Trip generation and attraction model and forecasting using machine learning methods

S M Saleh*, Lulusi, F Apriandy, J Fisiani, A Salmannur, R Faisal
Department of Civil Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia

*E-mail: sofyan.saleh@unsyiah.ac.id

Abstract. The necessity of people and goods to be moved from one place to another place has increased dramatically in recent years. It requires numbers of connectivity among the regions. Besides, policy changes in sea transportation sector including development of sea transport infrastructures as well as shipping/ferry lines to support fulfilling these needs are introduced. The ferry line from Kuala Langsa Port, Indonesia to Penang Port, Malaysia was introduced to encounter the need of mobility between those regions. In consequence, it is important to estimate future transport demand. This study is conducted to familiarize the use of machine learning methods in modelling and forecasting trip generation and trip attraction. Time-series trip generation and attraction data from Kuala Langsa to Penang and vice versa and socio-economic data were employed to develop the model. The result shows that gross domestic regional product (GDRP) and population variables has significant influence to generate trips between these ports.

1. Introduction
The expansion of low-cost air carrier has unlocked opportunities for ever-increasing people and goods movements across regions. With ASEAN-China Free Trade Agreement (ACFTA) in place, the South East Asia (SEA) region has seen border restriction loosen at a previously unseen level. However, since South East Asia Region has abundance of maritime wealth, there is a huge potential to develop seaborne transportation in the region.

United Nations Conference on Trade and Development (UNCTAD) states that approximately 90% of the global trade is attributed to maritime transport. Global ship density analysis by [1] shows that the sea-borne traffic in 2014 has quadrupled that of early 1990s, with the Indian Ocean and Western Pacific Seas experience the largest growth. The rise is related to maritime transport cost reduction, the decrease in the delivery times, and security of shipping [2]. Furthermore, as [3] states, in terms of CO2 emissions per cargo ton, maritime transport is environmentally more friendly. Larger cargo volume, both size and weight-wise, also plays a role.

Transportation in general is susceptible to policy change and technological developments. as a result of these, there is a need to estimate future transport demand. The forecast could be utilized to determine and analyse whether some adjustment, either engineering or management-wise, is needed. A number of factors widely used in transport demand estimation comprises Gross Domestic Product (GDP), population, household expenditure, fuel price, ticket fares, among others. These variables are deemed predominant and play significant roles in reflecting future transport demand. Furthermore, there is an intense correlation between GDP and transport demand, as it is considered to be the most suitable variable to demonstrate transport demand [4].

In this study, the features employed to estimate future trip generation and attraction are population and Gross Domestic Regional Product (GDRP). These variables are deemed exogenous, because they are external to the problem under study. GDRP varies from time to time due to goods and services prices fluctuation and volume changes [5]. The dataset is compiled from Langsa’s Centre for Statistics Board (BPS Langsa). Data prior to 2010 are not taken into consideration because, as in line with UN
recommendation, The GDP base year calculation has been changed from 2000 to 2010 taking into account such exceptional circumstances as ACFTA, international trade recording system, and capital markets services expansion. Based on the data used, it is considered an economics and time-series based method.

Transport demand modelling predominantly employs statistical approaches most notably regression, gravity model, and time-series method among others [6]. These approaches, which need to be mathematically proven, require the data to satisfy pre-existing assumptions beforehand. Unlike traditional method, machine learning technique exempt the data to follow any presumptive foundations and let the computers to learn from the observations. This offers flexibility for engineers to develop model [7].

Machine learning-based methods can be utilized for descriptive purpose by describing the system current status, predictive objective by predicting the system ensuing condition, or prescriptive aim by proposing system improvement [8]. Machine learning algorithms teaches computers to replicate the simplest way from which human learn: learning from experiences. Researches, such as [8-10], have shown that machine learning approaches often outperforms traditional modelling techniques. A number of most commonly used machine learning method in researches include regression, k-means clustering, Artificial Neural Network (ANN). The growing interest in the use of machine learning methods, in particular ANN, is attributed to its capability to deal with nonlinearities, discontinuities and polynomial aspects [9]. The output of ANN could be a nonlinear function of the variables which is a result of a linear integration of variables in the so-called hidden layers at the first place. It allows the model to learn and adapt by itself [11]. ANN is superior in its ability to train complex models, however a huge amount of data and iteration for training is often a prerequisite [12]. ANN also suits to model greatly varied data, missing data, or data that ignore cause and effect relationship [6].

2. Study area

The study was focused on the international Langsa-Penang sea route, as portrayed in Figure 1. The journey starts from Kuala Langsa Port, Langsa, Aceh, Indonesia, then docks in Penang Port, Penang, Malaysia, and vice versa. It was first opened in 2013, aiming to improve and strengthen bilateral cooperation in politics, economy, social, culture, investment, and tourism. It is the only direct transport mode connecting Langsa and Penang, and the only scheduled passenger ship linking Northern Sumatera Island and Penang Island. The 192-passenger seats Kenangan-3 ferry, owned by Fast Ferry Ventures Sdn Bhd, is the sole ship serving the route. The ship operates twice a week, departing from Penang at 8.30 am on Mondays and Wednesdays, and leaving Langsa every Tuesday and Thursday at 9 am. With a distance of around 300 km, a one-way voyage normally takes 6 to 7 hours, depending on sea and weather condition. A return trip costs Rp. 500,000.

3. Methodology

This study performs linear regression and ANN method to model trip generation and trip attraction given by GDPR and population as independent variables. Both methods are built using Python code. The algorithms are executed in Scientific Python Development Environment (Spyder). Dmatrices from patsy and statsmodels libraries are employed in linear regression method. Dmatrices is useful in separating independent and dependent variables and adding a constant value in regression analysis. Statsmodels has Ordinary Least Square (OLS) function that can model linear relationship between independent and dependent variables. It returns intercept, coefficient of each independent variable, and important statistical values such as R-squared and p-value. The 5% confidence level test is performed in this study.

