Modelling Rail Accident and Incident Causes by Using Zero-Inflated Poisson Approach

Joewono Prasetijo[1*], Ishak Baba[2], Khalid Hasnan[1], Shamsullarifin Ismail[3], Mohd Ali Ridho Khairul Anuar[1], Mohd Azri Mat Aron[1], M Ikhsan Setiawan[4], Sri Wiwoho Mudjanarko[4]

[1]Department of Transportation Engineering Technology (STSS), Faculty of Engineering Technology,
 Universiti Tun Hussein Onn Malaysia, 84600 Panchor, Johor, MALAYSIA

[2]Faculty of Engineering Technology,
 Universiti Tun Hussein Onn Malaysia, 84600 Panchor, Johor, MALAYSIA

[3]VS Industri Bhd.,
 81700 Pasir Gudang, Johor, MALAYSIA

[4]Department of Civil Engineering,
 Narotama University, Surabaya, Jawa Timur 60117, INDONESIA

*Corresponding Author

Abstract: The development of Railway industry has been growing rapidly until today, which make it as one of the popular choice of transportation mode to travel from one place to another and it becoming more complex. Thus, the complexity of rail network required high level of safety features to prevent any unwanted incident. Therefore, this study proposed a proper procedure on modelling accident which is conducted by using Poisson model. The most contributory factor that influenced the accident can be identified by using root cause analysis. "Ishikawa diagram" is a popular tool to identify problem occurring from the root where it begins. The data were taken from several sources which is secondary data where the data period was starting from 1999 to 2014. Analysis from Ishikawa shows there are ten main factors involved to influences an accident. Those factors are "train driver mistake", "other's human mistake", "weather influence", "track problem", "train problem", "signaling error", "maintenance error", "communication error", "procedure error", and "others". Then, the model was tested to know which model among regression model is suitable and give a better prediction result by carrying out Dispersion test and Vuong test. The results show that Zero-inflated model considered as a sophisticated model to predict accidents and incident cases by Vuong test with p-value of 0.19695481, 0.1301056 and 0.0689108. The most factors contribute to the cases are "collisions", "derailment" and "SPAD".

Keywords: Rail accidents, safety, prediction model, regression model, Poisson

1. Introduction

Rail transportation is one of the biggest transportation modes on the planet. This transportation was extremely mainstream among well-developed nation. It additionally has multi-reason of functionality. Right on time of their invention, rail transportation was used to bring coal, mineral and others industrial products to the distribution center and port. Rail system is not just to bring products, yet also as transportation mode to move individuals to wherever they

desire to go. The rate of advancement of rail system makes it more entangled and complex. Thus, risk assessment is essential in initial planning to enhance railway system as far as security for time ahead. Rail mishap and occurrence is a main consideration that will make intrusion the rail task. For that, more research that focused on this issue is needed. Such as in the United States, the Federal Highway Administration (FHWA) periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement at railroad crossing. They require each state to develop and implement a highway safety improvement program (HSIP), which consists of three components. Those are planning, implementation, and evaluation [1].

Through in deep analysis, underlying cause leading to the incident occurrence events can be distinguished. On the contrary, effectiveness of countermeasure had been taken before it can be assessed. As per, further effort can be done properly to improve safety performance and obliterate error occurrence from starting. By using mathematical modeling method, connection and coefficient estimation of each factor can be computed. Expectation model is used to enhance monitoring system of railway system to conquer and prepare of any risk and potential mishap happen. Nevertheless, statistical data will help administration to keep up and monitor the safety or well-being performance of rail track network. Since every country has an alternate geographic and operating system of rail in their networks, prediction model could be hard to be done for their independent variable or factor impact. Therefore, there is no appropriate investigation system to foresee the rail mishap and occurrence. Accordingly, this paper will focus on the strategy how to clarify the most impact factor will influence to the accident or incident and become independent variable for expectation model. It is blend between the root cause analysis to find the contributory factor and expectation model as predictor. Subsequently, priority of each factor can also be found. Since there are tons of model for regression, this procedure additionally to distinguish which model is most reasonable to use after certain steps and hypothesis test. At that point, this method can be standardizing in rail industry all around. This study develop investigative procedure to predict factor influence of railway accident, to identify contributory factor that influencing to the rail accident using Ishikawa diagram through case study of railway accident investigation report and to formulate the prediction model of railway accident and incident using regression formulae.

