TIME SERIES: APPLIANCES ENERGY PREDICTION

Group 8: Applied Statistics and Experimental Design Project

Team members:

20214876 - Vũ Lâm Anh

20214877 - Hồ Ngọc Ánh

20214893 - Trịnh Hoàng Giang

20214899 - Hoàng Thành Đạt

20214928 - Lăng Văn Quý

Content outline

1. Introduction
2. Dataset
3. Linear Models
4. Autoregressive model
5. ARMA model
6. VAR model
7. Experiment and Result

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.

Content outline

1. Introduction
2. Dataset
3. Linear Models
4. Autoregressive model
5. ARMA model
6. VAR model
7. Experiment and Result

DATASET

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.

DATA PRE-PROCESSING

Not many meanings for predicting energy in 10 minutes

Resampling the data from 10-minute frequency to hourly frequency

Resampling the dataset from 19735 × 28 to 3290 × 28

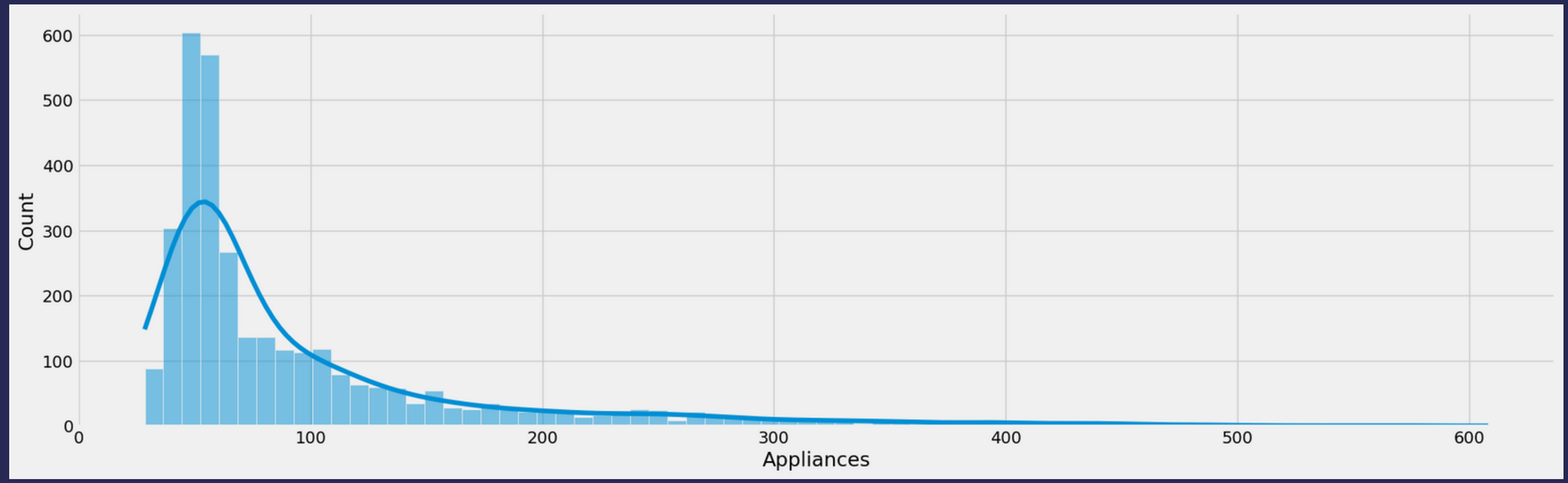
DATA EXPLORATION

In Appliances column:

- The standard deviation is 81 Wh
- 75% of the device consumption is less than 100 Wh.

