Vehicle’s Density Prediction Based on History Data of e-Toll in PT Jasamarga Pandaan Tol Using Hidden Markov Model

N I Asrori**, N Iriawan***, and W S Winahju***

1Department of Statistics, Faculty of Mathematics, Computing, and Data Science, Institut Teknologi Sepuluh Nopember, Indonesia

E-mail: *indahasrori@gmail.com, **nur_i@statistika.its.ac.id, and ***wiwiek@statistika.its.ac.id

Abstract. The Gempol-Pandaan toll is a strategic center for regional development. In 2017, the number of vehicles passing through this toll has increased. This condition could impact on traffic jam and road damage because of passes by many cars, especially trucks with loads that exceed their capacity. Therefore, it is necessary to do a vehicle’s density prediction not only from the Gempol toll gate but also from Kejapanan, Bangil and Rembang toll gates that will exit through Pandaan toll gate. In this research, vehicle's density is viewed from the probability and number of each vehicles category I, II, III, IV, V. The origin gate and vehicle category cannot be observed directly, but the car must pass the toll gate to do tapping the e-toll card, so the method used in this research is the Hidden Markov Model (HMM). Two estimate methods for HMM parameters are employed in this research, the Expectation-Maximization (EM) algorithm and Bayesian approach. The result shows that HMM parameters estimation using Bayesian is better than the EM algorithm. The Bayesian estimated parameter values are closer to the input parameters so that the model is more representative to explain the vehicle’s density prediction.

1. Introduction

Indonesia is dominated by land transportation. Total vehicles in 2017 were 138,556,669 units or 4% more than in 2016 [1]. This condition will impact on traffic jam. One of the government’s solutions to overcome this problem is the construction of toll roads. It has not been realized as decisions that solve problems properly because there are still many cars with overload capacity running at below the specified standard speeds i.e. 60-80 km/hour for tolls within the city, and 100 km/hour for inter-city toll so it will inhibit other vehicles [2]. The Gempol - Pandaan toll is part of the Surabaya-Malang toll construction plan which plays a strategic role for regional development, namely as a link for economic activity centers [3]. Based on information from Jasamarga, the volume of vehicles in the Gempol-Pandaan toll in 2017 increased by 10% more than in 2016. This condition will impact on traffic jam and road damage because of passes by many cars, especially overloaded capacity trucks. Therefore, the vehicle’s density is a problem that must be solved.

In this research, vehicle’s density is viewed from the probability and number of each vehicles category I, II, III, IV, V entering through Bangil, Gempol, Kejapanan, and Rembang toll gates. The origin of the toll gate and the vehicle category, however, cannot be observed directly.
Observable data are cars that have to pass through toll gates and tap their e-toll cards. From here the gate information of the origin and vehicle category can be identified. Therefore the proposed method to be applied in this study is the Hidden Markov Model (HMM). It is because the output depends on the observable state and the incident of vehicles arriving at the Pandaan toll gates following a Poisson process. Expectation-Maximization (EM) algorithm is usually employed to estimate HMM parameters. However, now it has developed, namely the use of the Bayesian method to estimate the HMM parameters [4]. Research on paired sherlock modeled as the Hidden Markov Model for disease interaction has also used this Bayesian approach [5].

The purpose of this study is to predict the vehicle’s density of Bangil, Gempol, Kejapanan, and Rembang toll gates to the Pandaan toll gate using the Hidden Markov Model, where the vehicle category data are recorded as a categorical time series. The hidden state used in this research is the origin toll gate and vehicle category, while the observable state is the Pandaan toll gate. The results of this study are expected to provide an overview of future vehicle densities that occur along toll roads from Bangil, Gempol, Kejapanan, and Rembang to the Pandaan exit toll gate, as well as providing recommendations for making toll road management policies for PT Jasamarga.

2. Literature Study
2.1. Goodness of Fit
The Goodness of Fit was a method to check the conformity between the distribution patterns of the observed data with the selected distribution pattern. Kolmogorov-Smirnov test was employed in this research to test the null hypothesis of data follows a particular distribution pattern versus the alternative hypothesis of data not follow a particular distribution pattern. Equation (1) is the test statistic for Kolmogorov-Smirnov:

\[ D_n = \max |F_n(x) - F_0(x)| \] (1)

where \( F_0(x) \) is the cumulative distribution function, and \( F_n(x) \) is the observed cumulative function of a random sample of \( N \) observations. The null hypothesis can be rejected when \( D_n \) is greater than the critical value of \( D_\alpha \) in Kolmogorov-Smirnov table at \((1 - \alpha)\) confidence level [6].

2.2. Mixture and Mixture of Mixture Models
A data has a mixed pattern if it has a combined pattern of several components, where each component has a different proportion [7] and [8]. As in Equation (2), mixture models can be expressed as the following notation:

\[ f(x|\theta, \delta) = \sum_{i=1}^{m} \delta_i p_i(x_i, \theta_i) \] (2)

where \( i=1,2,...,m \) is a mixture component, \( p_i(x_i, \theta_i) \) is density function, \( \theta_i \) is a parameter vector, and the parameter denoting the proportion of the population from the \( i \)th component is \( \delta_i \), with \( \delta_i \geq 0 \) and \( \sum_{i=1}^{m} \delta_i = 1 \).

The mixture of mixture models can be identified if there are \( k \) groups that have a mixture pattern [9]. A mixture of mixture models can be expressed as in Equation (3):

\[ f(x|\pi, \delta, \theta) = \pi_1 f_1(x|\delta_1, \theta_1) + \pi_2 f_2(x|\delta_2, \theta_2) + ... + \pi_k f_k(x|\delta_k, \theta_k) \] (3)

where \( \pi = \pi_1, \pi_2, ..., \pi_k \) is the proportion of the population from the \( k \)th group, with \( \sum_{j=1}^{k} \pi_j = 1 \), \( f_1(x|\delta_1, \theta_1) \) is a mixture model as stated in equation (2).
2.3. Hidden Markov Model

Hidden Markov Model (HMM) is a model in which the distribution that generates an observation depends on the state of an underlying and unobserved Markov process [4]. There are five elements of HMM, i.e. the number of hidden states in the model \(N\), the number of observable states \(M\), the state transition probability matrix \(\Gamma\), average number of vehicles \(\lambda\), and the initial state probability \(\delta\). The last three elements, i.e. \(\Gamma\), \(\lambda\), and \(\delta\), are called HMM parameters that have to be estimated to characterize the HMM [10]. One of the HMM algorithms used to predict the optimal hidden state sequence is the Viterbi algorithm.

2.4. Expectation-Maximization (EM) Algorithm

In the context of HMM, the EM algorithm is known as the Baum–Welch algorithm. The Baum–Welch algorithm is designed to estimate the parameters of an HMM whose Markov chain is homogeneous but not assumed stationary, that is, it is not assumed that \(\delta \Gamma = \delta\). The EM algorithm consists of two steps, i.e. E-step and M-step, that will be iteratively done until the HMM parameters have reached the convergence. The first step represents the stages for calculating the conditional expectations of the missing data using forward and backward algorithms. The second step is to maximize the log-likelihood function after the missing data have been estimated in the first step [4].

2.5. Bayesian Inference for Hidden Markov Model

As an alternative to the frequentist approach, one can also consider Bayesian estimation. Bayesian methods handle all unknown parameters by treating as random variables which have distribution. Bayesian would be coupled with the Markov Chain Monte Carlo (MCMC) algorithms which will estimate all of the posterior of parameters in the Bayesian model of HMM by employing the Gibbs sampler [11] and [12]. Suppose we have observations \(x^{(T)} = x_1, x_2, ..., x_T\) and the current values of the parameters \(\theta\) (represent both \(\Gamma\) and \(\lambda\)), then to simulate a sample path \(C^{(T)}\) of the Markov chain as in Equation (4):

\[
Pr(C^{(T)}|x^{(T)}, \theta) = Pr(C_T|x^{(T)}, \theta) \prod_{t=1}^{T-1} Pr(C_t|x^{(T)}, C_{t+1}^{(T)}, \theta)
\]  

(4)

In every generated sample path, the new estimated \(\theta\) can be gathered by drawing posterior \(\Gamma\) from Dirichlet distribution based on each row of the simulated sample path \(C^{(T)}\) of the Markov chain as a matrix of transition counts. Then, generate posterior \(\lambda\) from the gamma distribution by using its prior parameter which is updated based on the previously generated number of visit state in posterior of \(\Gamma\).

2.6. Accuracy

One measure that can be used to evaluate the accuracy of predictions is accuracy value. As in Equation (5), accuracy value can be expressed as the following notation: sample path \(C^{(T)}\) of the Markov chain as in Equation (4):

\[
\text{accuracy} = \frac{P_{11} + P_{22}}{P_{11} + P_{12} + P_{21} + P_{22}}
\]

(5)

where

- \(P_{11}\): the number of state 1 that are predicted correctly in state 1 on the next period,
- \(P_{12}\): the number of state 1 that are predicted in state 2 on the next period,
- \(P_{21}\): the number of state 2 that are predicted in state 1 on the next period,
- \(P_{22}\): the number of state 2 that are predicted correctly in state 2 in the next period.
2.7. Vehicle’s Density
Vehicle’s density is a condition of traffic jam that passes the road exceeded the capacity of the road plan and speed of this road approaching 0 km/hour [13]. It can be seen as the number of vehicles per unit length of the roadway. The two main impacts of the problem of vehicle’s density are travel time and road damage due to often bypassed by many vehicles.

3. Research Methodology
3.1. Data Source
The data used in this research is secondary data about the transaction of e-Toll at the exit gate PT Jasamarga Pandaan Tol. The data consist of 136,900 vehicles passing in Pandaan toll gate from May 15th to May 31st, 2018. The data will be divided into training and testing data. The training data used in this research is e-Toll transaction data from May 15th to May 28th, 2018 and testing data is e-Toll transaction data from May 29th to May 31st, 2018.

3.2. Research Variable
The research variables were the type of gate toll \((X_1)\) and combination of origin gate and vehicle category \((X_2)\) as described in Table 1:

| The Origin Gate | Vehicle Category | State | The Origin Gate | Vehicle Category | State |
|-----------------|------------------|-------|-----------------|------------------|-------|
| Bangil          | I                | B1    | I               | K1               |
|                 | II               | B2    | II              | K2               |
|                 | III              | B3    | Kejapanan       | III              | K3    |
|                 | IV               | B4    | IV              | K4               |
|                 | V                | B5    | V               | K5               |
| Gempol          | I                | G1    | I               | R1               |
|                 | II               | G2    | II              | R2               |
|                 | III              | G3    | Rembang         | III              | R3    |
|                 | IV               | G4    | IV              | R4               |
|                 | V                | G5    | V               | R5               |

The Hidden Markov Model as the structure scheme used in this study described in Figure 1.

![Figure 1. HMM Structure Scheme.](image)
3.3. Analysis Method
The steps of analysis in this study are as follows:
1. Descriptive statistical analysis and data exploration of vehicle’s density in Pandaan Toll.
2. Identifying mixture and mixture of mixture.
3. EM estimation for HMM parameters.
   a. Develop a computational algorithm to estimate the HMM parameters.
   b. Implement the algorithm into the R syntax.
   c. Analyze the model.
   d. Predict using the Viterbi algorithm.
   e. Calculate the accuracy value of this model.
4. Bayesian estimation for HMM parameters.
   a. Determine the likelihood function and prior distribution.
   b. Develop a computational algorithm to estimate the HMM parameters.
   c. Implement the algorithm into the R2OpenBUGS syntax and make a pilot run.
   d. Analyze the model.
   e. Predict using the Viterbi algorithm.
   f. Calculate the accuracy value of this model.
5. Choose the best of these two models by comparing their accuracy values.
6. Do the interpretation
7. Draw the conclusions and suggestions

4. Results and Discussion
4.1. Characteristics of Vehicle’s Density in Pandaan Toll
The Gempol-Pandaan toll is a strategic center for regional development, so find out the specific times of traffic jams in the Pandaan toll is very useful for toll road users.

Figure 2. Number of Vehicles in Pandaan Toll.

Figure 3. Number of Vehicles by Category.

Figure 4. Number of Vehicles by Gate Origin.
Figure 2 shows the density of vehicles on weekdays or weekend occurred at 05.00 a.m until 08.00 p.m. Meanwhile, Figure 3 and Figure 4 show that vehicles passing through the Pandaan toll are dominated by vehicles from the Kejapanan toll and included in category I.

4.2. Mixture and Mixture of Mixture
Results of the goodness of fit test to the interarrival time of 6 hidden states follow an exponential distribution, so the number of the vehicle per unit time will follow the Poisson distribution. The analysis, therefore, would employ a Bayesian Poisson approach. The mixture models used in this study are a conditional event which was the first vehicle comes only from one of the ith hidden states but the next vehicle that comes has the possibility of coming from all the jth hidden states, so there are 20 mixture models building the one mixture of mixture model. The model will be able to represent the probability of each vehicle came to the Pandaan toll exit gate.

4.3. HMM Parameters Estimation Using the EM Algorithm
HMM has 3 parameters, i.e. the average number of vehicles ($\lambda$), the initial state probability ($\delta$), and state transition probability matrix ($\Gamma$). In this study, there were 20 hidden states. It, therefore, had to estimate 440 parameters. HMM parameter estimation using EM algorithm stops at the 292 iterations with a tolerance value of 0.000001. As a result, this study has 20 Poisson mixture models where the contribution of each component in every mixture is represented as the transition probability from hidden state ith to hidden state jth. The Poisson mixture model of B1 (Bangil for vehicle category I), for example, can be written as Equation (6).

$$
\begin{align*}
\nu_{mix,B1} &= 5.2 \times 10^{-5} e^{-0.0928} + 4.69 \times 10^{-3} e^{-0.0939} + 5.18 \times 10^{-5} e^{-0.0939} + 4.27 \times 10^{-6} e^{-0.0939} + \ldots + 9.04 \times 10^{-6} e^{-0.0939} + 5.72 \times 10^{-7} e^{-0.0939} \\
&= 4.27 \times 10^{-6} \frac{1.0093^y e^{-1.0093}}{y!} + \ldots + 9.04 \times 10^{-6} \frac{1.0093^y e^{-1.0093}}{y!} + 5.72 \times 10^{-7} \frac{1.0093^y e^{-1.0093}}{y!}
\end{align*}
$$

(6)

and for writing the Poisson mixture of mixture model is as in Equation (7):

$$
\begin{align*}
\nu_{mixofmix} &= \delta_{B1}\nu_{mix,B1} + \delta_{B2}\nu_{mix,B2} + \ldots + \delta_{B5}\nu_{mix,B5} \\
&= 4.5 \times 10^{-6} \frac{1.0093^y e^{-1.0093}}{y!} + 1.34 \times 10^{-9} \frac{1.0093^y e^{-1.0093}}{y!} + \ldots + 5.98 \times 10^{-10} \frac{1.0093^y e^{-1.0093}}{y!}
\end{align*}
$$

(7)

where $y$ is the number of vehicles in an hourly interval of time.

4.4. HMM Parameters Estimation Using Bayesian Methods
The first stage in the Bayesian method is to determine the prior distribution as in Equation (8):

$$
\begin{align*}
\lambda_i &\sim \text{Gamma}(\alpha, \beta) \\
\Gamma &\sim \text{Dirichlet}(v_1, v_2, \ldots, v_{20}) \\
\delta &\sim \text{Dirichlet}(g_1, g_2, \ldots, g_{20})
\end{align*}
$$

(8)

where $\alpha, \beta$, and $v$ value are determined informatively based on data pattern, while $g$ value is a noninformative prior. In this study, we are using 10,000 iterations to obtain convergent parameters. The result of HMM parameters estimation using Bayesian shows that there were 440 parameters and have 20 Poisson mixture models where is the probability of the component of each mixture is its transition probability matrix. For a simple comparison, the Poisson mixture model of B1 can be written as Equation (9):

$$
\begin{align*}
\nu_{mix,B1} &= 0.0049 e^{-0.0034} + 9.95 \times 10^{-5} e^{-0.00049} + 5.98 \times 10^{-5} e^{-0.00049} + \ldots + 2.25 \times 10^{-6} e^{-0.00003} + 4.83 \times 10^{-7} e^{-0.00003} \\
&= 2.25 \times 10^{-6} \frac{0.00003^y e^{-0.00003}}{y!} + 9.95 \times 10^{-5} \frac{0.000049^y e^{-0.00049}}{y!} + \ldots + 4.83 \times 10^{-7} \frac{0.00003^y e^{-0.00003}}{y!}
\end{align*}
$$

(9)
while the Poisson mixture of mixture model can be written as Equation (10):

\[
f_{\text{mix of mix}} = \delta_{B1} f_{\text{mix}}_{B1} + \delta_{B2} f_{\text{mix}}_{B2} + \ldots + \delta_{B5} f_{\text{mix}}_{B5} = 0.03064 \left( 0.0049 e^{-0.00033} + \ldots + 4.83 \times 10^{-6} 0.000003 e^{-0.000003} \right) + \ldots + 0.00062 \left( 0.033 e^{-0.00034} + \ldots + 5.72 \times 10^{-6} 0.000003 e^{-0.000003} \right) \]

(10)

As an interpretation example of Equation (9), the Poisson mixture model said that the probability of vehicle categories I from the Bangil toll and for 19 other Poisson mixture models also have the same meaning as the mixture Poisson model for B1. The higher the probability value of a hidden state, the higher the vehicles with certain types of categories will pass through the origin of the toll gate. While the Poisson mixture of mixture model as in Equation (10) can be interpreted as the probability of vehicles coming in Pandaan toll exit. The higher the probability value, the more the vehicle category I, II, III, IV, and V which passes through the Pandaan toll.

Figure 5 shows that the estimated initial probability of each hidden has a pattern that is almost the same as the value of the input parameters. This condition occurs because the Bayesian approach treats all parameters as random variables and are estimated by generating their posterior individually.

### 4.5. Model Selection

Modeling to predict vehicle density in this study was conducted using two estimation methods for HMM parameters, i.e. EM algorithm and Bayesian approach. By selecting the higher accuracy value between those two estimation methods, as shown in Table 2, HMM parameters estimated by using Bayesian is better than the EM algorithm.

| Model            | Accuracy |
|------------------|----------|
| EM Algorithm     | 87.18%   |
| Bayesian Methods | 87.23%   |

| Figure 5. Parameter Comparison of Parameter δ with ˆδ |

Figure 5 shows the estimated initial probability of each hidden has a pattern that is almost the same as the value of the input parameters. This condition occurs because the Bayesian approach treats all parameters as random variables and are estimated by generating their posterior individually.
By using the selected best methods, vehicle’s density was also reviewed based on initial probabilities and the average number of vehicles. Figure 6 and Table 3 show that the density of vehicles in every category could occur in the Kejapanan toll to the Pandaan toll, followed by vehicles from the Gempol toll to the Pandaan toll, then vehicles from the Rembang toll, and finally, as the smallest possibility is vehicles from the Bangil toll to the Pandaan toll.

![Figure 6. Initial State Probability](image)

Table 3. Average Number of Vehicles

| Category | The Origin Gate | Bangil | Gempol | Kejapanan | Rembang |
|----------|----------------|--------|--------|-----------|---------|
| I        | 0.0003350      | 0.602900 | 10.29000 | 0.001697  |
| II       | 0.0000049      | 0.000306 | 0.003309 | 0.000038  |
| III      | 0.0000016      | 0.000109 | 0.000593 | 0.000017  |
| IV       | 0.0000002      | 0.000004 | 0.000189 | 0.000002  |
| V        | 0.0000005      | 0.000005 | 0.000089 | 0.000003  |

5. Conclusion

Based on the results of the analysis, it was obtained that vehicles passing through the Pandaan toll exit gate are dominated by vehicles category I from the Kejapanan toll. The estimation of the HMM parameters is more precise when they are estimated using Bayesian than the EM algorithm. It has been shown through the probability value of the components forming the mixture model or Poisson mixture of mixture is more representative to predict vehicle density. For future research, it was necessary to develop the Hidden Markov Model analysis using data at longer intervals by preparing hardware that supports big data analysis and for exponential mixture modeling to predict the interarrival time of vehicle at the Pandaan toll. One recommendation to PT Jasamarga is that with high vehicle density, especially those from the Kejapanan Toll Road, maintaining or increasing the accuracy and speed of the e-toll card reader machine can help the completeness of the Pandaan toll user data for supporting the HMM modeling and analysis as the future big-data vehicles density analytics research.
Acknowledgments
The authors would like to express our gratitude to reviewers for the criticism and suggestions for the improvement of this paper.

References
[1] Statistics Indonesia 2017 Land Statistics Transportation (Jakarta: BPS RI) p 22
[2] Government Regulation of The Republic of Indonesia 2005 Government Regulation of The Republic of Indonesia Number 15 Toll Road (Jakarta)
[3] Jasamarga Pandaan Tol 2015 Company Profile (https: pt.jpt.co.id/profil-perusahaan)
[4] Zucchini W, MacDonald I L and Langrock R 2016 Hidden Markov Models for Time Series (2nd ed) (London: Chapman & Hall) pp 29-116
[5] Sherlock C, Xifara T, Teller S and Begon M 2013 A Coupled Hidden Markov Model for Disease Interaction Journal of The Royal Statistical 62 609-627
[6] Sidney S and Castellan N J 1988 Nonparametric statistics for the behavioral sciences 2nd (New York: McGraw-Hill) pp 47-52
[7] Gelman A, Carlin J B Stern H S and Rubin D B 2004 Bayesian Data Analysis (2 nd es) (London: Chapman and Hall) pp 463-479
[8] Iriawan, N 2011 Pemodelan Mixture of Mixture dalam Pemilihan Portfolio Prosiding Seminar Nasional Statistika, Universitas Diponegoro
[9] Astuti, A. B., Iriawan, N., Ihamah, and Kuswanto, H 2015 An Algorithm for Determining the Number of Mixture Components on the Bayesian Mixture Model Averaging for Microarray Data Journal of Mathematics and Statistics 11 (2): 45-51, DOI: 10.3844/jmssp.2015.45.51
[10] Djuric, P.M. and Chun, J 2002 An MCMC Sampling Approach to Estimation of Nonstationary Hidden Markov Models IEEE TRANSACTIONS ON SIGNAL PROCESSING 50 1113-23
[11] Susanto, I., Iriawan,N., Kuswanto, H., Suhartono, Fithrisari, K., Ulama, B.S.S., Suryaningtyas, W, and Pravitiasari, A.A 2018 On The Markov Chain Monte Carlo Convergence Diagnostic of Bayesian Finite Mixture Model for Income Distribution IOP Conf. Series: Journal of Physics: Conf. Series 1090 (2018) 012014. DOI: 10.1088/1742-6596/1090/1/012014
[12] Mahmud, M.P. and Schliep, A 2011. Fast MCMC Sampling for Hidden Markov Models to Determine Copy Number Variations BMC Bioinformatics 12(1):428, 1-17
[13] Directorate General of Highways 1997 Indonesian Road Capacity Manual (Jakarta)