2. Methodology

2.1 Study area and data collection

The data is taken from several official rail authorities and the data are the investigation report of accident and incident starting from 2009 until 2014 which is published for public references. Investigation reports were concerned about several categories of accident and incident such as collision, derailment, signal pass at danger (SPAD), and others (mostly for the incident). All the categories are related to the type of transport such as passenger train, freight train, and others (road vehicle, rail vehicle, rail machine).

2.2 Extracting data to Microsoft Excel

The summarized and extracted data consisted of the contributory factor and the number of transport involved in an accident as shown in Table 1.

Table 1 - Number of transport involve in accident and incident.

	Collision	Derailment	SPAD	Others	Total
Accident Cases	47	59	20	31	157
Passenger Train	24	20	10	13	67
Freight Train	18	38	9	16	81
Others Vehicle	36	2	0	10	48

The data shows derailment was the highest with the number of accident reported by 59 cases, it is followed by collision with 47 cases, others with 31 cases and SPAD with 20 cases. The data then were extracted further containing the element of human error, system error and environment influence as shown in Table 2.

Table 2 - Number of error involve in accident and incident.

	Human Error	System Error	Environment influence
Collision	38	9	3
Derailment	25	37	12
SPAD	19	5	0
Others	19	15	1
Total	101	66	16

The data were extracted and tabulated to be clarified about the cases and causes. Furthermore, Ishikawa diagram is used to sort the data for root cause analysis.

2.3 Root cause analysis using Ishikawa diagram

Root cause analysis as shown in Fig. 1 help to understand and to trace the causes or factor influences towards each type of accident and incident cases [2]. Ishikawa Diagram make it easier to figure out the flow of effect for each other factor to the type of accident and incident through its element [3]. This paper would study to capture factor influence with higher impact in increment number of accident and incident cases. Three element (human, environment, and system) become main categories of Ishikawa diagram for each type (collision, derailment, SPAD, and Others) of accident and incident event. Contributory factor getting from the Ishikawa diagram is used to predict the model to evaluate the general form or specified by its element.
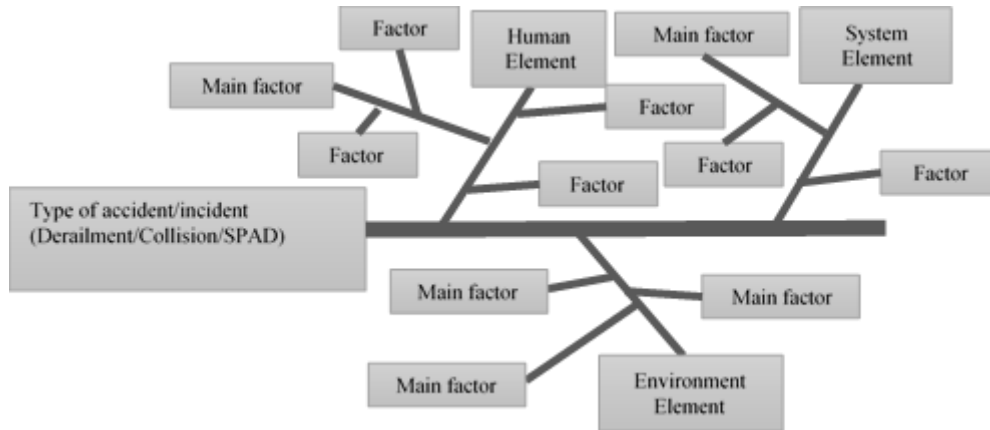

Fig. 1 - Sample Ishikawa diagram.