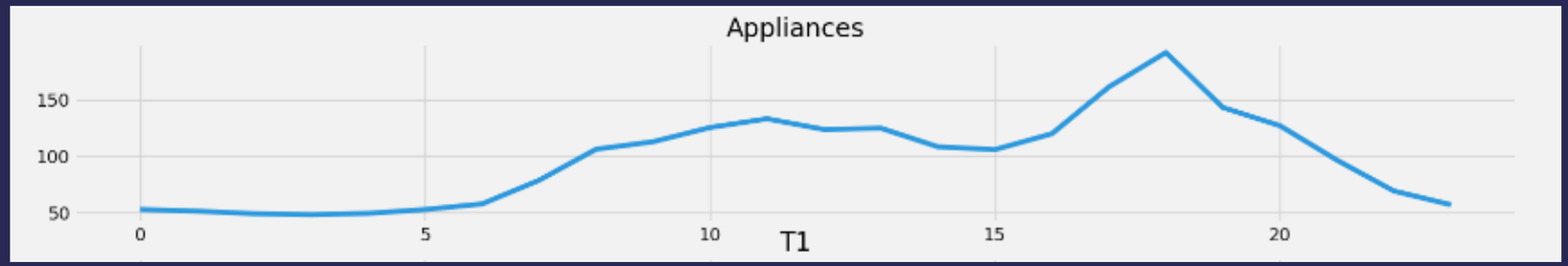
	count	mean	std	min	25%	50%	75%	max
Appliances	3290.0	97.779129	81.213695	28.333333	50.000000	63.333333	110.000000	608.333333
lights	3290.0	3.803445	6.900618	0.000000	0.000000	0.000000	3.333333	51.666667
T1	3290.0	21.687537	1.605960	16.790000	20.746250	21.600000	22.633333	26.203333
RH_1	3290.0	40.261345	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

DATA EXPLORATION



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)

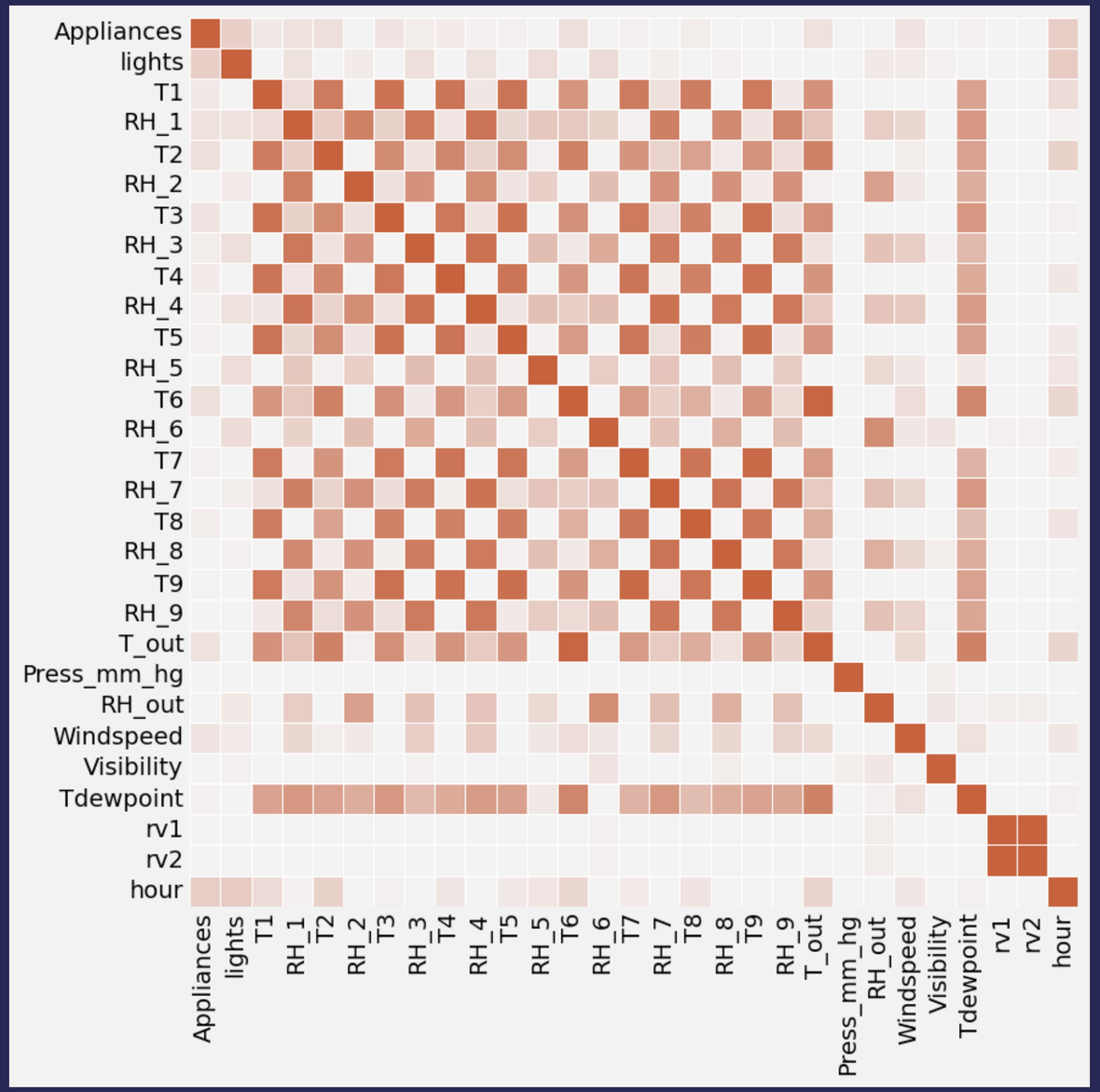
DATA EXPLORATION



- The appliance is mainly used from 8 am to 8 pm, particularly peaking at 6 pm

CORRELATION

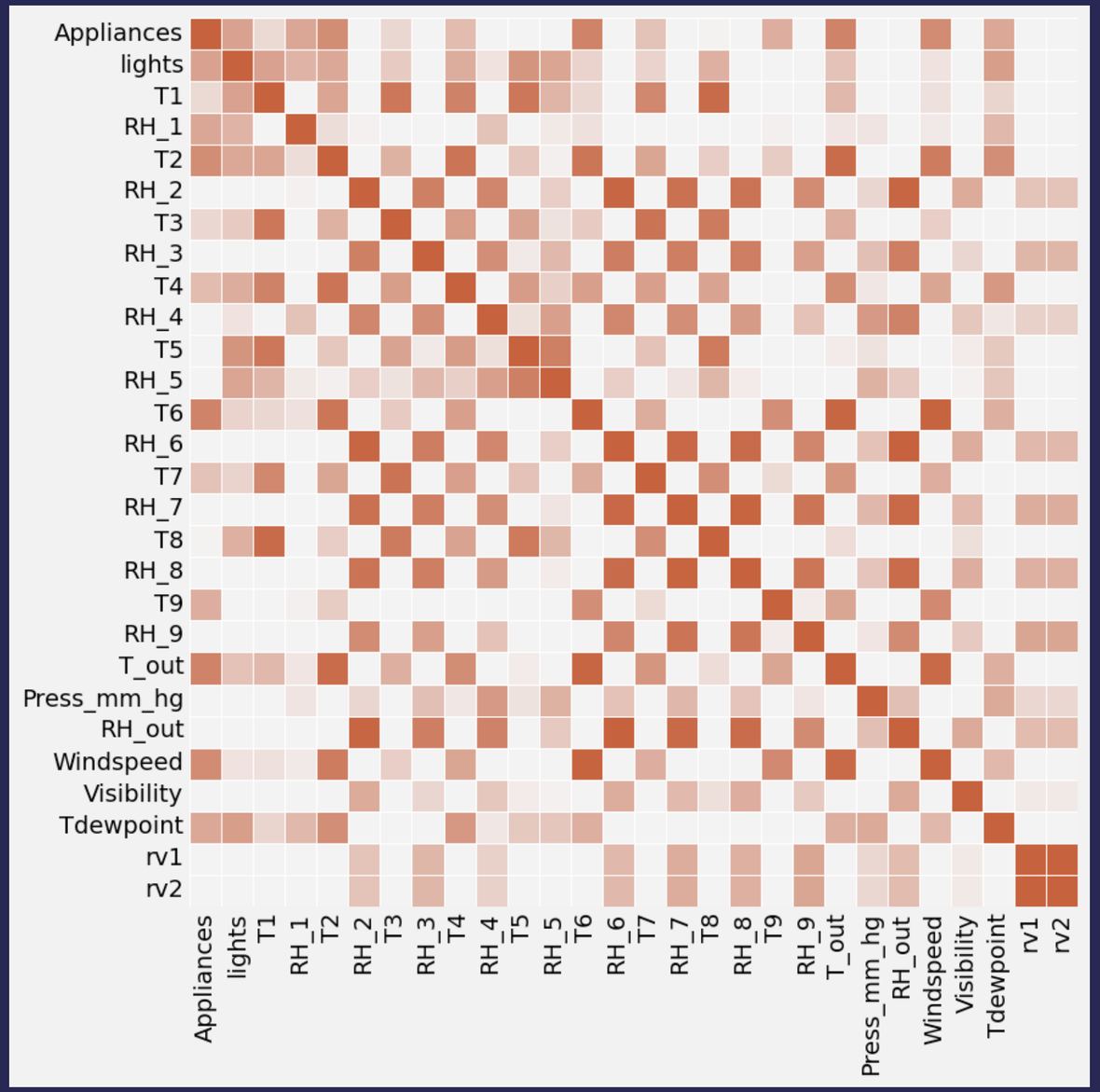
- The similarity between appliance and other time series is considerably low
- Appliances is relatively independent of the other variables



