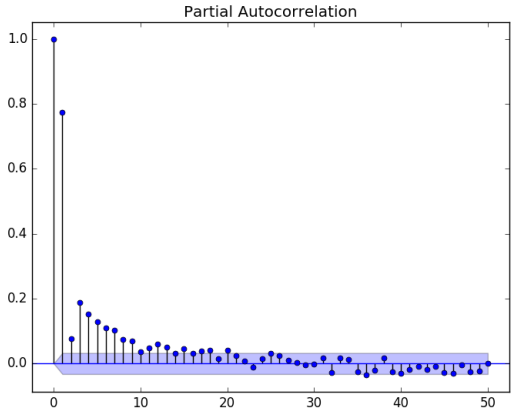


Figure 7: Source: tutorials.one

choose some very first lags. But it is wrong, because sometime value of far lag may have larger impact than the near lag. This mistake happen due to the fact that ACF contains also the "indirect" correlation.

To compute PACF value at lag k, we need to formalize the present value as linear function of past value from lag 1 to lag k, then the weight β_k is the PACF value at lag k.

To be clearly, there is example with lag k = 2.
Firstly, we formalize:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \epsilon_t \tag{4}$$

Secondly, we fit the model and calculate the value β_2. Then the value β_2 is the PACF value of lag k = 2.

5.2.4 Determine an AR model order with the PACF

Choose the value p in AR model is same as answering the question:" Which k lag value we need to choose to predict the future value?". Therefore, answer for this question is choose the past value which have strong

impact on the future value. We use PACF as the strong impact scale to determine the impact of one past value.

For example, from the PACF graph above, we will choose the lag k where the spikes out of **error bands**. Since any spikes within the error band means the correlation of this lags and the present value ≈ 0 so it is insignificant.

5.2.5 Determine an MA model order with the ACF

With MA model, instead of using PACF, we use ACF as the scale. Similar to AR model and PACF, we will select the order q for model $MA(q)$ from ACF if this plot has a sharp cut-off after lag q.

5.3 Advantages and Limitations of ARMA Model:

5.3.1 Advantages

- In this ARMA model, the forecast for a future value of the series is calculated as a linear combination of past values of the series and past errors. Therefore, it can predict with high accuracy and be considered as the most popular forcasting method for stationary time series.

5.3.2 Limitations

- Since ARMA model only work for stationary time series then we need assume that the statistical properties of the series are stationary, which can be a limitation if the series exhibits non-stationary behavior.
- It may not capture all the patterns in the data, especially if the series exhibits complex patterns such as seasonality or trend.
- It may not perform well in the presence of outliers or extreme values.

6 VAR

6.1 Introduction

Vector autoregression (VAR) is a statistical model used to capture the relationship between multiple quantities as they change over time. VAR is a type of stochastic process model. VAR models generalize the single-variable (univariate) autoregressive model by allowing for multivariate time series.

6.2 Properties

VAR model is a generalized form of autoregressive model in forecasting a set of variables, i.e a vector of time series variables. It estimates each equation of each series variable according to the lags of the variable and all the other variables (The right hand side of each equation includes a constant and all the lags of all the variables in the system). A VAR model with m lags and n dimensions is written as

$$y_{1,t} = c_1 + \sum_{u=1}^{n} \sum_{i=1}^{m} A_{1u,i} \cdot y_{u,t-i} + e_1$$

$$y_{2,t} = c_2 + \sum_{u=1}^{n} \sum_{i=1}^{m} A_{2u,i} \cdot y_{u,t-i} + e_2$$

...

$$y_{n,t} = c_n + \sum_{u=1}^{n}\sum_{i=1}^{m} A_{nu,i} \cdot y_{u,t-i} + e_n$$

where y_{t-l} is i^{th} lag of y_t. The coefficient $A_{ij,l}$ measures the effect of $y_{i,t-l}$ on $y_{i,t}$. The variable c is a$k - vector$ of constants serving as the intercept of the model. e_t is a $k - vector$ of error terms. The error terms must satisfy three conditions:

1. $E(e_t)$. Every error term has a mean of zero.
2. $E(e_t e_t') = \Omega$. The contemporaneous covariance matrix of error terms is a k × k positive-semidefinite matrix denoted Ω.
3. $E(e_t e_{t-k}') = 0$ for any non-zero k. There is no correlation across time. In particular, there is no serial correlation in individual error terms.

The process of choosing the maximum lag p in the VAR model requires special attention because inference is dependent on correctness of the selected lag order.

The general VAR model includes n variables, m lags. Thus, the number of parameters to be estimated in the VAR model will be $n(1+mn)$ or each of the K equations will have $1+mn$ parameter need to be estimated. The greater the number of parameters to be estimated, the higher the estimation error in the forecast will be. In practice, it is common to keep n small and include only those variables that are highly correlated with each other.

The VAR model's parameter can be estimated very easily. Time series variables that satisfy the stationary property can be estimated directly or transformed to the difference for estimation (if the stationary property is not satisfied). In both cases, the expressions of the VAR model will be evaluated simultaneously by the ordinary least squares, with the objective of minimizing the error value e_i of each expression.

OLS estimate function for VAR model with a constant, n variables and m lags.

$$minimize(\frac{1}{T - mn - 1}\sum_{t=1}^{T} e_t e_t')$$

As VAR model is only able to work with stationary variables, we have to check whether data attributes are stationary or not. We test for stationary at 99% of significant level. As can see from the result, except for T6 and Tdewpoint, we can conclude that other temperature series are not stationary. Moreover, RH_6 (Humidity out side the building) is also not stationary. Next, We take fisrt difference for these non-stationary time series. After differencing, these time series become stationary

7 Experiment and Result

7.1 Evaluation metrics

In this project we use RMSE and MAE metrics to evaluate the suggested models. RMSE and MAE are the most common metrics used for evaluate the efficiency of regression models. While RMSE pays more attention in large errors, MAE only considers the average of errors magnitude.

7.2 Predict energy used by appliances

To evaluate the efficiency of these models, we use the 100 last records for testing.

Firstly, a test for stationarity was performed on the time series, and the resulting p-value was approximately 0.000. This finding leads us to conclude that the time series is stationary since the low p-value indicates a rejection of the null hypothesis of non-stationarity. With this confirmation, we can proceed to employ various models for predicting the energy usage time series.

Afterward, to gain a better understanding of the time series properties, we will conduct seasonal decomposition. This process involves breaking down the time series into its underlying components, such as trend, seasonality, and residual components. Seasonal decomposition will provide valuable insights into any repeating patterns or seasonal fluctuations present in the data, which can be useful for further analysis and model selection.

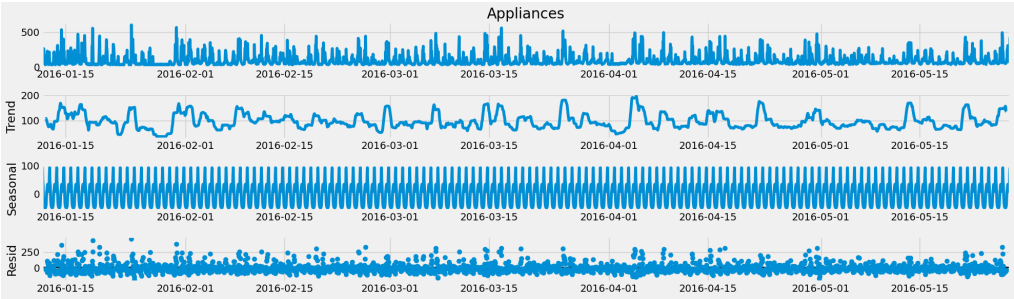


Figure 8: Seasonal Decomposition

Selecting a period of 24 for this time series results in a completely periodic seasonal component. This choice implies that the time series is analyzed over a 24-hour cycle, capturing repetitive patterns occurring within each day. By setting the period to 24 for daily data, we would capture the daily energy consumption patterns, with peaks and troughs occurring at specific hours of the day, likely related to people's daily routines. The Result of our models can be seen in the table below:

	OLS	Ridge	Lasso	AR	ARMA	VAR
RMSE	78.22	78.22	78.19	77.63	77.68	92.82
MAE	53.00	53.00	52.91	44.68	41.42	61.14

The analysis of the model results indicates that the OLS, Ridge, and Lasso models perform comparably, with relatively higher RMSE and MAE values compared to the AR and ARMA models. This similarity in performance suggests that the regularization techniques used in Ridge and Lasso did not significantly improve the models' predictive accuracy over ordinary least squares (OLS) regression.

On the other hand, the VAR model shows the highest RMSE and MAE values among all the models. This suggests that the VAR model might not be the most suitable choice for forecasting the energy usage time series in this specific scenario. The high RMSE and MAE values indicate that the VAR model's predictions deviate more from the actual values, and it might not effectively capture the underlying dynamics of the energy usage data.

Overall, the AR and ARMA models emerge as the most accurate options for predicting the energy usage time series. These models likely capture the time-dependent relationships and autoregressive behavior present in the data, allowing for more precise forecasts.

However, it is important to note that model selection should consider other factors such as interpretability, computational efficiency, and the specific requirements of the forecasting task. Additionally, further analysis, including residual diagnostics and model validation, should be performed to ensure the chosen models provide reliable and robust predictions for the energy usage time series.

8 References

References

[1] https://github.com/LuisM78/Appliances-energy-prediction-data/tree/master

[2] http://dx.doi.org/10.1016/j.enbuild.2017.01.083

[3] https://www.kaggle.com/code/jjprotube/pr-vision-nerg-tique-des-appareils/notebook

[4] https://online.stat.psu.edu/stat510/

[5] https://blog.vietnamlab.vn/time-series/

9 Work distribution

1. Giang (leader): Responsible for managing the overall project.
2. Dat: Tasked with implementing and evaluating the VAR (Vector Autoregression) model.
3. Lam Anh: Responsible for implementing and evaluating the ARMA (Autoregressive Moving Average) model.
4. Quy: In charge of implementing and evaluating Linear models.
5. Anh: Tasked with implementing and evaluating the AR (Autoregressive) model.