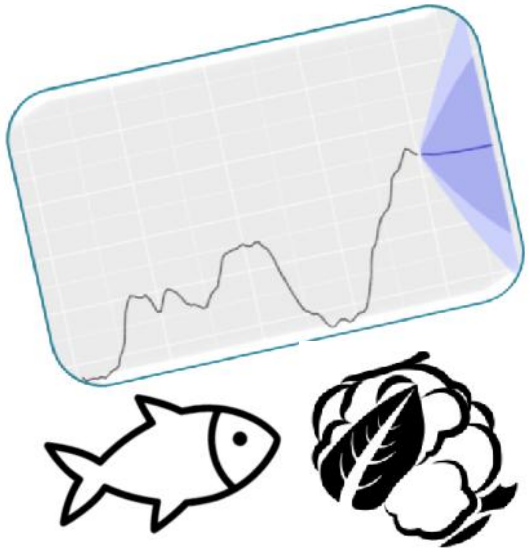
# FORECASTING PRICES OF FISH AND VEGETABLES USING WEB SCRAPED PRICE MICRODATA IN MALAYSIA: AN ARIMA APPROACH

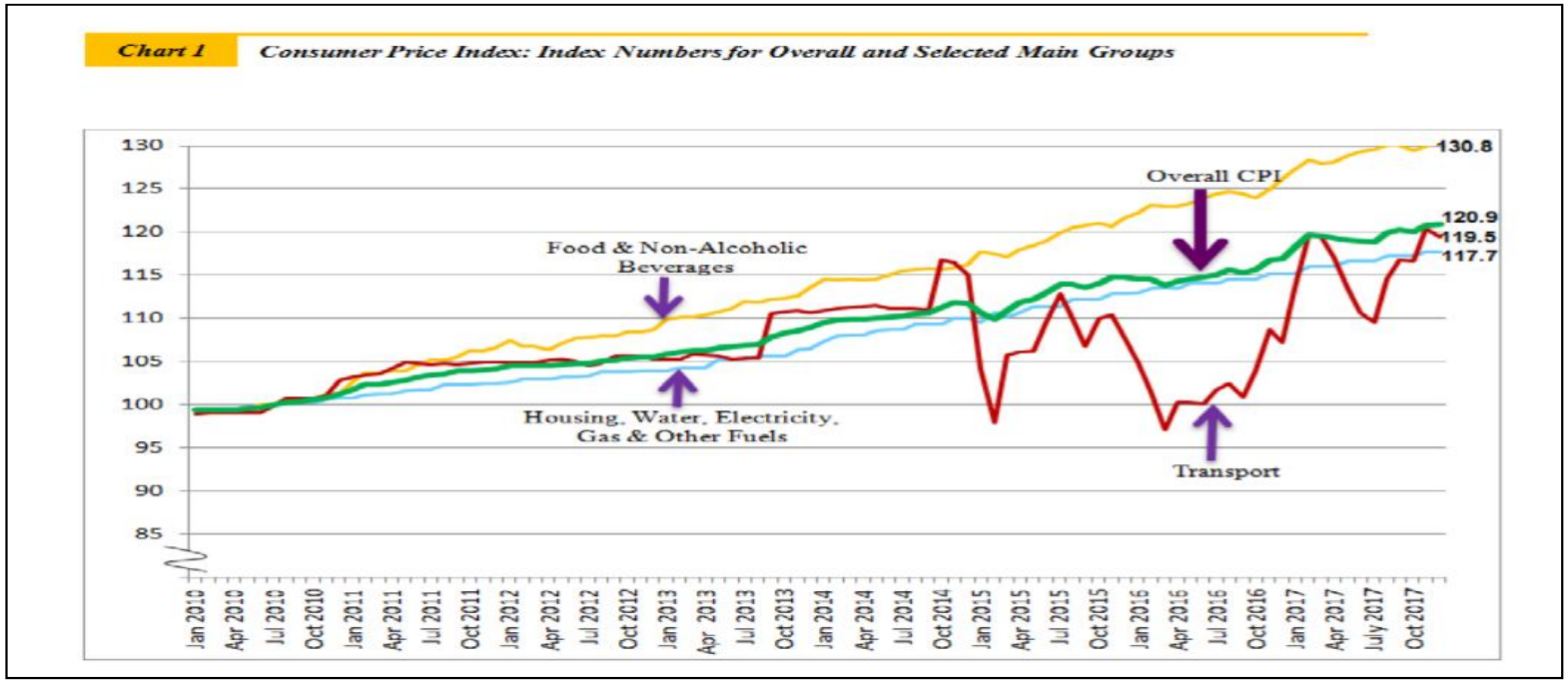


**Mazliana binti Mustapa**
Bahagian Metodologi dan Penyelidikan
Sesi Kolokium dan Research Poster 2018
**5 Oktober 2018**

# PRESENTATION OUTLINE

1. Introduction
2. Literature Review
3. Methodology
4. Implementation
5. Results
6. Conclusions

# 1. INTRODUCTION




- Consumer Price Index (CPI) in Malaysia is the price index used as a proxy for inflation in Malaysia

- It is compiled monthly based on the price data collection, which is conducted by the Department of Statistics, Malaysia (DOSM).

