PLATFORM - A Journal of Science & Technology

Received: 20 September 2021, Accepted: 25 December 2021, Published: 31 May 2022, Publisher: UTP Press, Creative Commons: CC BY 4.0

# DOMESTIC WATER CONSUMPTION FORECASTING WITH SOCIODEMOGRAPHIC FEATURES USING ARIMA AND ARIMAX: A CASE STUDY IN MALAYSIA

**Syahidah Mohd Khairi*, Izzatdin Abdul Aziz**

Department of Computer and Information Sciences, Universiti Teknologi PETRONAS

*Email: syahidah_20001705@utp.edu.my

## ABSTRACT

*Domestic water consumption in Malaysia exceeds the World Health Organisation's (WHO) recommendation and the target by National Water Services Commission (SPAN). This could cause problems to the future water supply. Domestic water consumption forecasting for each state proposed in this paper could be used to formulate targeted strategies for water management and conservation awareness. Time series annual data between 2010 to 2019 obtained from Malaysian government portals were used to forecast domestic water consumption based on its past values and socio-demographic data. Variance Inflation Factor was used to remove collinear variables, while non-collinear features were selected based on high Pearson and Spearman correlations and XGBoost and AdaBoost permutation-based feature importance. It was found that percentage of the population above 64 years old, gender ratio, and mean household size were the important features for all states. These features were included in fitting ARIMAX models in forecasting domestic water consumption one year ahead for each state. However, time series forecasting modelling results show that ARIMA models using only past data for domestic water consumption for each state have better accuracy than ARIMAX models, except for Kedah and Perak. The need to consider important features differently for each state, linearity assumption of exogenous variables and insufficient data could be the reasons for lower $R^2$ values and higher RRMSE and MAPE for ARIMAX models. Thus, the forecasts for the year 2020 were obtained using ARIMAX models for the two states and ARIMA models for others. The forecasted values show that only three states would meet the target by SPAN, while others were expected to exceed it. This shows that more efforts are needed to increase water conservation awareness, focusing on the states with high forecasted domestic water consumption.*

***Keywords:*** *ARIMA, ARIMAX, correlation, feature importance, Variance Inflation Factor (VIF)*

## INTRODUCTION

Clean water and sanitation are among the issues addressed in United Nations Sustainable Development Goals. One of the sub-goals in the area is to have efficient use of water in different sectors to ensure water availability for most people by 2030 [1]. In Malaysia, access to water is available to most people, with 96.7% having access to water supply in 2018 [2]. Although the country has abundant water, there is no exception to climate change's impact on Malaysia, such as the acts of droughts on dams [3] and agriculture [4]. Groundwater is also expected to decrease due to changes in land usage and the development of new areas [5]. Floods will also reduce the quality of water sources [4], increasing the cost of water treatment. However, there is a lack of awareness of water conservation in Malaysia. In 2018, domestic water consumption was 226 litres per capita per day (l/c/d) in Peninsular Malaysia and Labuan [2], which is considered high compared to the World Health Organisation (WHO) recommendation of 165 l/c/d, while the National Water Services Commission (SPAN) targets for 180 l/c/d by 2020 [2],[6]. Thus, there is

an urgency to reduce consumption to avoid water insufficiency.

The awareness of water consumption problems could be tracked down to socio-demographic factors. Several studies have shown that younger generations [8]-[9] and bigger household sizes were found to consume lesser water per capita [9], although there are conflicting findings on the effects of income and family size [8]. A study in a southern state of Peninsular Malaysia also shows that people in the middle-income class have less effort to consume water wisely [10]. In a Malaysian university, females were also consuming significantly more water than males [11]. Foreign citizens' percentage was also found to be higher with increasing water consumption, probably because the location of the study was usually dry and the local citizens were more aware of the limitations of water [9]. These socio-demographic factors may be used to develop better water management policies and conduct awareness campaigns [8] through personalised recommendations and data-driven actions. However, it is also important to determine which factors affect water consumption significantly.

Several studies also show that time series models can be used to forecast water consumption. Duerr et al. used the Autoregressive model (AR) and Autoregressive Integrated Moving Average model (ARIMA), together with other machine learning models, including Gradient Boosting Machine (GBM), Random Forest and Bayesian Additive Regression Trees (BART), to forecast monthly household water consumption in Tampa Bay city area. The AR model has the best performance in the study compared to other methods [12]. Statistical time series models require lesser computational resources yet have better accuracy for one-step-ahead forecasting compared to machine learning and deep learning approaches [13]. On the other hand, ARIMAX models were also used by including "dew point depression and average temperature" [14], and in analogue-based weather forecasts [15] to forecast water consumption in Las Vegas and Tampa Bay city areas, respectively. However, these studies focused on cities, and different areas may have different factors affecting water consumption. The model with different features was also selected based on errors after model fitting, which could have many possible combinations to be included when multiple features are considered. In this paper, the feature selection method was used prior to model fitting to determine the importance of specific socio-demographic features to forecast average annual water consumption in different states in ARIMA and ARIMAX models, and the selected features were included as exogenous variables since it was shown previously that they affect water consumption in Malaysia.

This paper aims to identify the relationship between socio-demographic features and domestic water consumption using correlation values and feature importance scores. Using the selected features, we developed time series models for next-year domestic water consumption forecasting using available datasets and validated the performance of time series models.

## MATERIALS AND METHODOLOGY

Figure 1 shows the workflow, from obtaining data to forecasting the values for domestic water consumption, which will be explained further in this section.

### Data Cleaning

The datasets for water consumption forecasting were obtained from Malaysia Informative Data Centre (MysIDC) [16] and data.gov.my [17] portals, summarised in Table 1. All data were sourced from the Department of Statistics Malaysia, except metered domestic water consumption by state, sourced from SPAN. MysIDC provides time-series structured datasets for all datasets, except Dataset 8, obtained through the data.gov.my portal with a similar format. There are 14 states included in this paper, with data for Kuala Lumpur and Putrajaya combined with Selangor.

The time range that covers all datasets is from the year 2010 to 2019, which will be used for forecasting the domestic water consumption for the year 2020 (one-time step). It is also important to note that mean monthly household gross income and Gini coefficient data are not collected annually. Since replacing them with mean or median values may affect the trend, imputations would be needed. A study by Wubetie [18] found that linear spline interpolation was the method

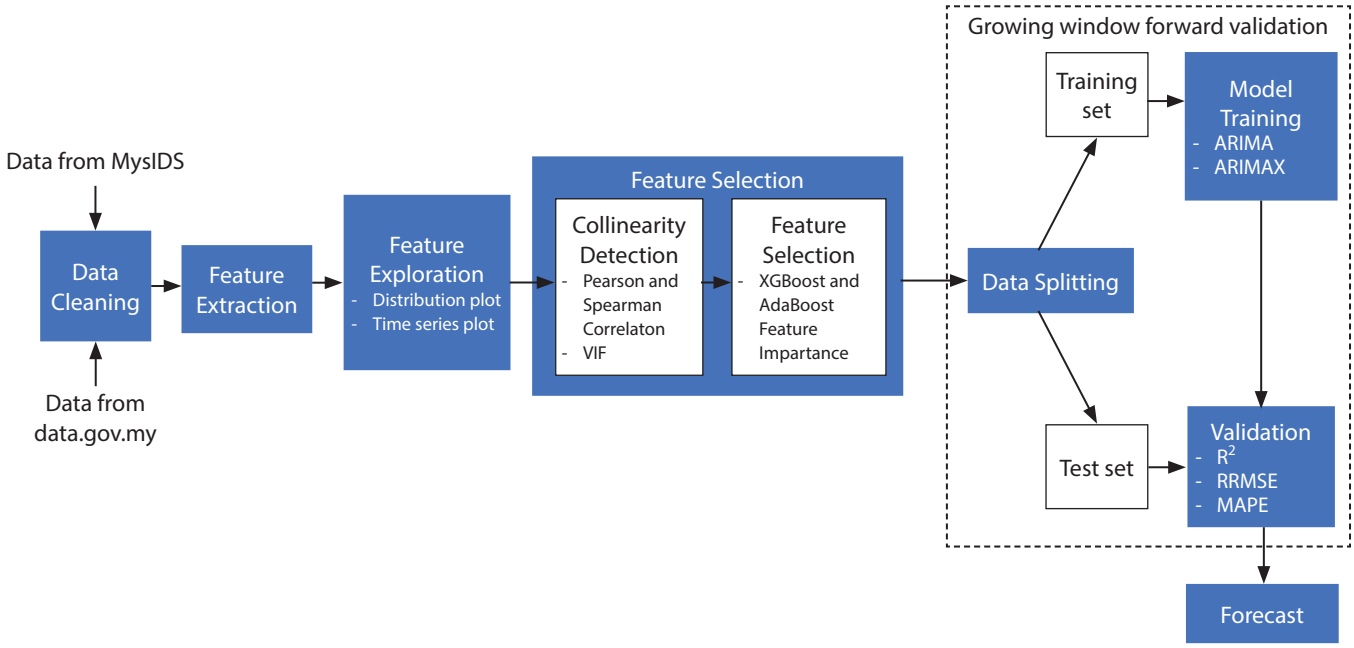


**Figure 1** Workflow summary for the time-series forecasting study

**Table 1** Available datasets for the study

| No | Dataset Title | Description | Time range (year) |
|---|---|---|---|
| 1 | Metered domestic water consumption by state | Metered domestic water consumption in a million litres per day for each state annually | 2003-2019 |
| 2 | Population with treated piped water by strata by state | Percentage of population with treated piped water in urban, rural and overall average for each state annually | 2000-2019 |
| 3 | Principal statistics of the labour force by State 1982–2019 | Number of persons who are employed, unemployed, labour force, and outside the labour force, percentage of labour force participation rate and unemployment rate annually in each state | 1982-2019, except W.P. Labuan (1984-2019) and W.P. Putrajaya (2011-2019) |
| 4 | Mean Monthly Household Gross Income by state, 2002 - 2019, Malaysia | Mean monthly household gross income for each state annually | 2002, 2004, 2007, 2009, 2012, 2014, 2016 and 2019 |
| 5 | Gini coefficient by state, 2002 - 2019, Malaysia | Gini coefficient for each state annually | 2002, 2004, 2007, 2009, 2012, 2014, 2016 and 2019 |
| 6 | Number of households by state, Malaysia, 2010 - 2020 | Number of households for each state annually | 2010-2020 |
| 7 | Population by age, sex and ethnic group, [State], [Year] | Population by age, group, sex and ethnic group annually, with a dataset for each state | 1980-2020 (W.P. Kuala Lumpur), 1991-2020 (W.P. Labuan), 2010-2020 (W.P. Putrajaya), 1970-2020 (other states) |
| 8 | Population by Age Group, Sex and Ethnic Group, Selangor, 1970-2020 | Population by age, group, sex and ethnic group annually in Selangor | 1970-2020 |

with the lowest error for non-random sequential missing socio-economic data, compared to linear regression, possibly because of the non-linear trend. Thus, it will be used to impute the missing values in the sixth and seventh datasets, using Equation 1, with as the index for the upper neighbour and as the index for the missing values [18].

$$Y_{t_k} = Y_{t_{k-1}} + \frac{t_k - t_{k-1}}{t_n - t_{k-1}} (Y_{t_n} - Y_{t_{k-1}}), \quad 2 < k < n-1 \qquad (1)$$

## Feature Extraction and Exploration

The cleaned data were transformed into features to be used in the time series models. Metered domestic water consumption data provided by SPAN is million litres per day (MLD). However, domestic water consumption per capita per day (DWC) was used in this paper to compare with the SPAN target and WHO recommendation [2],[6]. This value can be calculated by dividing metered domestic water consumption by the population with treated piped water, assuming all treated piped water is metered.

For socio-demographic data, the age groups were recombined into four: below 15 years old, 15-34 years old, 35-64 years old, and above 64 years old, following the recommendation in [9]. For the four age groups (P15, P1534, P3564, P64) and non-Malaysian or Foreign Citizen (FR) features, the units were in percentage over the total population. On the other hand, Gender (GR) would be represented by the ratio of male-to-female population as suggested in [19]. Unemployment Rate (UR) and Labour Force Participation rate (LFP), were also calculated based on the percentage of the respective number of persons over the total of the respective and opposite (employed or outside of labour force) data. The last calculated variable is the mean Household Size (HS), the total population over the number of households (number of persons per household). There were no transformations for mean Monthly Household Gross Income (MHGI), in Ringgit Malaysia, and Gini Coefficient (GC).

After the features were extracted, distribution plots and time series plots were developed for each feature and forecast variable to examine the normality and trends of the data.

## Feature Selection

Although there are 11 features extracted from the socio-demographic data, only relevant features should be used to forecast DWC.

### *Scaling*

Scaling features may reduce possible errors in the analysis due to different ranges across variables, especially for regression-based methods that involve the estimation of parameters based on data. Although different scaling techniques for regression were found to have minor differences [20], another study shows that XGBoost performed the best when RobustScaler was used [21]. RobustScaler is robust to outliers since it uses interquartile ranges instead of absolute minimum or maximum values in the calculation [22]-[23]. On the other hand, standardisation assumes a normal distribution of data [24], but RobustScaler does not. Thus, RobustScaler is suitable to be used to scale features while retaining non-stationarity and non-normality, calculated using Equation 2 [23].

$$x_{scaled} = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)} \qquad (2)$$

where $x_{scaled}$ is the scaled value, is the original value $Q_1(x)$ and $Q_3(x)$ are the first and third interquartile ranges.