ANN uses sklearn neural network’s MLP Regressor, sklearn model selection’s train test split. MLP Regressor train the data to fit a neural network model using Rectified Linear Unit (ReLU) as activation function for the hidden layer and the family of quasi-Newton methods-based optimizer called ‘lbfgs’ as the solver for weight optimizer. ReLU activation function returns the output value as f(x) = max (0, x). ReLU is commonly used to train numerical value-based neural network. Lbfgs explores the most suitable weights for each node in ANN by using iterative method: it trains the model by inserting random weights to initiate training, and then it updates the weights for the next iteration considering previous iteration results until the model converge. Lbfgs works well and converges fast for small dataset. Train test split module split the dataset into two subsets: training and test sets which is necessary
in ANN to validate the neural network model. The number of maximum iterations is set manually, but the system will return non-convergence warning should the number is deemed insufficient. Once the model converged, it will return the value of R-squared.

Figure 1. Study area

4. Results
The summary of the linear regression model is shown in Table 1 and Table 2. At first glance, the small number of each independent variable’s coefficient appears to indicate its insignificance. However, its p-value is small, suggesting that the independent variables, in this case GDRP and population, do have a strong connection to trip generation and trip attraction. Small coefficient does not necessarily mean meaningless. It might result from an imbalance dataset in the sense that population is a six-digit number, GDRP is a seven-digit number, whereas trip itself is a five-digit number. For every 10000 increase in population, 121 more trips will be generated, and an extra 33 trips will be attracted. On the other hand, one million rupiahs of GDRP growth is expected to result in 2700 and 700 added trip generation and trip attraction respectively. Both trip generation and trip attraction models produce an R-squared of almost 1.

After a few trials in ANN model, setting the maximum number of iterations to 20 does return the model converged. It can be seen from Figure 2 and 3 that the algorithm reaches its convergence at around 17 iterations in trip generation model and 14 iterations in trip attraction model. Fitting the models into the test subset, both models yield an R-squared values of 0.999, similar to those of the linear regression. ANN has no specific coefficient for each variable, because it assigns the weight in hidden layers. The only way to determine whether this ANN model perform well is through the value of R-squared. Hence for simple model, linear regression could be considered perform better because it specifies coefficient for each variable and its attributed statistical values such as p-values.
Table 1. Trip generation models summary

| Model                        | Independent variables | Coefficient | P-value | R-squared |
|------------------------------|-----------------------|-------------|---------|-----------|
| Linear regression            | Intercept             | -1304.88    | 0.038   | 0.999     |
|                              | Population            | 0.0121      | 0.043   |           |
|                              | GDRP                  | 0.0027      | 0.000   |           |
| Artificial Neural Network (ANN) | -                     | -           | -       | 0.999     |

Table 2. Trip attraction models summary

| Model                        | Independent variables | Coefficient | P-value | R-squared |
|------------------------------|-----------------------|-------------|---------|-----------|
| Linear regression            | Intercept             | -358.385    | 0.038   | 0.999     |
|                              | Population            | 0.0033      | 0.043   |           |
|                              | GDRP                  | 0.0007      | 0.000   |           |
| Artificial Neural Network (ANN) | -                     | -           | -       | 0.999     |

Figure 2. ANN R-squared values for each iteration in trip generation forecasting

Comparing both methods’ R-squared values, both linear regression and ANN models fit the data almost perfectly. From statistical perspective, it is considered the ideal model. However, it also hints a possible overfitting. It often happens for a small number of datasets.
Figure 3. ANN R-squared values for each iteration in trip attraction forecasting

5. Conclusion
In conclusion, the application of machine learning methods in modelling and forecasting trip generation and attraction is confirmed as a beneficial instrument to be used. It helps to analyse the accuracy of developed model by using artificial neutral network (ANN). Socioeconomic data (GDPR and population data) has high influence to determine number of trip generation and attraction between the ports. Given the R-squared resulted, both regression and ANN models perform well in forecasting trip generation and attraction. However, due to the limited number of datasets used in the modelling, there is an indication of a possible overfitting.

References
[1] Tournadre J 2014 Geophys. Res. Lett., 41 (22) 7924-7932.
[2] Cordón-Lagares E and Garcia-Ordaz F 2020 Res. Transp. Bus. Manag., 10 100520.
[3] Grant D B, Elliott M 2018 Ocean Coast. Manag., 163 162-172.
[4] Tsui W H, Fung M K 2016 J. Air Transp. Manag., 50 1-11.
[5] Langsa’s Centre for Statistics Board (BPS Langsa) 2020 Langsa Municipality in Figures 2020. Langsa: BPS Langsa.
[6] Profiliidis V A, Botzoris G N 2019 Modeling of Transp. Demand. Amsterdam: Elsevier.
[7] Zhao X, Yan X, Yu A, van Hentenryck P 2020 Travel Behaviour and Soc., 20 22-35.
[8] Barua L, Zou B, Zhou Y 2020 Res. Transp. Bus. Manag., 34 100453.
[9] Vijai P, Sivakumar P B 2018 Procedia Computer Sci., 143 258-266.
[10] Alekseev K P and Seixas J M 2009 J. Air Transp. Manag., 15 (5) 212-216.
[11] Liu Y, Zou B, Ni A, Gao L, Zhang C 2020 Transp. Lett., 1-13.
[12] Jiang J, Trundle P, Ren J 2010 Computerized Medical Imaging and Graphics., 34 (8) 617-631.