2.4 Develop prediction model

Prediction analysis was created to predict each factor influence for future. The best estimation is found for prediction value, and some statistical hypothesis will be tested. Then, each type of prediction model would be compared to hypothesis test result in purpose to find out which model was most fit. Due to the complex prediction calculation, therefore; R-studio will be used as a tool with extension of "pscl" and "AER" package.

2.5 Correlation coefficient value

Correlation coefficient is a measure of association between two variables. The sign of correlation coefficient depends on whether the variables are positively or negatively related between each others with a value range -1 to 1. The Pearson product-moment correlation coefficient is more widely used in measuring the association between two variables. Given paired measurements $(X1,Y1)$, $(X2,Y2)$, . . . ,(Xn,Yn), the Pearson product-moment correlation coefficient is a measure of association given by [4].

$$r_p = \frac{\sum_{i=1}^{n}(X_i-\bar{x})(Y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{x})^2 \; \sum_{i=1}^{n}(Y_i-\bar{y})^2}} \qquad (1)$$

Where X and Y are the sample mean of X1, X2, . . ., Xn and Y1, Y2, . . . , Yn, respectively. For this case, two variable to be measured is between type of accident and incident with the factor influence. Correlation coefficient with positive value will be accepted as the most influential factor and it would be consider as independent variable for prediction model. This process is one of the preliminary filter of selecting the variable for independent variable before start doing predicting process.

2.6 Regression model

Following the procedure, Poisson regression model was calculated first. First, group of variable need to be separated between dependent variable (y_i), then independent variable (λ_i) as below:

Dependent variable (y_i) = Accident cases (type of accident) Independent Variable (λ_i) = factor 1 + factor 2 + factor 3 +.......+ factor n... (from correlation coefficient result).

Thus, Poisson regression model will be form as below.

$$P(y_i) = \frac{e^{(-\lambda_i)}(\lambda_i^{y_i})}{y_i!} \qquad (2)$$

With parameter value;

$$\log\lambda_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3+} \ldots\ldots + \beta_k x_k \qquad (3)$$

Where is $\beta_0, \beta_1, \ldots \beta_n$ is a parameter of contributory factor. Other than normal distribution model, Zero-Inflated model also was calculated in this research. Usually, this model is used for data with excess zero data. Explanation this model as per below[5].

In case of Zero-Inflated Poisson model, the dependent variable Y_i become:

$$Y_i = \begin{cases} 0 & \text{with probability } P_i \\ \text{Poisson } (\lambda_i) & \text{with probability } 1 - P_i \end{cases} \qquad (4)$$

Thus, the density function of Y_i is follow

$$P_r(Y_i = y_i) = \begin{cases} P_i + (1 - P_i)\exp(-\lambda_i) & y_i = 0 \\ (1 - P_i)\dfrac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} & y_i > 0 \end{cases} \qquad (5)$$

Where $0 \le P_i \le 1$

Mean,

$$E[Y_i] = \lambda_i (1 - P_i) \qquad (6)$$

Var,

$$Var[Y_i] = \lambda_i (1 - P_i)(1 + \lambda_i P_i) \qquad (7)$$

Thus, parameter of λ_i and P_i as follow:

$$\ln(\lambda_i) = x_i\beta \qquad (8)$$

$$\lambda_i = \exp(x_i\beta) \qquad (9)$$

$$logit(P_i) = Z_i\gamma \qquad (10)$$

$$P_i = \frac{1}{1+\exp(-Z_i\gamma)} \qquad (11)$$

Where β is the regression coefficient from log-linear generalized linear mix model γ is the regression coefficient from logistic generalized linear model x_i and Z_i are vectors of independent variable.

2.7 Dispersion test

Dispersion test would be tested to know either data collection was under dispersion or over dispersion. This test will calculate and comparing mean $E[Y_i]$ and variance $Var[Y_i]$ of the model. In R- studio, this test use "pscl" package. When dispersion test was < 1, it means data collection was under dispersion with mean value $E[Y_i]$ was less than variance value $Var[Y_i]$. Thus, Poisson model is appropriate. But, when dispersion test was > 1, it shows that data collection was over dispersion, with a mean value $E[Y_i]$ was greater than variance value $Var[Y_i]$ [6]. Thus, Negative binomial model is appropriate [7]. Hence, parameter value will become;

$$log \lambda_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3+} \ldots \ldots + \beta_k x_k + \xi \qquad (12)$$

2.8 Vuong Test

Two models (Poisson model with Zero-Inflated Poisson model, or Negative Binomial model with Zero-Inflated Negative Binomial model) will be compared and this test was carried out after getting the results of zero-inflated Poisson model (ZIP). This test use "pscl" package in R-studio. It compares log-likelihood value of model and measure P-value, AIC number and BIC number to find which model were more fitted. Referring to the statistical analysis, P value will show either model was significant or not. Model with P-value < 0.05 mean model was statistical significant, thus, it would be rejecting null hypothesis. While model with P-value > 0.05, data was accepting null hypothesis. It could be happen when the data taken is less in number. The best prediction model among regression model was selected as predictor model use in this research.

3. Results and Discussion

3.1 Ishikawa diagram

Fig. 2 shows the Ishikawa diagram of derailment with 8 main tributary factor. Those are "communication error", "other's human mistake", and "driver mistake" under human element; "train problem", "poor maintenance"; "track problem" and "signaling error" under system element; and "bad weather" under environment element. It also shows current procedure is not enough to prove this system in high level of safe condition. Result from root cause analysis it can be conclude that there are 10 contributory factor has been selected to be independent variable for prediction model process.
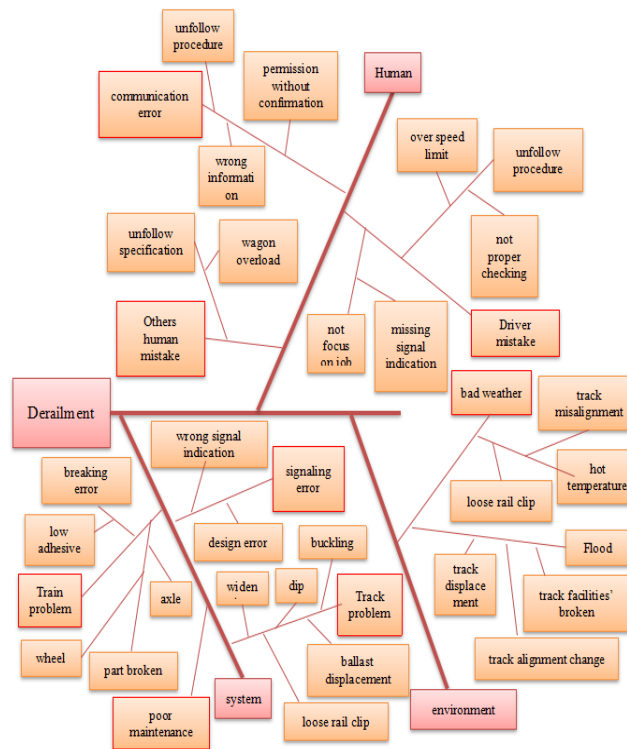

Fig. 2 - Ishikawa diagram of derailment.

3.2 Correlation coefficient value

Table 3 shows the result of correlation coefficient for those 10 independent variable of contributory factor to the type of accident and incident.

Table 3 - Correlation coefficient value for all variables.

	Collision	Derailment	SPAD
Train driver mistake	-0.0462	-0.2026	0.3551
Others human mistake	0.5596	-0.3188	-0.1741
Weather Influence	-0.0919	0.2441	-0.1148
Track Problem	-0.2915	0.5297	-0.2122
Train Problem	-0.1654	0.0936	-0.1663
Signal Error	0.0255	0.0243	-0.0380
Maintenance Error	-0.2119	0.2746	-0.1501
Communication Error	0.1516	-0.2839	0.0790
Procedure Error	0.0897	-0.0980	-0.0160
Others Factor	-0.0781	-0.0440	0.1239

3.3 Poisson regression model

From the correlation coefficient value, the dependent variable is type of accident, which is collision, derailment, and SPAD. An independent variable is contributory factor with positive value of correlation coefficient to the type of accident. Thus, Poisson regression model would become [8]:

$$P(y_i) = \frac{e^{(-\lambda_i)}(\lambda_i{}^{y_i})}{y_{i!}}$$

With parameter value;

$$\log \lambda_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{3+} \ldots + \beta_k x_k$$

Where is $\beta_0, \beta_1, \ldots \beta_n$ is a parameter of contributory factor and $x_1, x_2, \ldots, x_n$ is a estimated coefficient value of parameter.

Poisson model for collision can be derive in below Equation 13, derailment in Equation 14 and SPAD in Equation 15.

1.   Collision, Pc

One identified factor as collision is based on the several derailment accident and incident, (y_i), the following factor is considered, λ_i.

λ_i = Independent variable of "other's human mistake", "signal error", "communication error", and "procedure error".
λ_i = exp [-2.0417 + 1.7489(other's human mistake) + 0.7013(signal error) + 0.4324(communication error) +0.5027(procedure error)   (13)

2.   Derailment, Pd

Derailment factor, (y_i) is considered by the following factor, λ_i.

λ_i = Independent variable of "weather influence", "track problem", "train problem", and "signal error", and "maintenance error".
λ_i = exp [-1.7634 + 0.4957(weather influence) + 1.3308(track problem) + 0.7694(train problem) + 0.6075(signal error) + 0.1742(maintenance error)]………...(14)

3.   Signal Pass at Danger (SPAD), Ps

SPAD is another potential cases of accident/incidents and predicted as

λ_i = Independent variable of "weather influence", "train driver mistake", "communication error", and "other's factor".
λ_i = exp [-3.2194 + 1.8519 (train driver mistake) + 0.6606(communication error) + 1.3194(other's factor)]   (15)

Equation 13, Equation 14, and Equation 15 shows there are four variable required to be focused and pay attention. Those are "other's human mistake" for collision, "track problem" for derailment, and "train driver mistake" and "other's factor" for SPAD. This is because from the prediction result shows their p-value is significant to their dependent variable. Increment of significant value of p-value with less than 0.05 would increase the estimated coefficient value of the variable. Expected log count for one-unit cases increase in other's human mistake is 1.7489 for collision, track problem with 1.3308 for derailment, train driver mistake with 1.8519 and other's factor with 1.3194 for SPAD. Thus, this estimated coefficient also able to use to prioritizing the factor need to be solved in order to improve the safety in railway.

3.4 Dispersion Test

Dispersion test shows in Table 4 state that all category has under dispersion value which is the value is < 1. Thus, some of possibility prediction model can be used or fit to this data is Poisson and zero-inflated Poisson.

Table 4 - Dispersion test result.

Category	Dispersion Test
Collision, Ps	0.7070064
Derailment, Pd	0.6242038
SPAD, Ps	0.8726115

3.5 Zero-inflated poisson model (ZIP)

The following model of ZIP have been done in several studies [9]. The result of Zero-inflated other model for collision, derailment and SPAD.

1. Collision, other

λ_i = Independent variable of "other's human mistake", "signal error", "communication error", and "procedure error".

a) Poisson model log link

$$\lambda_i = \exp [-1.4450 + 1.2287(\text{other's human mistake}) + 0.2433(\text{signal error}) + 0.2069(\text{communication error}) + 0.2505 (\text{procedure error})] \qquad (16)$$

b) Zero-inflation model with logit link

$$P_i = \frac{1}{1 + \exp(-(z_i \gamma))}$$

Where:

$$Z_i \gamma = 0.2082 + [-13.2968(\text{other's human mistake})] + [-6.1826 (\text{signal error})] + [-11.0464(\text{communication error})] + [-2.2365 (\text{procedure error}) \qquad (17)$$

2. