MALAYSIAN AGE STRUCTURE FORECAST USING THE LOG-RATIO TRANSFORMATION AND ARIMA MODELS

LOW YEE SEN

Azizi Abdullah

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRACT

Ageing population is a concerning topic. It effects the operation of a country from the angle of society, culture, economic and so on. In order to well prepared about ageing population, age structure of a country need to be observed. There are many choices can be chosen for time series forecasting and ARIMA is chosen based on literature review on this paper. Past papers had used isometric log ratio (ILR) transformation to transform compositional data before prediction. In this paper, summated log ratio(SLR) is proposed to compared with ILR for the usage of transform compositional data before prediction. Function for summated log ratio transformation and inverse transformation built as no existing python library contain this function. Performance of optimize ARIMA model using auto ARIMA and grid search ARIMA is also investigated. For the result, the ratio of elder is expected to double in the next twenty years. One of the SLR, SLR2 proposed in this paper did perform better compared with ILR and original data. Grid search is also seen as a better optimization method for ARIMA compared with auto ARIMA in this paper.

1 INTRODUCTION

With better health care, population nowadays is expected to live longer. Proportion of old people in many countries also increase due to lower birth rate and other reasons. Based on a online journal from World Health Organization website, one out of six person will be aged more than sixty by 2030. World’s population of people aged sixty and above will double which is 2.1 billion when compare 2020 with 2050. This phenomena which composition of population shift to older people is called population aging. Some high income countries like Japan already experience this phenomena with 30 percent population over sixty years old. Two-thirds of whole world’s population over sixty years old predicted locate at middle income and low income countries. Therefore, as a middle income country, Malaysia should be well aware of population aging. (WHO 2022) Based on the investigation of Population Division, United Nations in 2018, ageing population is a phenomena obvious in high and middle income country compared with low income country. The reasons of population ageing happen is not necessary a bad thing. Population ageing is an achievement of advance society.

(Lloyd-Sherlock P, McKee M, Ebrahim S, Gorman M, Greengross S, Prince M, et al. 2021) People nowadays generally live healthier compared to the past, this lead to longer live expectancy at both birth and older ages. More educated population start to utilize protection when undergo sexual intercourse, able to access contraception and giving birth according to plan, this lead to lower fertility and birth rate. Although the reasons are good reasons, impacts of population ageing cannot be underestimated. Several reports show that the chance is population aging and human development has a negative relationship is high. (Maestas N, Mullen KJ, Powell D. 2016) Several ways population ageing influences human development. Old people is relatively less economically active when compare side by side with young people. Working population shrinking, this leads to shortage of workers and demand roles are harder to be filled in. There are consequences, if the hole of human resource cannot be fulfilled. These include poor productivity, shrinking business, declining market competitiveness. Labor costs are also expected to be higher and while this happen, companies might increase their products price to compensate increased labour expenses eventually lead to inflation. Population savings and consumption patterns are also altered, market have to build different financial infrastructure and social security systems to deal with it. Moreover, government need to increase expenditure on public health to make sure our elders are well taken care. In a report of the effect of population aging on economic growth, the labor force and productivity in 2016, growth rate of GDP has negative relationship with growth rate of elders group. This indicate a society with more elder could have slower economic.

2 PROBLEM STATEMENT

Malaysia is a middle-income country and it is hoped that this country will become a high-income country in the near future. Malaysia should be well prepared about becoming a high income country. Population aging is one of the phenomena that occurs in all high-income countries. Knowing about population distribution can be useful for long-term planning and policy implementation. The aim of this study is to predict the composition of the Malaysian population by age using those aged less than 15 years, between 15 to 64 years and 65 years and above. Each of these proportion, represents children, youth and the elderly. There are many choices of models that can be useful for time series forecasting. But many aspects can be investigated before choosing one. Is the selected model suitable for the target data size?

Are multivariate time series better than univariate time series forecasting? What models generally perform well?

Age structure data is compositional data. Compositional data is compositional data is data that describes the relative proportions of different components within a whole. Compositional data arises in various fields, including biology, geology, and environmental science, among others. Log transformation is one of the practice to transform compositional data before prediction to prevent compositional data structures from breaking after prediction. Among the papers reviewed, the isometric log ratio (ILR) was used for transformation. However, SLR transformation method might do a better job of transformation base on the papers. ILR transformation that applied on compositional data before prediction is used for contrast groups of parts and preserve the relationship between ratio. In paper, 'The isometric log ratio transformation in compositional data analysis: a practical evaluation' SLR is described as a more practical log ratio transformation method.

The ILR transformation is transformation represented by correspond geometric mean for the ratio and scalar constant. It can maps compositional data into a new space where the differences between compositions are proportional to the log-ratios of the components. This has the advantage of preserving the ratios between the components and making it easier to compare compositions. However, in examples given in the paper, not only ILR do not preserve the relationship as well, it has completely wrong interpretation of group contrast after transformation.

SLR transformation do not use any scaling factor or geometric mean, it simply log ratio of sum of components. The SLR transformation is similar to the ILR transformation, but it maps compositional data into a space where the distances between compositions are proportional to the symmetric log-ratios of the components. This has the advantage of making it easier to interpret the results and making it easier to perform statistical analysis. Compared with ILR, SLR in the other hand preserve the relationship better and accurately contrast the group in examples of the paper. This happen because compared with straightforward contrast in SLR, geometric mean and scalar constant in ILR formula might alter the contrast. Therefore, ILR can be substituted by SLR with a clearer and unambiguous interpretation.

3 RESEARCH OBJECTIVE

In general, this project has three purposes. To examine possible future age structures for better long-term plans using ARIMA models. Develop SLR transform functions that may be better than ILR over compositional data and transform those features into a new feature space to improve the prediction of compositional data. Measure model performance with different optimization methods and different types of data transformations.

4 METHODOLOGY

Methodology is important for any study. It shows how the study should be conducted. This is to ensure that the study runs smoothly, is organized and can meet the required specifications. In this study, there are six phases. This phase includes the problem identification phase, the data collection phase, the data pre-processing phase, the model development phase,the prediction phase.

4.1 Problem Identification Phase

Every mainstay in all study is always identify problem that need be solved. This study want to forecast the composition of Malaysia of population by age using machine learning methods. In this phase, what approach to choose the algorithm to build the model, what variables to use should be identified. This phase also determine the study objective as well as problem statement.

4.2 Data Collection Phase

Two source of data will be used here. World bank contain population related data. It will be downloaded in CSV form. Demographics indicators all around the world will be download from Department of Economic and Social Affairs Population Division of United Nation in CSV form.

4.3 Data Preprocessing Phase

In this phase, data cleaning, data transformation, scaling and data splitting. In data cleaning, mortality, fertility, migration rate in United Nations data and composition of Malaysia population by age in will be left, other columns will be dropped. Rows will also be dropped based on the year needed. Data cleaning will deal with missing value too. In data transformation, type of data is examine and changed if needed. In case the value of independent variables are way too different in perspective of value, scaling will be done. In data splitting, data will split into train and test data.

4.4 Model Development Phase

This phase is mainly about optimization. ARIMA model will optimized to predict composition of population by age by using base form and various of transformed data such as ILR, SLR and SLR2. This phase will optimize the model by choosing best performance hyperparameters.

Performance of various models will be evaluated in this phase. Root Mean Squared Error (RMSE) is the metrics will be used to evaluate models. Rolling cross validation will be the method to evaluate the performance of models by comparing actual data and predicted data in test set.