Data correlation over all data

CORRELATION

- The correlation between Appliances and lights and temperature attributes are higher



Data correlation by day-hour

Content outline

1. Introduction
2. Dataset
3. Linear Models
4. Autoregressive model
5. ARMA model
6. VAR model
7. Experiment and Result

LINEAR MODEL

1. INTRODUCTION
2. IMPLEMENTATION
3. ADVANTAGES & LIMITATIONS

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

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:

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

- y is the dependent variable
- x is the independent variable
- β0 is the y-intercept (the value of y when x is zero)
- β1 is the slope (the change in y for a unit change in x)
- ϵ is the error term

2. IMPLEMENTATION

Data Preparation

We split the data into 2 parts, the last 100 records for testing and the rest for training

Estimating Model Coefficients

The coefficients (β1, β2, …, βn) in the regression equation are estimated using various techniques. In this project, we used Ordinary Least Squares (OLS) method, which aims to minimize the sum of squared differences between the observed values and the predicted values.

The slope coefficients (β1, β2, …, βn) represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant

3. ADVANTAGES & LIMITATIONS

Advantages

- Simplicity and interpretability.
- Efficient estimation methods for large datasets.

Limitations

- Assumes a linear relationship, which may not hold.
- Relies on assumptions of independence and homoscedasticity
- Sensitive to outliers and influential observations.
- Limited flexibility for capturing complex relationships.

Content outline

1. Introduction
2. Dataset
3. Linear Models
4. Autoregressive model
5. ARMA model
6. Experiment and Result

AR MODEL

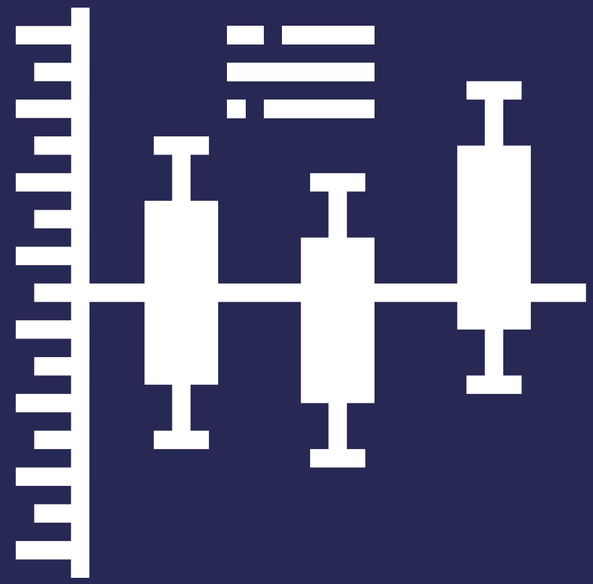
Introduction

The AR (AutoRegressive) model is a powerful tool in time series analysis, enabling prediction and understanding of trends and fluctuations in time series data. Widely applied, the AR model plays a crucial role in supporting decision-making and forecasting across various fields.

AR MODEL

Definition

The AR (Autoregressive) model specifies that the output variable depends linearly on its own previous values and a stochastic term.



Equation

The notation AR(p) indicates an autoregressive model of order p. The AR(p) model equation is:

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

- Xt is the value of the variable at time t.
- c is the constant term (intercept) or the mean value of the time series.
- φi are the autoregressive coefficients corresponding to the previous values X(t-i)
- ε is the stochastic term at time t, representing the unpredictable component.

AR MODEL

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, ...)

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 model and 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.

AR MODEL

Advantages and Limitations

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.

Content outline

1. Introduction
2. Dataset
3. Linear Models
4. Autoregressive model
5. ARMA model
6. VAR model
7. Experiment and Result

ARMA MODEL

INTRODUCTION

HOW TO BUILD GOOD MODEL

ADVANTAGE & LIMITATION

1.INTRODUCTION

ARMA model is popular time series forecasting method. It can ”explain” a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that it can be used to forecast future values.

2. OVERVIEW OF ARMA MODEL

2.1 Autoregressive model (AR model)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \varepsilon_t$$

2.2 Moving averge model (MA model)

$$Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_p \varepsilon_{t-q} + \varepsilon_t$$

2.3 ARMA model

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_p \varepsilon_{t-q} + \varepsilon_t$$

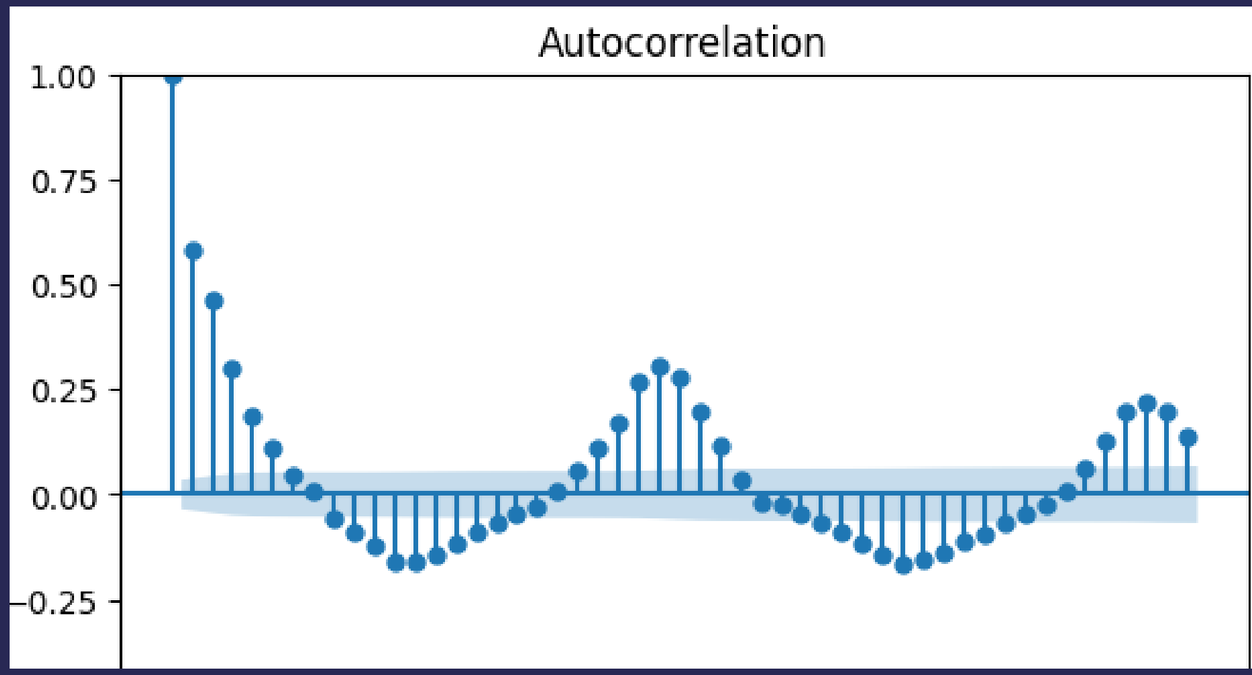
3. BUILD GOOD ARMA MODEL

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).

3.1 Determine a MA model order with the ACF

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.

We will select the order q for model M A(q) from ACF if this plot has a sharp cut-off after lag q.

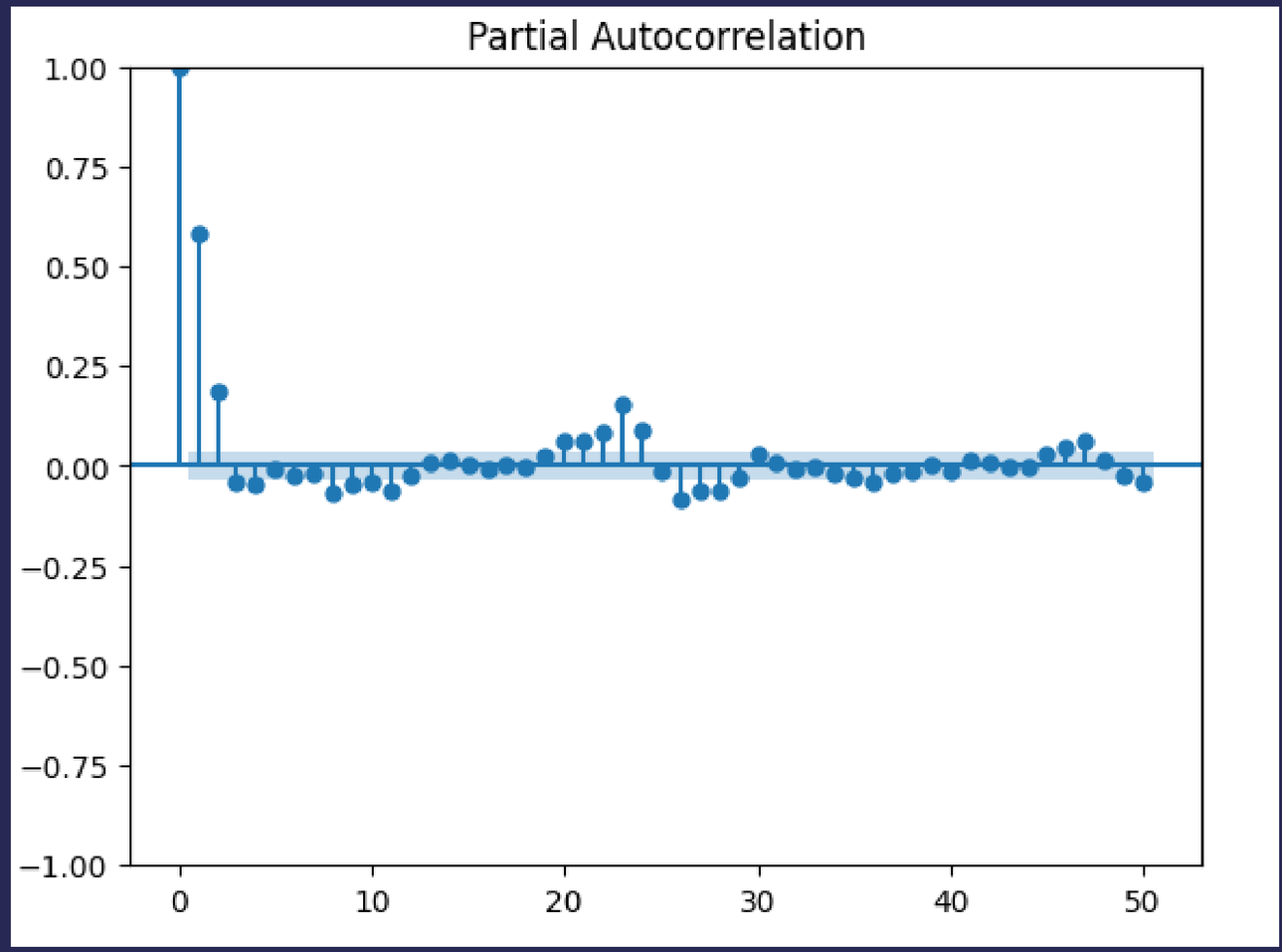


3. BUILD GOOD ARMA MODEL

3.2 Determine an AR model order with the PACF

PACF is same as ACF. But now, instead of concerning in both ”direct” and ”indirect” correlation, PACF only focus on ”direct” correlation. From PACF we can choose lag value which is a good predictor for future value.

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.



4. ADVANTAGE & LIMITATION

4.1 Advantage

- With ARMA model we can predict with high accuracy and be considered as the most popular forecasting method for stationary time series.

4.2 Limitation

- ARMA model only work for stationary time series and 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

Content outline

1. Introduction
2. Dataset
3. Linear Models
4. Autoregressive model
5. ARMA model
6. **VAR model**
7. Experiment and Result

VAR MODEL

Definition

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.

Properties

- VAR model is a generalized form of AR 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

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

VAR MODEL

Properties

- The general VAR model includes n variables, m lags.
- In practice, it is common to keep n small and include only those variables that are highly correlated to reduce time complexcity.
- Variables can be estimated by OLS with the objective of minimizing the error value

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

- VAR model is only work with stationary variables so we have to check the the stationarity.
- For non-stationary attributes, we take first difference and then these time series become stationary.

CONTENT OUTLINE

1. Introduction
2. Dataset
3. Linear Models
4. Autoregressive model
5. ARMA model
6. VAR mode
7. **Experiment and Result**

7. EXPERIMENT AND RESULT

Evaluation metrics: RMSE and MAE

	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 VAR model shows the highest RMSE and MAE values
- The AR and ARMA models are the most accurate options

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/

THANK YOU FOR LISTENING!