### *Collinearity Detection*

Collinear features could reduce their feature importance values [25]. To identify the collinearity between the features, correlation matrices were created for all features and DWC. Then,

1. The pairs of features with both correlation values higher than 0.7 (strong correlation) [27] were identified.
2. One of the two features with lower correlations with DWC was removed for each pairas conducted in [9].
3. VIF was calculated for all remaining features.
4. If VIF values greater than 10, the feature with the highest VIF value was identified. Any correlation value of that feature with others greater than 0.7 was noted.
5. After that, steps 2, 3 and 4 were repeated until all VIF values were lower than 10 to reduce multicollinearity [26].

The remaining features were then compared using feature importance scores.

***Feature Importance***

Feature importance of using AdaBoost and XGBoost algorithms were calculated for the remaining features. XGBoost, or eXtreme Gradient Boosting, is a type of a system using tree boosting that specialises in scalability using lesser resources [28]. XGBoost feature importance was used as one of the selection measures suggested in [29] to determine important locations to forecast solar irradiance. Additionally, Nuţă et al. also use AdaBoost feature importance, which uses decision trees as weak learners with a random subset of features in each tree to determine the important factors for carbon emissions [30]. In this paper, to select important features for domestic water consumption, Python scikit-learn's permutation-based feature importance was used for the calculation since it is less biased towards features with higher cardinality or more possible values [25].

Since the permutation importance function depends on the model trained, cross-validation was used to find the hyperparameters using Python xgboost's xgboost.cv() and scikit-learn's GridSearchCV() for both XGBoost and AdaBoost respectively by minimising Mean Absolute Error (MAE). The permutation feature importance scores were also compared with Pearson and Spearman correlations of the features with DWC to gain insights into the relationship between features and forecast variables. Features with consistently high scores were selected.

***Model Training and Validation***

For small data size as used in this paper (10 years of annual data), statistical models were found to have better prediction scores than neural network and random forest machine learning methods [31]. Two-time series statistical models, ARIMA and ARIMAX models are chosen in this paper based on their usage in multiple kinds of literature for time series forecasting [12], [14]-[15].

Hyndman et al. [32] explain the Autoregressive Integrated Moving Average (ARIMA) model as the combination of the autoregression (AR) model and Moving Average (MA) model with differencing. The AR model uses past values of the response variable to forecast the next value using multiple linear regression, which requires stationary data. However, non-stationary data can be transformed or differenced to keep the mean of the time series stable. On the other hand, the MA model uses past forecast errors. ARIMA combines the AR(p) model, differencing and MA(q) model, resulting in Equation 3.

$$y'_t = c + \phi_1 y'_{t-1} + \ldots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_1 \varepsilon_{t-q} + \varepsilon_t \quad (3)$$

where $y'_t$ is the differenced time series data with differencing order $d$ at time $t$, $c$ is a constant and $\varepsilon_t$ is the white noise at time $t$. For the given $p$, $q$ and $d$, the parameters $c$, $\phi_1$, ..., $\phi_p$, $\theta_1$, ... $\theta_q$ are obtained through maximum likelihood estimation, which is similar to minimising the sum of squares of $\varepsilon_t$.

Meanwhile, ARIMA with exogenous variables, or ARIMAX, incorporates the effects of exogenous variables into the ARIMA model. According to Hyndman et al. [32], the ARIMAX model can be represented using the following Equation 4.

$$y_t = \beta_0 + \beta_1 x_{1,t} + \ldots + \beta_k x_{k,t} + \eta_t \quad (4)$$

where $y_t$ is the forecast value at time $t$, $x_{k,t}$ represents the $k^{th}$ exogenous variable measured at time $\beta_k$, is the $k^{th}$ coefficient, and $\eta_t$ follows ARIMA process with orders $p$, $d$ and $q$. $x_{k,t}$ can also represent lagged exogenous variables.

The forward validation technique with a growing training window was used since it has the best performance [31] and is suitable for the models [33]. The test set at time was forecasted using the model trained based on the data from the first data to time $t - 1$ [32]. Each dataset for each state was split into training and test sets with a 70-30 train-test ratio, which is optimal for time series forecasting [34].

***Model Order Selection***

The models were trained separately for each state, so there are 14 models for each technique (ARIMA and ARIMAX), with possibly different orders that minimise the errors of the test set. Among different validation measures of $R^2$, RRMSE and MAPE, which can compare different states' values in percentages or ratios, RRMSE was selected to ensure forecasting with minimum

errors. RRMSE, compared to MAPE, penalises large deviations of forecasted from actual values because the errors are squared. Minimising RRMSE could reduce the probability of deviating far away from the actual values, which is important in this paper.

The order ranges that were explored in this paper were between 0 to 4 for $p$ and $q$, and between 0 to 2 for $d$. Moreover, to ensure the forecasting ability for the next step of this data, the models were also fitted to the whole data for each state and order. For ARIMA models, only time series data of DWC were used. On the other hand, ARIMAX models require additional exogenous variables. However, since future data are not available, lagged data would be used as the exogenous data as suggested by Hyndman et al. [32]. Thus, the model to predict DWC at time would consist of previous data of DWC and data of selected features at time $t - 1$. However, it is important to note that the size of the training set was reduced by 1, since the first data was not included due to the absence of the lagged selected features data for that time.

***Validation between ARIMA and ARIMAX models***

RRMSE, $R^2$ and MAPE were calculated for each selected model order for each state. These values were compared to identify the model, either ARIMA or ARIMAX, with the highest $R^2$ and lowest RRMSE and MAPE for each state. The chosen models were then used to forecast the DWC for the next time step.

***Forecast***

All the available DWC and selected features (for ARIMAX) data from 2010 to 2019 were used to fit the chosen model, and the values of DWC were forecasted for the next time step (in this case, the year 2020). The forecasted values for each state were compared with the targeted DWC by SPAN and WHO.

## RESULTS AND DISCUSSION

Using the methodology explained previously, only certain socio-demographic features were determined to be important. Then, ARIMA and ARIMAX models were fitted and compared based on the validation measures. Lastly, the forecasted values of domestic water consumption were compared with the targeted value to propose targeted strategies.

***Feature Exploration***

Distribution plots for all features and forecast variables for all states and years were constructed, as shown in Figure 2. From the distribution plots, most variables are approximately normal or bimodal. However, FR data are very skewed to the right. Since there are multimodal and skewed distributions, normality assumptions may not be valid for these data. Moreover, scaling is needed since the range of values differ across variables.

Time series plots were also constructed for each feature and forecast variable, as shown in Figure 3. Generally, the difference in values across states are similar, except for FR and P1534 which Sabah has a higher values compared to other states. This might explain the presence of very high values in the two variables, as shown in Figure 2. Moreover, several variables may not be assumed stationary based on the plots since there are trends with time.

Based on these plots, data might not be assumed to be normal and stationary. Thus, ARIMA and ARIMAX models would be suitable for time series forecasting since differencing is necessary to be included for non-stationary data.

***Feature Selection***

The socio-demographic features used to forecast DWC were selected based on non-collinearity and feature importance scores. Only the selected features were used in fitting ARIMAX models.

***Collinearity Detection***

Pearson and Spearman correlation matrices for all the features and DWC are shown in Figure 4. Generally, the values for both correlations are similar. However, some pairs have differences in the correlations greater than 0.2. Some pairs have a much lower Pearson correlation (FR and MHGI, FR and LFP, and P1534 and MHGI), which could be because of the non-linear trend, although monotonic. There are also pairs with much higher Pearson correlation (P1534 with DWC, HS, UR, FR and P64, and FR with DWC and UR). The outliers or non-normality may result in a high Pearson correlation because a few points with extreme values could increase the correlation values [35].

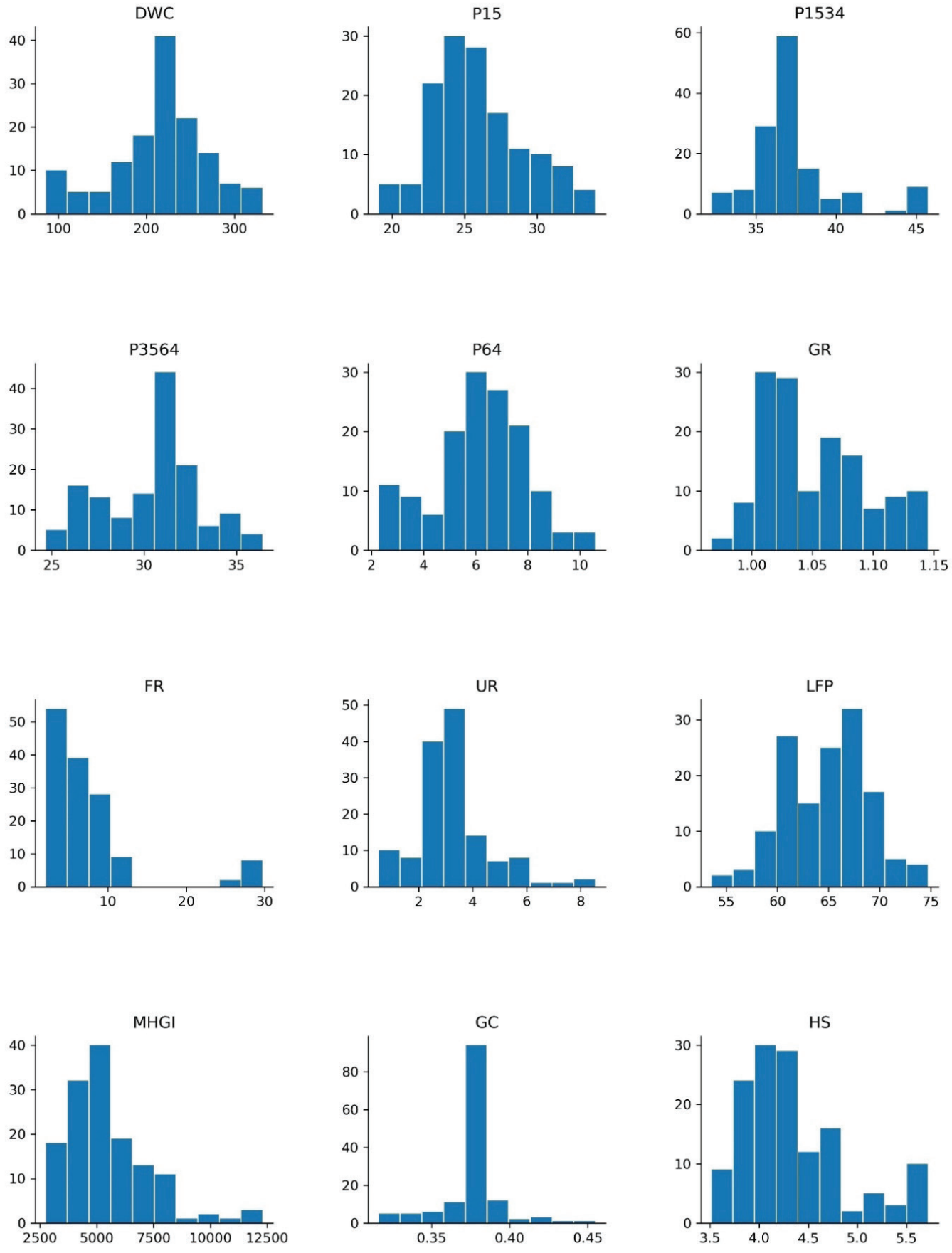


**Figure 2** Distribution plot of variables

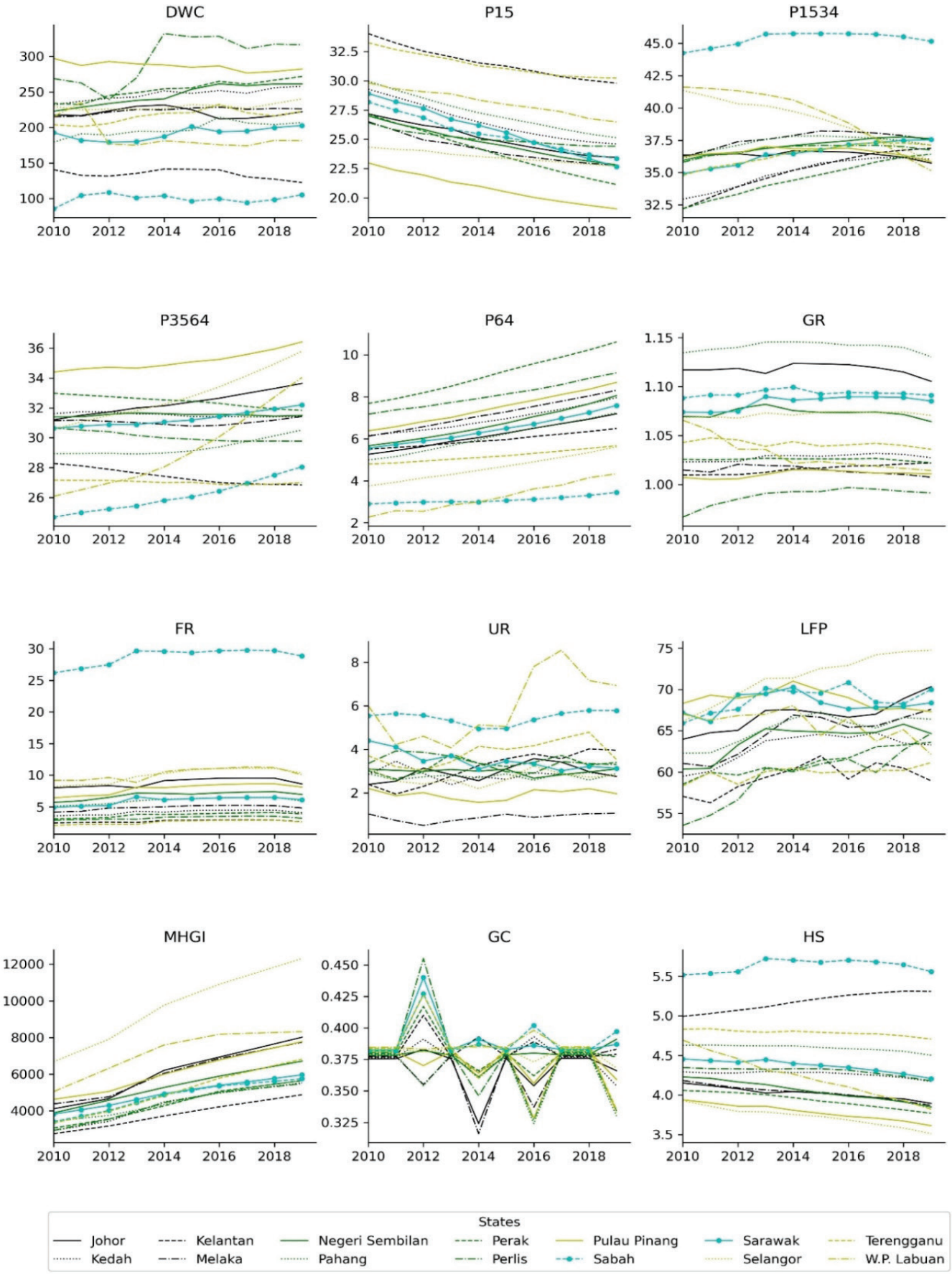


**Figure 3** Time series plots of variables

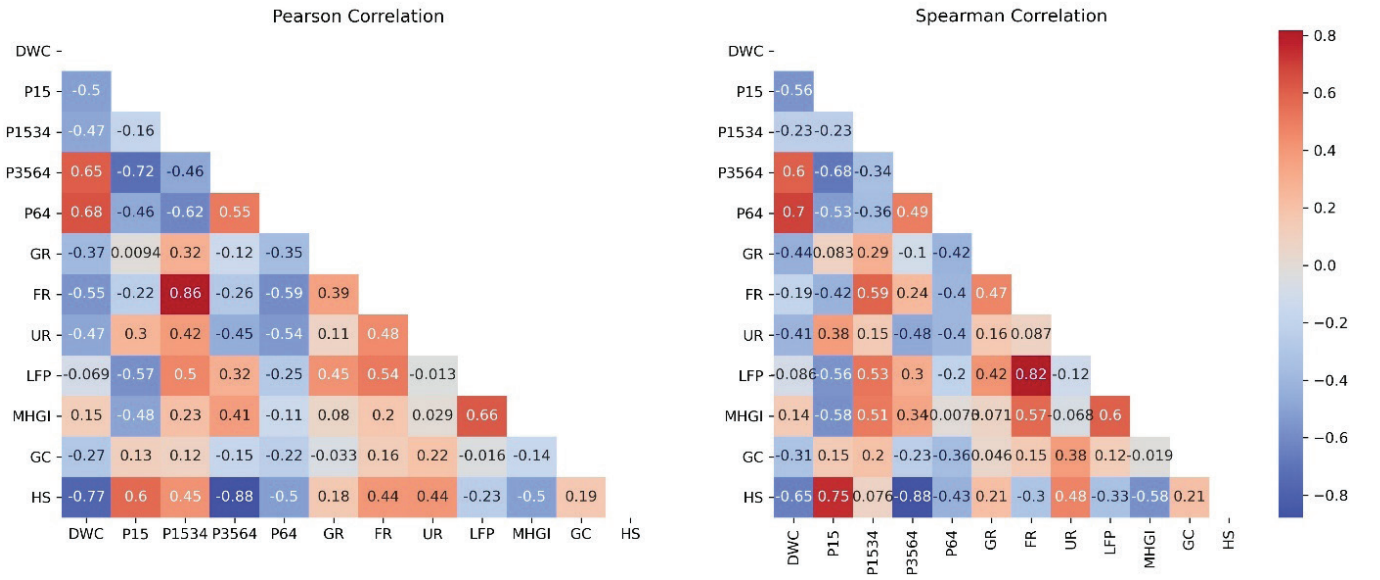


**Figure 4** Pearson and Spearman correlation matrices

One pair of features was found to have both absolute correlation values greater than 0.7, which are P3564 and HS. This shows that a higher percentage of the population between 35 to 64 years old correlates with a smaller mean household size, which could be due to their preference to live independently with their own income, while younger ones move out to study and work at other places. Since HS and P3564 are collinear, one of them needs to be removed. P3564 has a lower correlation value with DWC compared to HS with DWC so that HS could be more important than P3564, and P3564 was removed as one of the features.

VIF values were then calculated for all other features, as summarised in Table 2. It was found that P15 and HS have VIF values greater than 10, showing the presence of possible multicollinearity. P15, with the highest VIF, was found to strongly correlate with HS based on Spearman correlation (0.75), and P3564 based on Pearson correlation (–0.72). However, since P3564 was removed in the previous step, only HS and P15 would be considered. Both Pearson and Spearman correlations of P15 with DWC are lower than HS with DWC, so P15 was removed as one of the features since HS could be more important than P15. VIF for all the remaining features were recalculated and all the values were lower than 10, as shown in Table 2. Thus, these features could be compared and selected for forecasting DWC.

**Table 2** VIF calculation results

| Features | VIF | VIF after removing P15 |
|---|---|---|
| P15 | 18.4302 | - |
| P1534 | 7.21619 | 4.91367 |
| P64 | 9.64312 | 2.19134 |
| GR | 1.76574 | 1.49477 |
| FR | 9.9695 | 5.52107 |
| UR | 1.98499 | 1.94057 |
| LFP | 5.09649 | 3.4526 |
| MHGI | 2.96553 | 2.60755 |
| GC | 1.16402 | 1.16265 |
| HS | 10.8492 | 3.4334 |

***Feature Importance***

XGBoost and AdaBoost feature importance scores were calculated for the remaining nine features. The scores, and Pearson and Spearman correlations with DWC, were summarised in Table 3. The total feature importance scores and absolute values of correlations were also included for comparison.

P64, GR and HS are the features with scores higher than 1.0. The score is consistent with the feature importance scores higher than 0.1, including non-linear or non-monotonic relationships between features and DWC. Thus, the three features were selected as important ones. Only P64 has positive correlations among the selected features, while GR and HS have negative correlations. In other words, domestic water consumption per capita is generally higher with a larger

**Table 3** Feature importance and correlation values of features

| Features | AdaBoost | XGBoost | Pearson Correlation | Spearman Correlation | Total Absolute Score |
|---|---|---|---|---|---|
| P1534 | 0.018092 | 0.010856 | -0.47092 | -0.23431 | 0.734172 |
| P64 | 0.190831 | 0.081183 | 0.678385 | 0.697444 | 1.647843 |
| GR | 0.173872 | 0.191536 | -0.37408 | -0.43751 | 1.176996 |
| FR | 0.005644 | 0.006499 | -0.55498 | -0.18573 | 0.752852 |
| UR | 0.022808 | 0.067853 | -0.47242 | -0.41062 | 0.973702 |
| LFP | 0.004663 | 0.004536 | -0.06935 | -0.08579 | 0.164333 |
| MHGI | 0.004709 | 0.014474 | 0.148346 | 0.144195 | 0.311724 |
| GC | 0.010422 | 0.01224 | -0.27253 | -0.31268 | 0.607873 |
| HS | 0.601759 | 0.82251 | -0.77045 | -0.64779 | 2.84251 |

population above 64 years old and a lower gender ratio and household size, which is consistent with previous findings. Older generations may consume more water because of lower awareness or difficulties saving water in daily activities due to a lack of energy or ability to move [8].

On the other hand, smaller households may consume more water per capita since they may use similar amount of water to bigger households for daily activities such as dishwashing and laundry [9]. Domestic water consumption is also higher with a lower gender ratio, corresponding to more females than males. The finding is consistent with the finding that female students consume more water than male students [11]. This could be explained when females do household activities using water, such as watering plants and washing dishes and clothes more frequently and for a longer time than males.

***Time Series Forecasting Models***

ARIMA and ARIMAX models were then fitted and compared using the growing window forward validation technique. Figure 5 shows the plot of the actual and forecasted data. Overall, RRMSE and MAPE values are below 5% for all states except the ARIMAX models for Kelantan, Pahang, Perlis, Sabah, and Terengganu. For ARIMAX models with high errors, the predicted values were very far away from the actual values and most do not follow the general trends. On the other hand, many $R^2$ values for the test set are negative, except Johor (ARIMA), Sarawak (ARIMA), Perak (ARIMAX), Kedah (both) and Selangor (both), which have forecasts close to actual values and mostly following the monotonic trends.

The validation measures for ARIMA and ARIMAX models were compared. Figure 6 shows the comparison of validation measures for each state's model. For easier visualisation, the percentage score changes of ARIMA models compared to ARIMAX models for all validation measures were calculated using Equation 5, where negative values indicate higher values for ARIMA models and vice versa.

$$\text{Percent Score Change (\%)} = \frac{\text{ARIMA Score} - \text{ARIMAX Score}}{\text{ARIMAX Score}} \times 100 \qquad (5)$$

All states have consistent validation measures indications, either RRMSE and ARE greater and $R^2$ lower for ARIMA, or vice versa, except Sabah. ARIMA model for Sabah has lower RRMSE, and greater MAPE and $R^2$ than ARIMAX model. MAPE may show different model preferences because of absolute difference instead of squared difference. Thus, the errors for ARIMAX for Sabah might mostly come from large errors that are penalised in squared difference but also include a few minor errors. Nevertheless, since the per cent difference in MAPE is lower than RRMSE, the two measures indicate that the ARIMA model has better accuracy. The ARIMA model is determined as the model with better accuracy for Sabah.

Kedah and Perak showed better accuracy for ARIMAX models than ARIMA models. In these states, the inclusion of selected socio-demographic features increases the performance of the models compared to only using previous DWC data. This shows that the feature selection method benefits the models for these two states. On the other hand, models without


**Figure 5** Plot for train, test and ARIMA and ARIMAX forecast DWC data

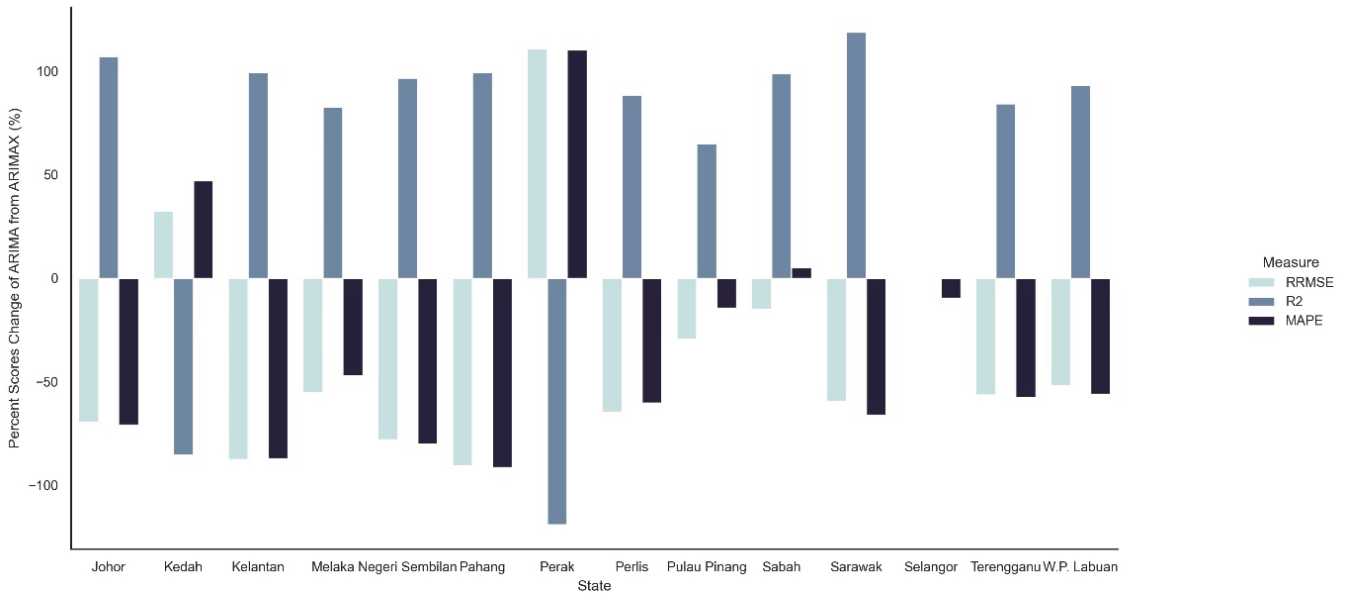


**Figure 6** Comparison between ARIMA and ARIMAX models for each state

the selected socio-demographic features show better accuracy for other states. Thus, the effects of socio-demographic features on the forecasting performance of DWC differ for each state. The assumption that selected features are important for all states may need to be revised since different states may be affected by different factors. Moreover, although the features were selected based on high scores for AdaBoost and XGBoost feature importance and the combination of both Pearson and Spearman correlations, the ARIMAX model assumes a linear relationship for exogenous variables, which is not necessarily true for important features selected by AdaBoost and XGBoost algorithm. ARIMAX models may also estimate more coefficients, where data insufficiency could affect the performance of the models. Based on the comparison, ARIMAX would be used to forecast DWC for Kedah and Perak, while ARIMA would be used for others.

***Forecasted Values***

Table 4 shows the forecasted values of DWC for 2020 along with the models and orders selected previously for each state and the change of DWC from 2019 to 2020. The average forecasted DWC for all states is about 219 l/c/d, which is above SPAN's target while also close to the value reported in 2018, which is 226 l/c/d [2]. This shows that the forecasted values are reasonable, and the overall DWC might decrease with time to indicate better awareness of saving water. However, the value varies with states, with the range of about 210 l/c/d. There are only three states with forecasted DWC lower than the target: Kelantan, Sabah, and WP Labuan. This is expected since the previous data also shows low values of DWC for the three states.

Moreover, Kelantan and Sabah have the lowest population with treated piped water compared to other states [16]. Although DWC data was extracted only based on the population with treated piped water, people in these states may have multiple water sources, causing possible inaccuracy in water consumption measurement. It is also possible to consume less water due to a limited piped water supply. The trends of DWC can also be observed across time. DWC for Sarawak, Perak, Johor and Kelantan are expected to increase from 2019 to 2020, while others decrease or approximately the same.

States using ARIMAX models also show the effects of socio-demographic features. Perak is forecasted to have higher domestic water consumption than Kedah, which can be linked to higher P64 and lower GR and HS, other than higher past values of DWC. Thus, the selected socio-demographic features are important and should be considered together in the targeted strategies for water management and awareness campaigns, especially for these states.

The results show that different states have different trends and forecasted DWC values. Therefore, there is

**Table 4** Forecasted DWC for each state for the chosen model and orders

| State | Model | Orders | Forecasted DWC (l/c/d) | DWC Change (l/c/d) |
|---|---|---|---|---|
| Johor | ARIMA | (0, 0, 4) | 223.980218 | +1.754376 |
| Kedah | ARIMAX | (1, 1, 2) | 246.8380 | -11.03060 |
| Kelantan | ARIMA | (4, 1, 0) | 128.7824 | +6.39593 |
| Melaka | ARIMA | (3, 2, 3) | 221.6141 | -4.5815 |
| Negeri Sembilan | ARIMA | (0, 1, 0) | 261.3162 | +0.000009 |
| Pahang | ARIMA | (3, 0, 3) | 201.9806 | -4.98264 |
| Perak | ARIMAX | (1, 1, 1) | 287.7448 | +16.01054 |
| Perlis | ARIMA | (1, 0, 0) | 309.8499 | -6.69345 |
| Pulau Pinang | ARIMA | (4, 1, 0) | 274.3821 | -8.01246 |
| Sabah | ARIMA | (4, 1, 1) | 98.47259 | -6.8537 |
| Sarawak | ARIMA | (0, 2, 1) | 204.6200 | +1.87984 |
| Selangor | ARIMA | (2, 1, 3) | 235.7343 | -4.48972 |
| Terengganu | ARIMA | (0, 0, 0) | 215.8891 | -6.75721 |
| W.P. Labuan | ARIMA | (4, 1, 0) | 178.1453 | -3.12358 |

a need to conduct targeted strategies and campaigns for each area. One of the strategies is to prioritise areas with very high and increasing forecasted DWC (such as Perak) for multiple water-saving efforts, such as installing IoT-based smart meters and giving feedback on water usage and its categories [8]. Higher budgets can be allocated to encourage the local community to realise the importance of water conservation and practice water-efficient daily activities based on feedback on their usage. The effectiveness of these efforts can be monitored by comparing the forecasted value and the actual data to be observed. Lower measured data may indicate that the targeted actions work effectively.

## CONCLUSION

This paper discusses the approaches to forecast DWC for each state in Malaysia using time series data of domestic water consumption and socio-demographic features. Three socio-demographic features were selected based on VIF, Pearson, and Spearman correlations, AdaBoost and XGBoost permutation feature importance scores for data from all states. The correlations show that DWC increases with increasing population percentage above 64 years old and decreasing gender ratio and mean household size. Moreover, the different techniques were consistent and complemented each other in detecting collinearity and ranking important features. However, this may be specific to this type of data since correlations are suitable for monotonic or linear data, while feature importance techniques also consider other relationships. It was also found that ARIMA models perform better than ARIMAX models with the selected for most states, except Kedah and Perak. This shows that the selected socio-demographic features may not necessarily increase the performance of the models, and the previous data of DWC would be more important for most cases, which is consistent with the finding by Duerr et al. [12]. However, this finding might be affected by limited data and the assumption of the same important features across states. It is worth to note that the percent difference in performance measure for most states are greater than the value of 50%, showing the great effects of the selected features. A future suggestion would be to conduct feature selection on each state data for each state forecasting model to better understand the effects of feature inclusion. In this paper, the forecasts for each state were obtained based on their most accurate models. Since most states exceed the target by SPAN, it should be highlighted with the need for more efforts while focusing more on the states with increasing and high DWC.

## REFERENCES

[1] "Progress on household drinking water, sanitation and hygiene 2000-2020: Five years into the SDGs.", *World Health Organization (WHO) and the United Nations*

*Children's Fund (UNICEF),* 2021. https://www.who.int/publications/i/item/9789240030848

[2] "Peninsular Malaysia & FT Labuan: Water And Sewerage Statistics 2019", *Suruhanjaya Perkhidmatan Air Negara*, 2019.

[3] C. Payus, L.A. Huey, F. Adnan, A.B. Rimba, G. Mohan, S.K. Chapagain, G. Roder, A. Gasparatos & K. Fukushi, "Impact of Extreme Drought Climate on Water Security in North Borneo: Case Study of Sabah", *Water,* 12, 4, pp. 1135, 2020.

[4] Z. Anang, J. Padli, N.K. Abdul Rashid, R. Mat Alipiah & H. Musa, "Factors Affecting Water Demand: Macro Evidence in Malaysia", *Jurnal Ekonomi Malaysia,* 53, 1, pp. 17-25, 2019.

[5] G.C. Mridha, M.M. Hossain, M.S. Uddin & M.S. Masud, "Study on availability of groundwater resources in Selangor state of Malaysia for an efficient planning and management of water resources", *Journal of Water and Climate Change,* 11, 4, pp. 1050–1066, 2020.

[6] "Strategies to Enhance Water Demand Management in Malaysia", *Academy of Sciences Malaysia, Academy of Sciences Malaysia*, Kuala Lumpur, 2016.

[7] "Innovation & Technology," *Air Selangor*, [Online]. Available: https://www.airselangor.com/about-us/innovation-technology/. [Accessed 26 July 2021].

[8] M.S. Rahim, K.A. Nguyen, R.A. Stewart, D. Giurco & M. Blumenstein, "Machine Learning and Data Analytic Techniques in Digital Water Metering: A Review," *Water,* 12, 1, pp. 294, 2020.

[9] M.C.Villarín, "Methodology based on fine spatial scale and preliminary clustering to improve multivariate linear regression analysis of domestic water consumption", *Applied Geography,* 103, pp. 22-39, 2019.

[10] N.S. Muhammad, J. Abdullah, N. Abd Rahman & N.A. Razali, "Water usage behaviour: Case study in a southern state in Peninsular Malaysia", *IOP Conference Series: Earth and Environmental Science*, 2021. DOI:10.1088/1755-1315/646/1/012017

[11] F. Mohd Daud & S. Abdullah, "Water Consumption Trend Among Students in a University's Residental Hall", *Malaysian Journal of Social Sciences and Humanities (MJSSH),* 5, 9, pp. 56-62, 2020.

[12] I. Duerr, H. R. Merrill, C. Wang, R. Bai, M. Boyer, M.D. Dukes & N. Bliznyuk, "Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study", *Environmental Modelling & Software,* 102, pp. 29-38, 2018.

[13] M.Spyros, S. Evangelos & A. Vasssilios, "Statistical and Machine Learning forecasting methods: Concerns and ways forward", *PLOS ONE,* 13, 3, pp. e0194889, 2018. Doi:10.1371/journal.pone.0194889

[14] P. Huntra & T. C. Keener, "Evaluating the Impact of Meteorological Factors on Water Demand in the Las Vegas Valley Using Time-Series Analysis: 1990–2014", *ISPRS International Journal of Geo-Information,* 6, 8, pp. 249, 2017.

[15] D. Tian, A. Christopher J. Martinez & T. Asefa, "Improving Short-Term Urban Water Demand Forecasts with Reforecast Analog Ensembles", *Journal of Water Resources Planning and Management,* 142, 6, pp. 04016008, 2016.

[16] "MysIDC," *Malaysia Informative Data Center*, [Online]. Available: https://mysidc.statistics.gov.my/. [Accessed 23 May 2021].

[17] "Population by age, sex and ethnic group, Selangor," *data.gov.my*, [Online]. Available: https://www.data.gov.my/data/ms_MY/dataset/population-by-age-sex-and-ethnic-group-selangor. [Accessed 31 May 2021].

[18] H.T. Wubetie, "Missing data management and statistical measurement of socio-economic status: application of big data," *Journal of Big Data,* 4, 1, pp. 47, 2017.

[19] L. Fan, L. Gaib, Y. Tonga & R. Lia, "Urban water consumption and its influencing factors in China: Evidence from 286 cities," *Journal of Cleaner Production,* 166, pp. 124-133, 2017.

[20] Z. Shang, E. Zgraggen, P. Eichmann & T. Kraska, "Niseko: a Large-Scale Meta-Learning Dataset," *3rd Workshop on Meta-Learning at NeurIPS 2019, Vancouver, Canada,* 2019.

[21] Q. Gu, A. Kumar, S. Bray, A. Creason, A. Khanteymoori, V. Jalili, B. Grüning & J. Goecks, "Accessible, Reproducible, and Scalable Machine Learning for Biomedicine," *bioRxiv,* 2020. Doi:10.1101/2020.06.25.172445

[22] "sklearn.preprocessing.RobustScaler," *scikit-learn,* [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html [Accessed 15 June 2021].

[23] D.K. Thara, B.G PremaSudha & X. Fan, "Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques", *Pattern Recognition Letters,* 128, pp. 544-550, 2019.

[24] "6.3. Preprocessing data", *scikit-learn*. [Online]. Available: https://scikit-learn.org/stable/modules/preprocessing.html [Accessed 15 June 2021].

[25] "4.2. Permutation feature importance", *scikit-learn*, [Online]. Available: https://scikit-learn.org/stable/modules/permutation_importance.html [Accessed 18 June 2021].

[26] C. Robinson & R. E. Schumacker, "Interaction effects: centering, variance inflation factor, and interpretation issues," *Multiple linear regression viewpoints,* vol. 35, no. 1, pp. 6-11, 2009.

[27] H. Akoglu, "User's guide to correlation coefficients", *Turkish Journal of Emergency Medicine,* 18, 3, pp. 91-93, 2018.

[28] T. Chen & C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, *New York*, 2016.

[29] B. Brahma & R. Wadhvani, "Solar Irradiance Forecasting Based on Deep Learning Methodologies and Multi-Site Data," *Symmetry,* 12, 11, pp. 1830, 2020.

[30] F. M. Nuţă, A. C. Nuţă, C. G. Zamfir, S.-M. Petrea, D. Munteanu & D. S. Cristea, "National Carbon Accounting—Analysing the Impact of Urbanisation and Energy-Related Factors upon CO2 Emissions in Central–Eastern European Countries by Using Machine Learning Algorithms and Panel Data Analysis," *Energies,* 14, 10, pp. 2775, 2021.

[31] M. Schnaubelt, "A comparison of machine learning model validation schemes for non-stationary time series data", *FAU Discussion Papers in Economics,* 11, 2019.

[32] R. Hyndman & G. Athanasopoulos, Forecasting: principles and practice, 3rd ed., *Melbourne: OTexts*, 2021.

[33] S. Siami-Namini, N. Tavakoli & A. S. Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series", *2018 17th IEEE International Conference on Machine Learning and Applications*, 2018.

[34] A. Ç. O. K. T. D. a. Ç. A. I. Yenidoğan, "Bitcoin forecasting using ARIMA & prophet," *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 2018.

[35] J. C. de Winter, S. D. Gosling & J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data.", *Psychological methods,* 21, 3, pp. 273, 2016.

[36] J. Frost, "How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis", *Statistics by Jim*, [Online]. Available: https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression/. [Accessed 9 July 2021].