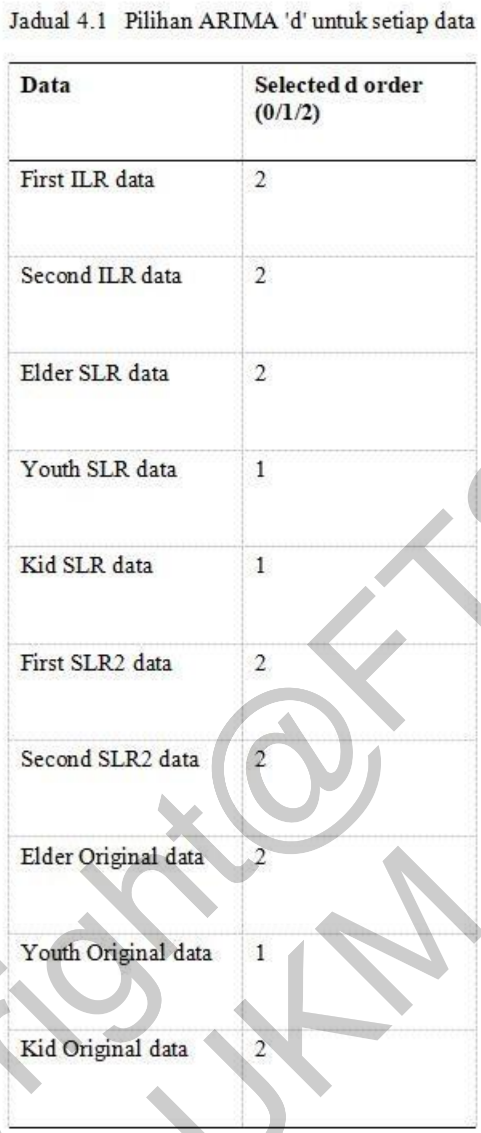
4.5 Prediction Phase

Utilizing the best performance parameters in model development phase, model fit with full data instead of training data. Composition of population by age in next 20 years predicted. The result will be visualized and briefly analyzed.

5 IMPLEMENTATION AND OUTCOME

Forecasting the age structure of Malaysia was developed using the Python programming language. The software used is Visual Studio Code.

First of all, the messy data was cleaned to simplify the model building process. The target variable is transformed into 3 types, namely ILR, SLR and SLR2. To use the ARIMA model, 'd' was determined using the stationarity test.

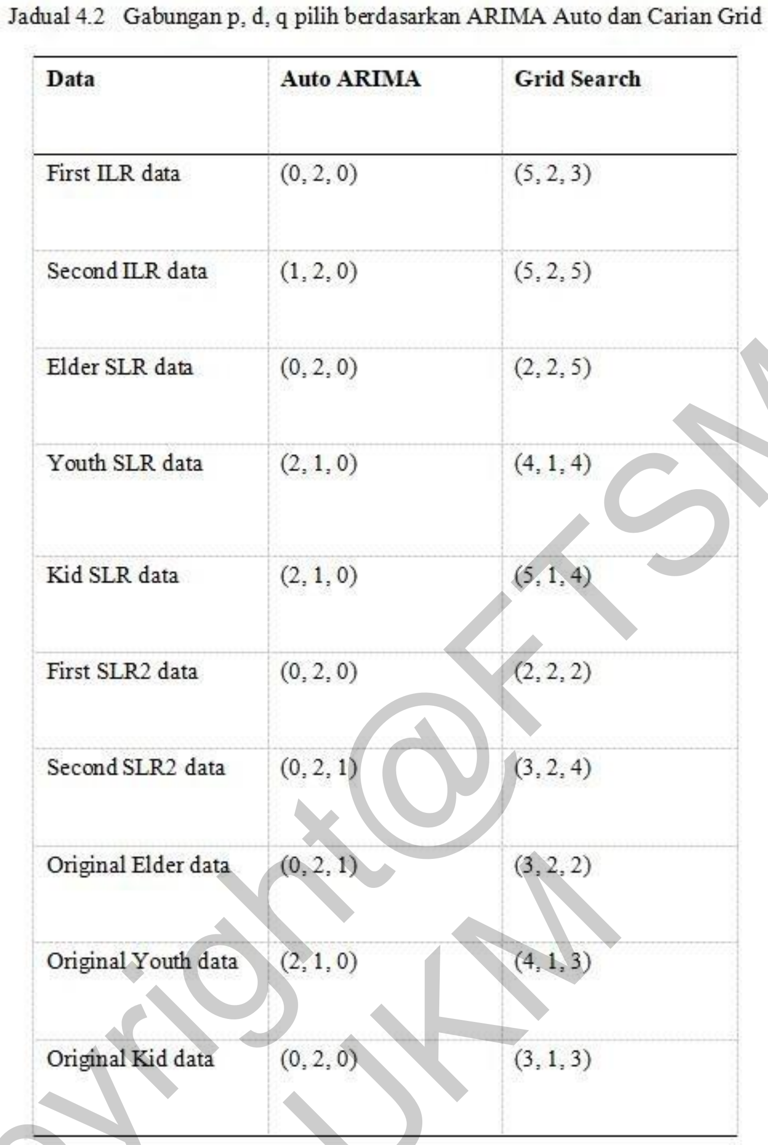


Jadual 4.1 Pilihan ARIMA 'd' untuk setiap data

Data	Selected d order (0/1/2)
First ILR data	2
Second ILR data	2
Elder SLR data	2
Youth SLR data	1
Kid SLR data	1
First SLR2 data	2
Second SLR2 data	2
Elder Original data	2
Youth Original data	1
Kid Original data	2

Figure 1 ‘d’ order determined using stationary test

After determine 'd', 'p' and 'q' are determined using auto ARIMA and grid search optimization.

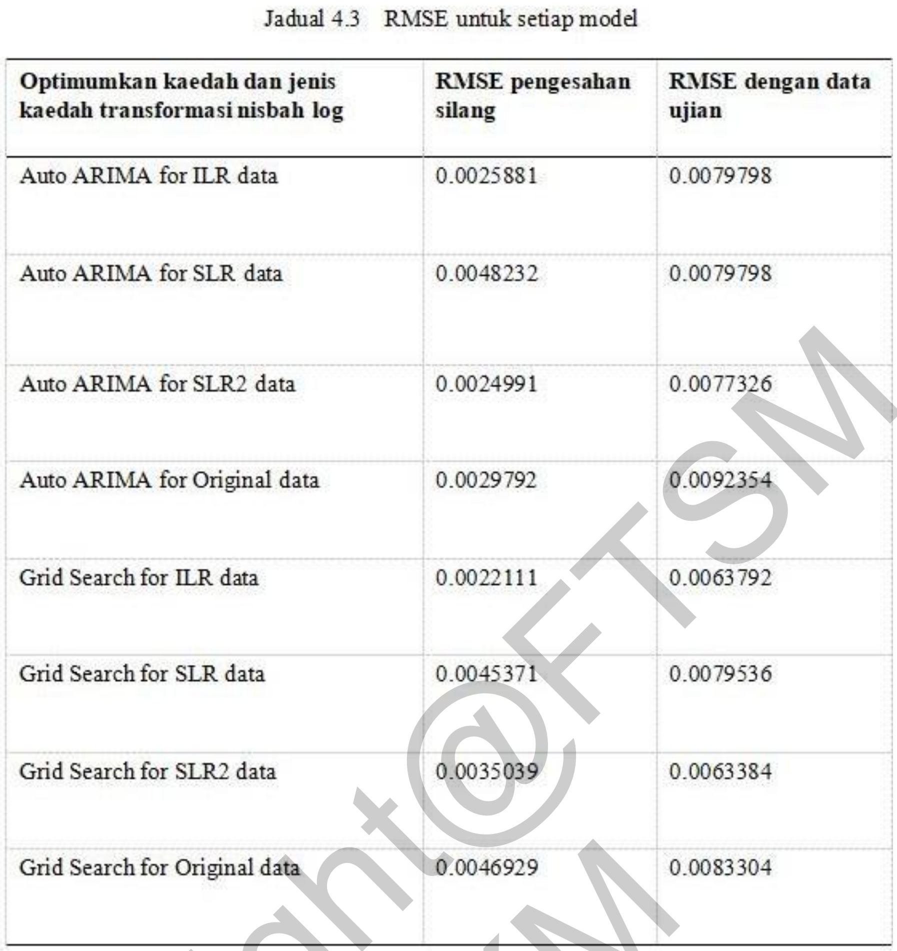


Jadual 4.2 Gabungan p, d, q pilih berdasarkan ARIMA Auto dan Carian Grid

Data	Auto ARIMA	Grid Search
First ILR data	(0, 2, 0)	(5, 2, 3)
Second ILR data	(1, 2, 0)	(5, 2, 5)
Elder SLR data	(0, 2, 0)	(2, 2, 5)
Youth SLR data	(2, 1, 0)	(4, 1, 4)
Kid SLR data	(2, 1, 0)	(5, 1, 4)
First SLR2 data	(0, 2, 0)	(2, 2, 2)
Second SLR2 data	(0, 2, 1)	(3, 2, 4)
Original Elder data	(0, 2, 1)	(3, 2, 2)
Original Youth data	(2, 1, 0)	(4, 1, 3)
Original Kid data	(0, 2, 0)	(3, 1, 3)

Figure 2 ‘p’ dan ‘q’ determined using auto ARIMA and grid search

Using transformed data types and optimization methods will be tested. However, before testing, the data needs to be transformed back. After the data is transformed back, the performance of the model is shown in the form of RMSE.



Jadual 4.3 RMSE untuk setiap model

Optimumkan kaedah dan jenis kaedah transformasi nisbah log	RMSE pengesahan silang	RMSE dengan data ujian
Auto ARIMA for ILR data	0.0025881	0.0079798
Auto ARIMA for SLR data	0.0048232	0.0079798
Auto ARIMA for SLR2 data	0.0024991	0.0077326
Auto ARIMA for Original data	0.0029792	0.0092354
Grid Search for ILR data	0.0022111	0.0063792
Grid Search for SLR data	0.0045371	0.0079536
Grid Search for SLR2 data	0.0035039	0.0063384
Grid Search for Original data	0.0046929	0.0083304

Figure 3 Performance of different transformed data and optimization methods

After ARIMA, the XGBoost model will be built. Before building the ARIMA, the predictor for the XGBoost model will be predicted using the ARIMA model with better optimization methods. Correlations between predictors for XGBoost were investigated. Using predictors and regressor chain strategies, XGBoost with gblinear boosters undergoes cross-validation to select parameters. XGBoost predictions are tested and ARIMA has better performance. Residual analysis was conducted for the ARIMA model.

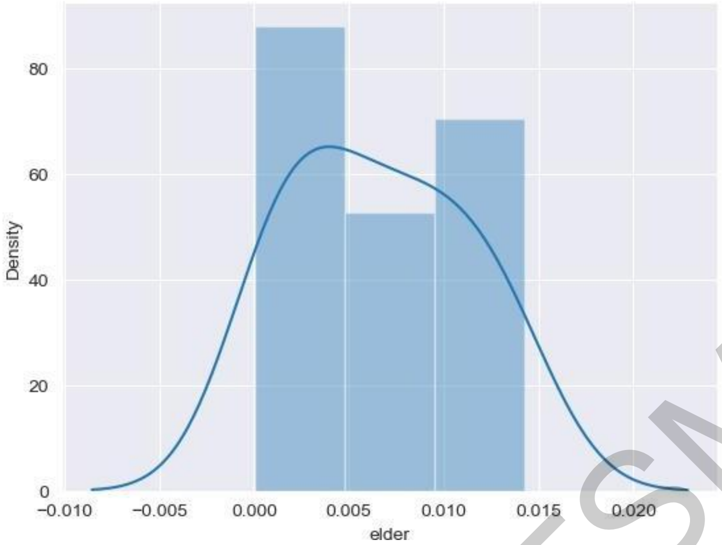


Figure 4 Residual plot for elder proportion

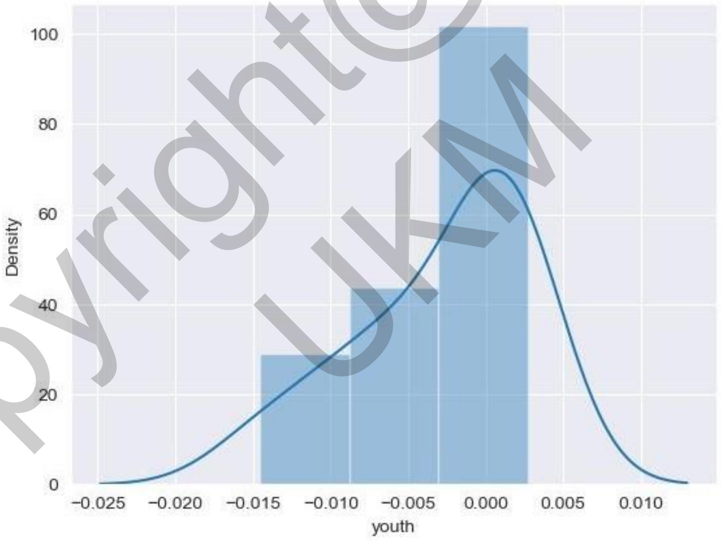


Figure 5 Residual plot for youth proportion

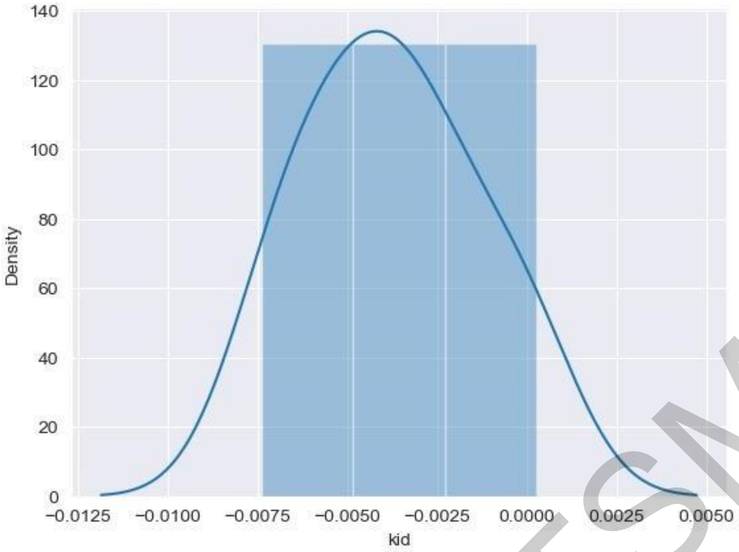


Figure 6 Residual plot for kid proportion

Using ARIMA, final forecasts up to 2041 are made.