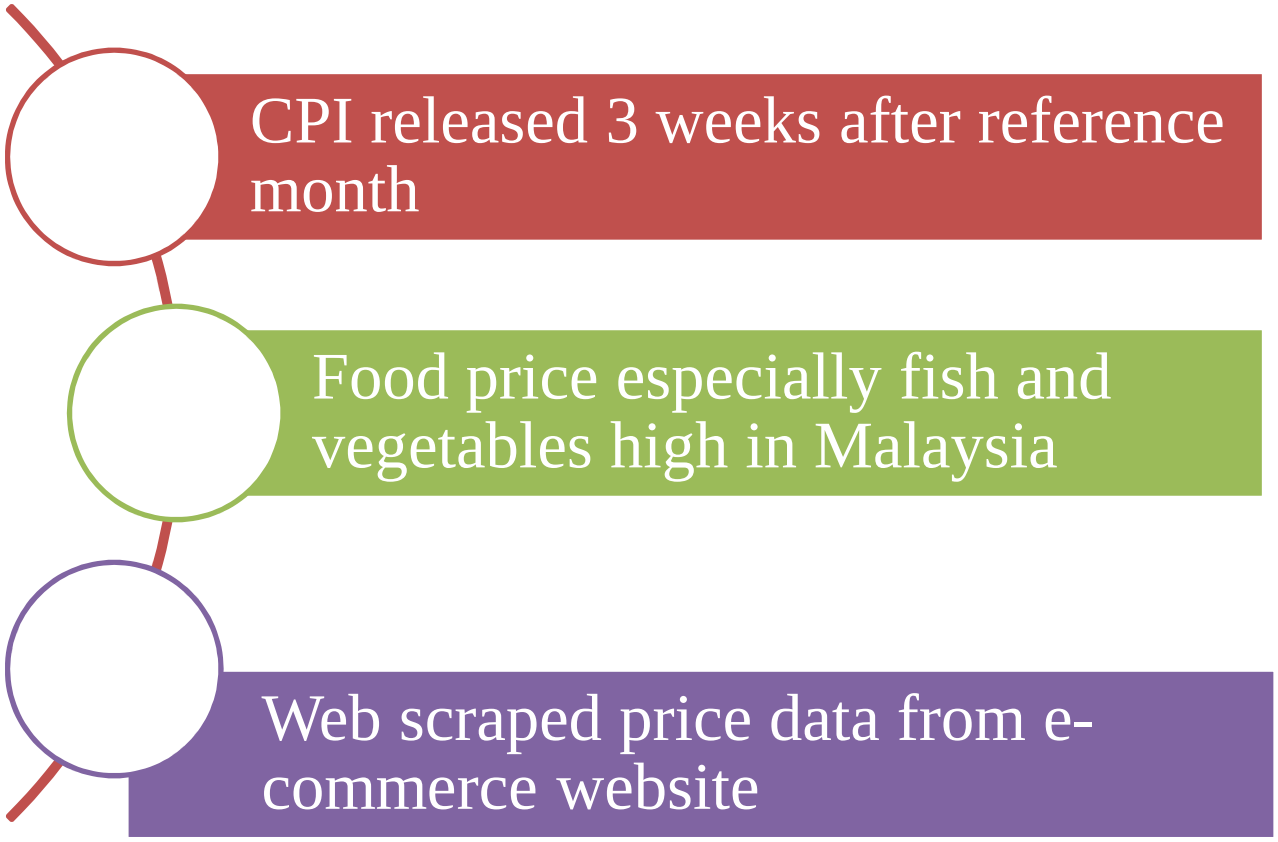
# 1. INTRODUCTION

- E-commerce website is growing with the trends of consumers purchase online.


- Web scrapping is the process of getting information from the website using special tools named web scraper. The web scraped data has the possibility to become new source of compiling the CPI.

- Forecasting price using the web scraped data helps the official statistics office to predict future value and can be used to control the situation of supply and demand side.

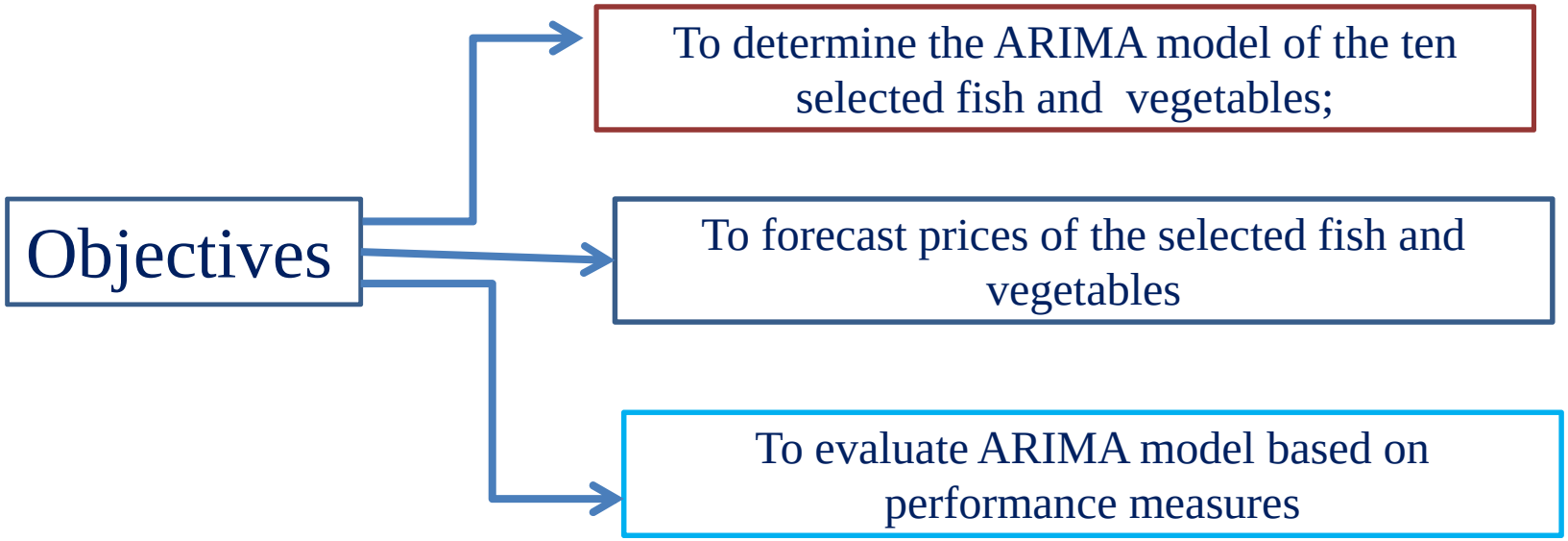
# PROBLEM STATEMENT




Thus, this study aims to utilize the web scraped data in forecasting ten selected fish and vegetables in Malaysia using ARIMA approach

# AIM

To utilize the web scraped price data as an alternative data source and to forecast or predict the selected fish and vegetables using time series forecasting technique.

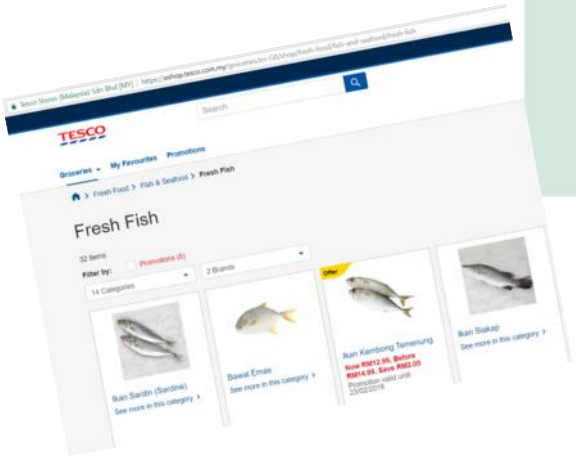


# SCOPE OF THE RESEARCH

1

Tesco e-shop

Fish & Vegetables




2

July – November 2017

Web Scraped data

# SIGNIFICANCE

Improve effective use of updated data to determine better estimations on product pricing

To reduce establishment burden

Improve the scope of the CPI baskets and give better statistics.

To generate automated signal or alert when the price(s) of products are increasing

# 2. LITERATURE REVIEW

| Author | Title | Findings |
|---|---|---|
| Cavallo & Rigobon (2016) | The Billion Prices Project: Using Online Prices for Measurement and Research | Online price offers many benefits such as real-time, the frequency of data is daily basis, economical, full of information, easy to collect anywhere and comparable in the country |
| Polidoro et al. (2015) | Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. Statistical | The outcome of this research shows a good results where the process of data collection becomes efficient. Using the web scraped data helps decrease of the error of measurement and sampling error |
| Mayhew (2016) | Imputing Web Scraped Prices | The results of the simulation study, the geometric average growth is recommended to impute price as the data is part of CPI. The author recommends the number of days for the imputation of the price. |

# 2. LITERATURE REVIEW

| Author | Title | Findings |
|---|---|---|
| Metcalfe et al. (2016) | Research indices using web scraped price data: clustering large datasets into price indices | The Clustering Large dataset into Price Indices (CLIP) is to produce price index with web scraped data. CLIP has a limitation where it depends on the product availability. |
| Powell et al. (2017) | Tracking and modelling prices using web-scraped price microdata: towards automated daily consumer price index forecasting | Hyperparameters estimates are produced to tackle the different product prices of 33 food item products of CPI. To determine the correlation web scraped price data with the normal price collection survey. |
| Cavallo (2013) | Online and official price indexes: Measuring Argentina's inflation | The online price index, which is a mixture of official methods and web price data. The author claims that the online price index is equivalent to official price statistics (inflation) in Colombia, Venezuela, Argentina and Brazil. |

# 2. LITERATURE REVIEW

| Author | Title | Findings |
|---|---|---|
| Chuanyang & Joseph (2016) | Experiences with the use of Online Prices in Consumer Price Index. | Two issues were discussed along the pilot work which are no consistency of product type and lack of staff, which do not have the capability to handle the web crawler. |
| Breton et al., (2016) | Research indices using web-scraped data | The Gini, Eltetö, Köves and Szulc (GEKS) index is proposed to overcome the chain issue. The GEKS is better than the unit price index |

# 2. LITERATURE REVIEW

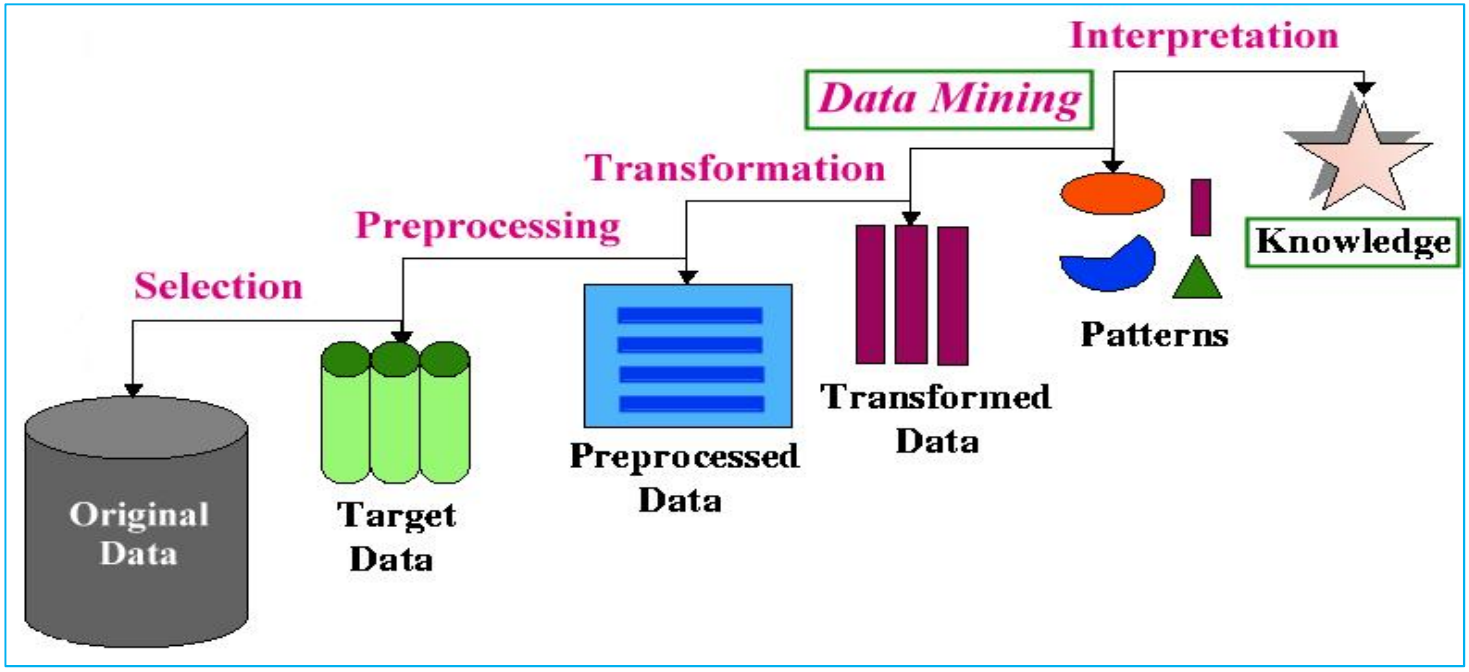
| Author | Title | Time Period | Findings |
|---|---|---|---|
| Michinaka et al., (2016) | Forecasting Monthly Prices of Japanese Logs | Jan 2002 – Sept 2015 (monthly)<br>- 165 observations | The aim of the research is to forecast monthly price of the selected logs of 6 and 12 months ahead. Based on the accuracy, ARIMA is outperformed as compare to ETS method. |
| Paul, Hoque & Rahman, (2013) | Selection of Best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh: A Case Study on Square Pharmaceutical Ltd. | 2011 (236 working days) | The best ARIMA model is determine using the AME, RMSE,AIC, MAPE, SIC, AICc based on least value of the measures. The best model for forecasting share price is ARIMA (2, 1, 2). |
| Abdullah (2012) | ARIMA Model for Gold Bullion Coin Selling Prices Forecasting | 2002 - 2007 (daily selling price) | The ARIMA (2, 1,2) model is outperformed with the minimum error 10%. |

# 3. METHODOLOGY

- Knowledge Discovery in Database (KDD)
- Tools : R Studio
- Forecasting Technique: ARIMA
- Statistical Analysis
- Performance Measure

# 3. METHODOLOGY

Knowledge Discovery in Databases (KDD)



# 3. METHODOLOGY

Tools: R Studio

1. Package ‘tseries’
2. Package ‘auto.arima’
3. Package ‘forecast’
4. Statistical Test – Box.test( ) : Ljung Box test & adf.test( ) : Augmented Dickey Fuller

# MODEL SPECIFICATION

**ARIMA (p,d,q)**

$$\phi_p(B)\nabla^d Y_t = \delta + \theta_q(B)\varepsilon_t$$

where :

$\nabla^d$ = regular/non - seasonal differences

$\delta$ = constant

$Y_t$ = time series data

$\varepsilon_t$ = white noise/random error

$\phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$

$\theta_q(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$

# 3. METHODOLOGY

Statistical Analysis

| Augmented Dickey-Fuller test | Ljung Box test |
| --- | --- |
| • To check stationary of time series data | • To check model adequacy |

# 3. METHODOLOGY

## Performance Measure

1

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

Root Mean Squared Error (RMSE)

2

$$MAE = \frac{\sum_{t=1}^{n} | e_t |}{n}$$

Mean Absolute Error (MAE)

3

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{Y_t}\right| \times 100\%$$

Mean Absolute Percentage Error (MAPE)

# 4. IMPLEMENTATION

## 1 Data

- Tesco e-shop website

## 2 Period

- 1 July – 30 November 2017

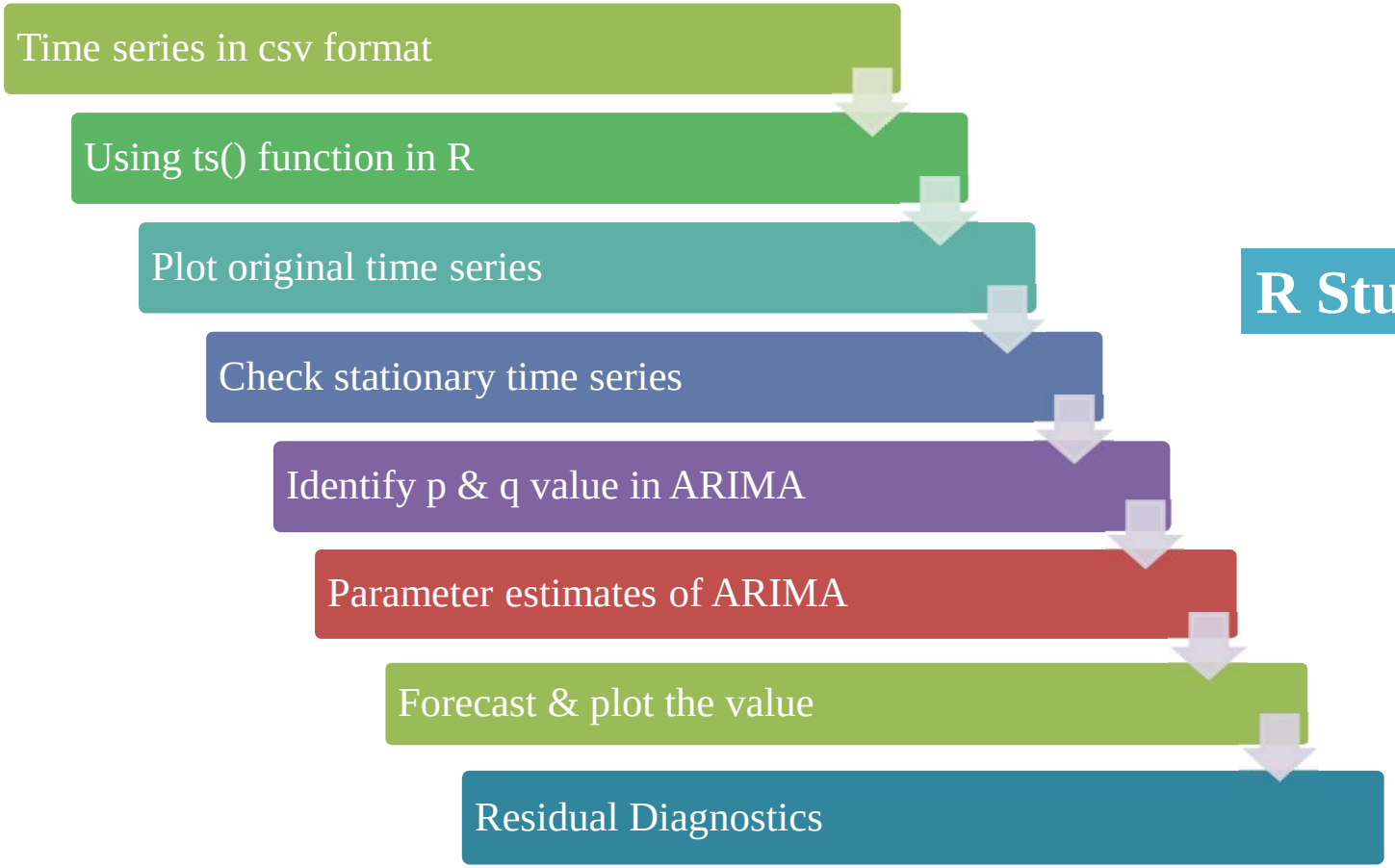
## 3 Variable

- Price

## Data

| No. | Item | Unit of Measurement | Number of Observation |
|---|---|---|---|
| **1** | Bawal | per pieces | 153 |
| **2** | Cencaru | per pieces | 153 |
| **3** | Kembong | per pieces | 153 |
| **4** | Red Bream | per pieces | 153 |
| **5** | Selar Kuning | per pieces | 153 |
| **6** | Green Spinach | per bundle | 153 |
| **7** | Kangkung | per bundle | 153 |
| **8** | Long Beans | per 100g | 153 |
| **9** | Round Cabbage | per pieces | 153 |
| **10** | Sawi Jepun | per 100g | 153 |

# 4. IMPLEMENTATION




**R Studio**

# 5. RESULTS

| No. | Item | ARIMA (p,d,q) |
|---|---|---|
| **1** | Bawal | ARIMA (0, 1 ,0) |
| **2** | Cencaru | ARIMA (0, 1 ,0) |
| **3** | Kembong | ARIMA (0, 1 ,0) |
| **4** | Red Bream | ARIMA (2, 0 ,1) with non-zero mean |
| **5** | Selar Kuning | ARIMA (2, 0 ,1) with non-zero mean |
| **6** | Green Spinach | ARIMA (2, 0 ,1) with non-zero mean |
| **7** | Kangkung | ARIMA (2, 0 ,1) with non-zero mean |
| **8** | Long Beans | ARIMA (1, 0 , 2) with non-zero mean |
| **9** | Round Cabbage | ARIMA (0, 1 ,0) |
| **10** | Sawi Jepun | ARIMA (0, 1 ,0) |

# MODEL ARIMA

Model 1: Red Bream – ARIMA (2,0,1)

$$Y_t = 2.4205 - 0.0333 Y_{t-1} + 0.6382 Y_{t-2} + 0.0594 + 0.9688 e_{t-1}$$

Model 2: Selar Kuning – ARIMA (2, 0, 1)

$$Y_t = 1.4794 - 1.6648 Y_{t-1} - 0.7431 Y_{t-2} + e_t - 0.7924 e_{t-1}$$

Model 3: Green Spinach – ARIMA (2,0,1)

$$Y_t = 1.1813 + 1.7246 Y_{t-1} - 0.7774 Y_{t-2} + 0.1008 - 0.8126 e_{t-1}$$

Model 4: Kangkung – ARIMA (2, 0,1)

$$Y_t = 1.0995 + 1.7293 Y_{t-1} - 0.7776 Y_{t-2} + 0.1023 - 0.8315 e_{t-1}$$

# MODEL ARIMA

Model 5: Long Beans – ARIMA (1,0,2)

$$Y_t = 0.5460 + 0.3667\, Y_{t-1} + 0.0780 + 0.4322\, e_{t-1} + 0.2638\, e_{t-2}$$

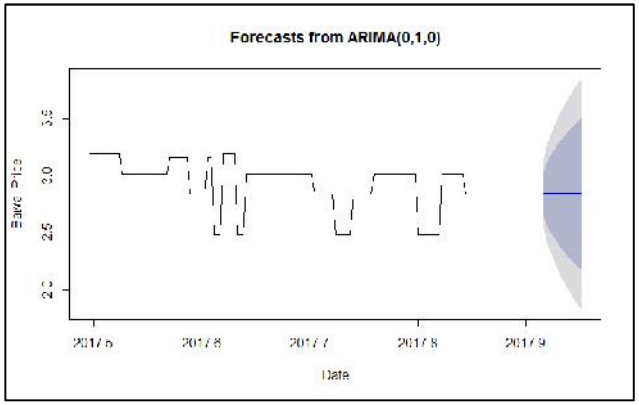
Model 6: Bawal, Cencaru, Kembong, Round Cabbage and Sawi Jepun (ARIMA(0, 1,0)

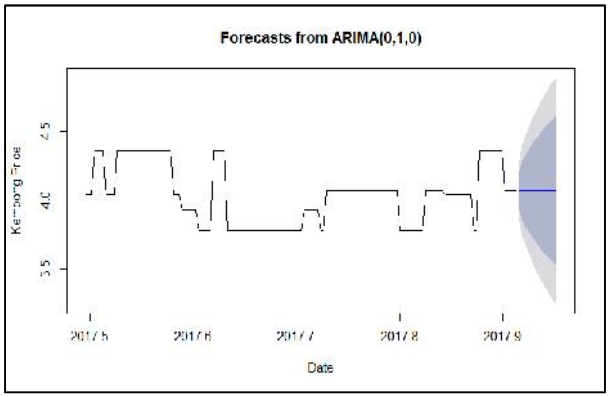
$$Y_t = \mu + Y_{t-1}$$

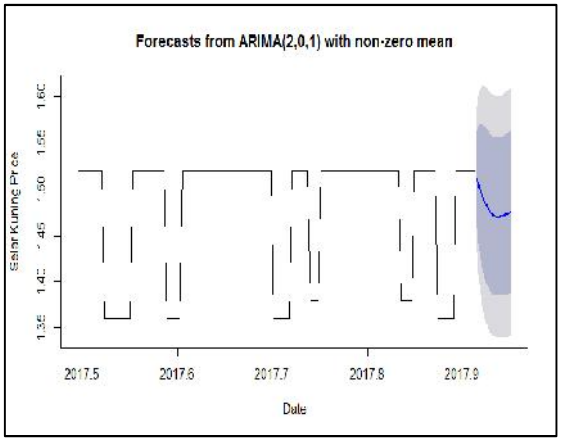
where $\mu$ : mean of the changes of period to period

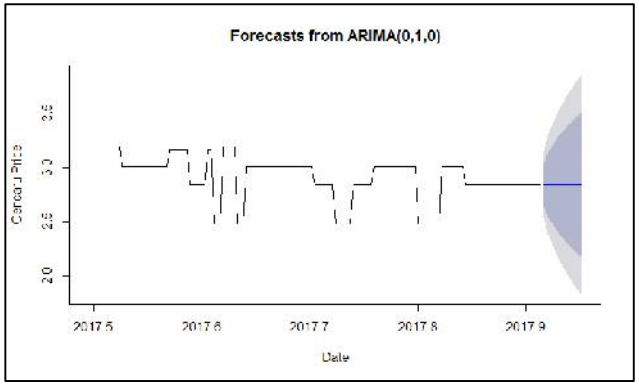
# FORECAST PRICES

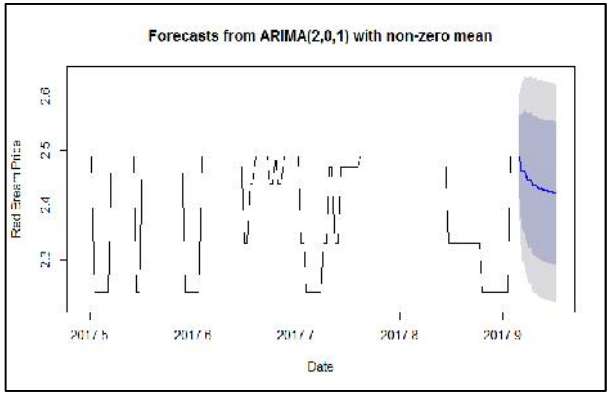
**Bawal, Cencaru, Kembong, Red Bream, Selar Kuning, Green Spinanch**

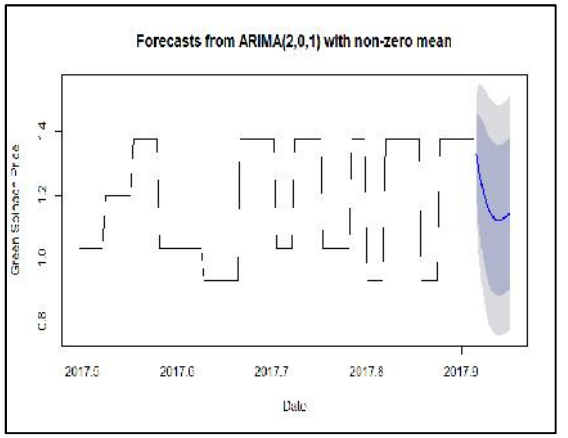


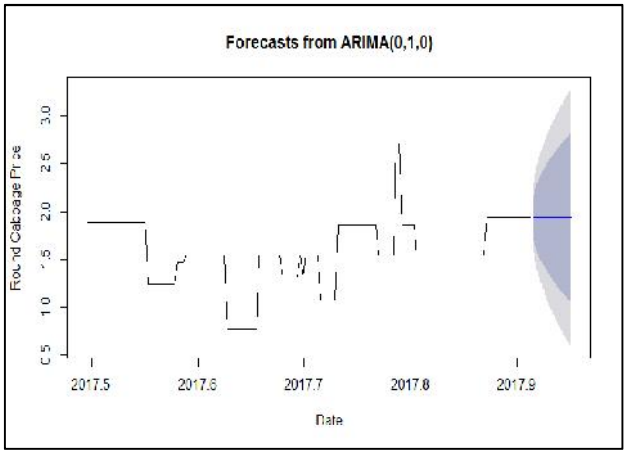



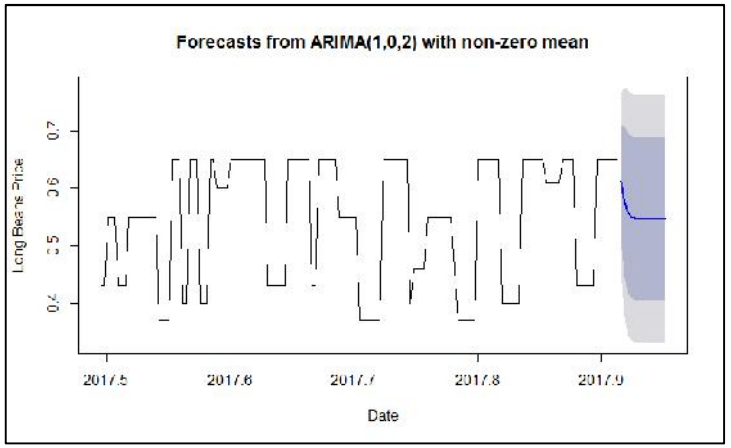



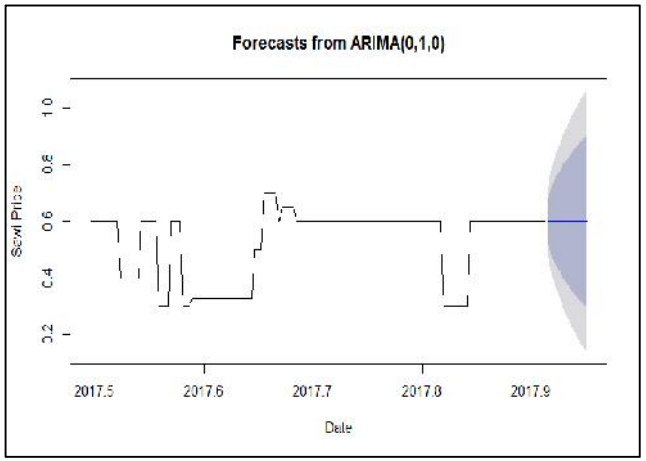



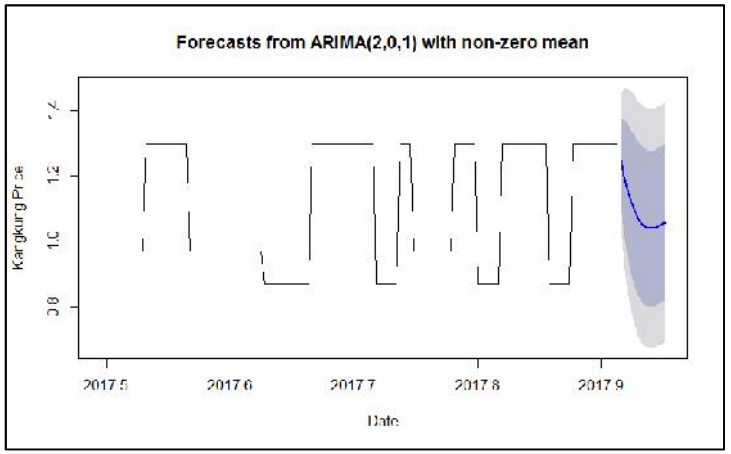
# FORECAST PRICES

**Round Cabbage, Sawi Jepun, Long Beans, Kangkung**



# FORECASTED PRICE FOR NEXT 14 DAYS : FISH

| Date | Bawal | Cencaru | Kembong | Red Bream | Selar Kuning |
|---|---|---|---|---|---|
| **2017-12-01** | 0.60 | 2.84 | 4.07 | 2.49 | 1.51 |
| **2017-12-02** | 0.60 | 2.84 | 4.07 | 2.46 | 1.50 |
| **2017-12-03** | 0.60 | 2.84 | 4.07 | 2.46 | 1.49 |
| **2017-12-04** | 0.60 | 2.84 | 4.07 | 2.45 | 1.49 |
| **2017-12-05** | 0.60 | 2.84 | 4.07 | 2.45 | 1.48 |
| **2017-12-06** | 0.60 | 2.84 | 4.07 | 2.44 | 1.48 |
| **2017-12-07** | 0.60 | 2.84 | 4.07 | 2.44 | 1.47 |
| **2017-12-08** | 0.60 | 2.84 | 4.07 | 2.43 | 1.47 |
| **2017-12-09** | 0.60 | 2.84 | 4.07 | 2.43 | 1.47 |
| **2017-12-10** | 0.60 | 2.84 | 4.07 | 2.43 | 1.47 |
| **2017-12-11** | 0.60 | 2.84 | 4.07 | 2.43 | 1.47 |
| **2017-12-12** | 0.60 | 2.84 | 4.07 | 2.42 | 1.47 |
| **2017-12-13** | 0.60 | 2.84 | 4.07 | 2.42 | 1.47 |
| **2017-12-14** | 0.60 | 2.84 | 4.07 | 2.42 | 1.48 |

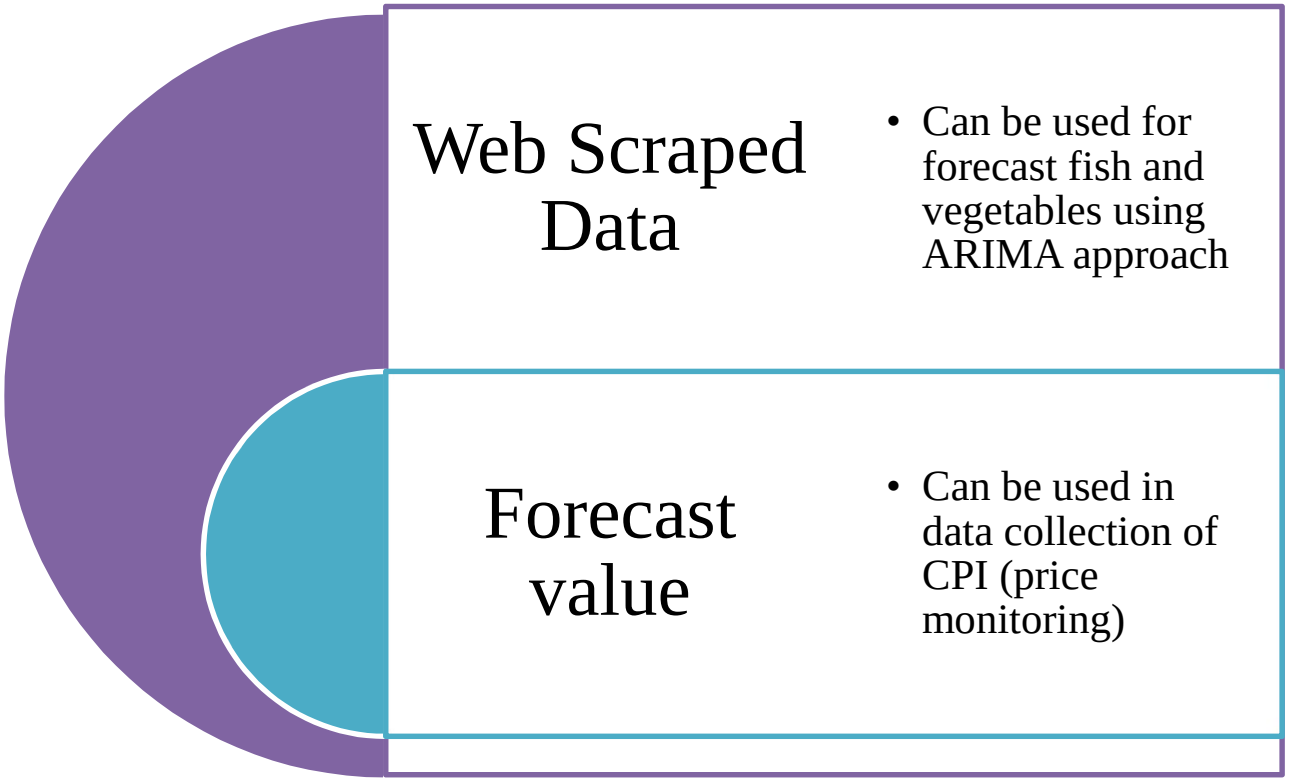
# FORECASTED PRICE FOR NEXT 14 DAYS : VEGETABLES

| Date | Green Spinach | Kangkung | Long Beans | Round Cabbage | Sawi Jepun |
|---|---|---|---|---|---|
| **2017-12-01** | 1.33 | 1.25 | 0.61 | 1.94 | 0.60 |
| **2017-12-02** | 1.28 | 1.20 | 0.58 | 1.94 | 0.60 |
| **2017-12-03** | 1.24 | 1.16 | 0.56 | 1.94 | 0.60 |
| **2017-12-04** | 1.20 | 1.12 | 0.55 | 1.94 | 0.60 |
| **2017-12-05** | 1.17 | 1.09 | 0.55 | 1.94 | 0.60 |
| **2017-12-06** | 1.15 | 1.07 | 0.55 | 1.94 | 0.60 |
| **2017-12-07** | 1.14 | 1.06 | 0.55 | 1.94 | 0.60 |
| **2017-12-08** | 1.13 | 1.05 | 0.55 | 1.94 | 0.60 |
| **2017-12-09** | 1.12 | 1.04 | 0.55 | 1.94 | 0.60 |
| **2017-12-10** | 1.12 | 1.04 | 0.55 | 1.94 | 0.60 |
| **2017-12-11** | 1.12 | 1.04 | 0.55 | 1.94 | 0.60 |
| **2017-12-12** | 1.13 | 1.05 | 0.55 | 1.94 | 0.60 |
| **2017-12-13** | 1.14 | 1.05 | 0.55 | 1.94 | 0.60 |
| **2017-12-14** | 1.14 | 1.06 | 0.55 | 1.94 | 0.60 |

# PERFORMANCE MEASURES

| No. | Item | RMSE | MAE | MAPE |
|---|---|---|---|---|
| 1 | Bawal | 0.32 | 0.09 | 1.12 |
| 2 | Cencaru | 0.14 | 0.04 | 1.39 |
| 3 | Kembong | 0.11 | 0.03 | 0.81 |
| 4 | Red Bream | 0.06 | 0.03 | 0.81 |
| 5 | Selar Kuning | 0.04 | 0.02 | 1.57 |
| 6 | Green Spinach | 0.10 | 0.05 | 4.67 |
| 7 | Kangkung | 0.10 | 0.05 | 5.07 |
| 8 | Long Beans | 0.08 | 0.05 | 10.57 |
| 9 | Round Cabbage | 0.18 | 0.05 | 3.43 |
| 10 | Sawi Jepun | 0.06 | 0.02 | 3.84 |

# 6. CONCLUSION



# THANK YOU