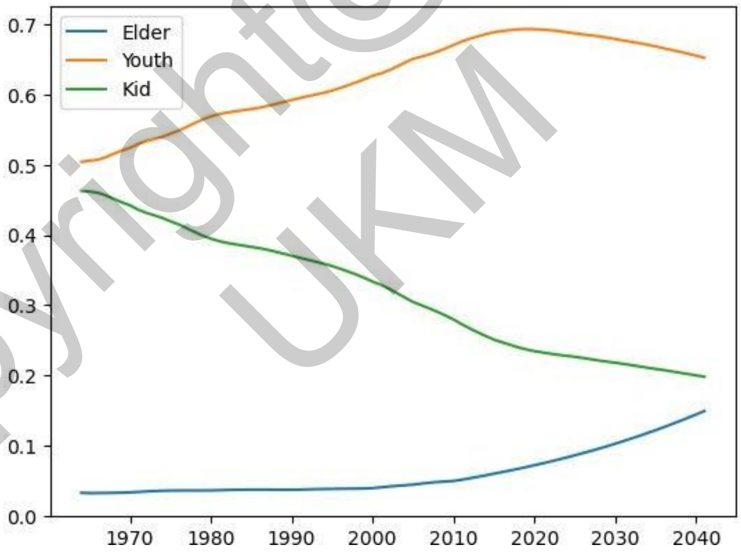


Figure 7 Prediction of population proportion by age group

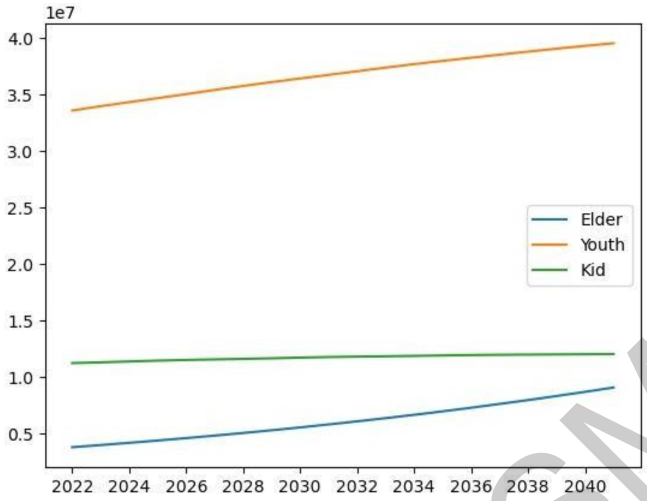


Figure 8 Estimated population

Based on the forecast the rate of elderly people seems to double by 2041.

6 CONCLUSION

Overall, different types of log-ratio transformations of the data are worth trying to make better predictions. Transformation before prediction for compositional data proven to be perform better than base data. Different ways of log ratio transformation able to preserve relationship between ratio differently therefore self made SLR2 perform better than ILR when under the same model, ARIMA. At the same time, apart from ARIMA, ACF, PACF auto graphs which are more often used to optimize ARIMA models, grid search can be tried. Based on that prediction, the doubled composition of elderly people in our communities should be a concern for policymakers.

7 REFERENCE

A.K. Biswas, S. I. Ahmed, T. Bankefa, P. Ranganathan and H. Salehfar. 2021. Performance Analysis of Short and Mid-Term Wind Power Prediction using ARIMA and Hybrid Models. 2021 IEEE Power and Energy Conference at Illinois (PECI), pp. 1-7

Atchadé MN, Sokadjo YM. 2022. Overview and cross-validation of COVID-19 forecasting univariate models. Alexandria Engineering Journal.

AZUAR, A. Z. A. L. E. A. 2022. Malaysia attained ageing nation status - themalaysianreserve.com. Retrieved December 27, 2022, from https://themalaysianreserve.com/2022/10/11/malaysia-attained-ageing-nation-status/

Hongyan Li, Zhong Wu. 2009. The Application of Markov Chain into the Forecast for Population Age Structure in Shanghai. 2009 International Conference on Computational Intelligence and Software Engineering.

Greenacre, Michael & Grunsky, Eric. 2018. The isometric logratio transformation in compositional data analysis: a practical evaluation. 10.13140/RG.2.2.10817.20322.

A. Iwok, A. S. Okpe. 2016. A Comparative Study between Univariate and Multivariate Linear Stationary Time Series Models. American Journal of Mathematics and Statistics, Vol. 6 No. 5, pp. 203-212.

A. S. Abu Amra and A. Y. A. Maghari. 2018. Forecasting Groundwater Production and Rain Amounts Using ARIMA-Hybrid ARIMA: Case Study of Deir El-Balah City in GAZA. 2018 International Conference on Promising Electronic Technologies (ICPET), pp. 135-140

K. Sethi, M. Mittal. 2020. Analysis of Air Quality using Univariate and Multivariate Time Series Models. 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence),pp. 823-827

K. Sreehari, M. Adham, T. D. Cheriya and R. Sheik. 2021. A Comparative Study between Univariate and Multivariate Time Series Models for COVID-19 Forecasting. 2021 International Conference on Computational Performance Evaluation (ComPE), 2021, pp. 485-491

Kuan, M.-M. 2022. Applying sarima, ETS, and hybrid models for prediction of tuberculosis incidence rate in Taiwan. PeerJ, 10. https://doi.org/10.7717/peerj.13117

Li, Pin & Zhang, Jin-Suo. 2018. A New Hybrid Method for China’s Energy Supply Security Forecasting Based on ARIMA and XGBoost. Energies. 11. 1687. 10.3390/en11071687.

Lloyd-Sherlock P, McKee M, Ebrahim S, Gorman M, Greengross S, Prince M, et al. 2012. Population ageing and health. Lancet. 379:1295–6.

Maestas N, Mullen KJ, Powell D. 2016. The effect of population aging on economic growth, the labor force and productivity. Natl Bureau Econ Res.

Maxim Andreevich Novak, Elena Ivanovna Kozlova. 2019. Application of the Method of Moving Ages to Predict the Population of the Lipetsk Region. 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA)

M. Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe and S. Bhirud. 2017. Forecasting of sales by using fusion of machine learning techniques. 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), pp. 93-101

Perone, G. 2022. Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. Eur J Health Econ 23, 917–940. https://doi.org/10.1007/s10198-021-01347-4

R. Chuentawat and Y. Kan-ngan. 2018. The Comparison of PM2.5 forecasting methods in the form of multivariate and univariate time series based on Support Vector Machine and Genetic Algorithm. 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 572-575, doi: 10.1109/ECTICon.2018.8619867.

Soto-Ferrari, Milton & Chams-Anturi, Odette & Escorcia Caballero, Juan Pablo. 2020. A Time-Series Forecasting Performance Comparison for Neural Networks with State Space and ARIMA Models.

Sun, Zhanao. 2020. Comparison of Trend Forecast Using ARIMA and ETS Models for S&P500 Close Price. 57-60. 10.1145/3436209.3436894.

WeiGuo Ma. 2020. Prediction and analysis of population aging based on computer Leslie model. MSIEID. DOI: 10.1109/MSIEID52046.2020.00027

Wei Y, Wang Z, Wang H, Li Y, Jiang Z. 2019. Predicting population age structures of China, India, and Vietnam by 2030 based on compositional data. PLoS ONE 14(4): e0212772. https://doi.org/10.1371/journal.pone.0212772

WHO. 2022. Ageing and health. Geneva: World Health Organization.

Widayani, Heni. 2020. Pyramid Population Prediction using Age Structure Model. CAUCHY. 6. 66. 10.18860/ca.v6i2.8859.

X. Zheng, J. Cai and G. Zhang. 2022. Stock Trend Prediction Based on ARIMA-LightGBM Hybrid Model. 2022 3rd Information Communication Technologies Conference (ICTC), 2022, pp. 227-231, doi: 10.1109/ICTC55111.2022.9778304.

Y. Wang and Y. Guo. 2020. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Communications, vol. 17, no. 3, pp. 205-221, doi: 10.23919/JCC.2020.03.017.

LOW YEE SEN (A175741)
Dr Azizi Abdullah
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia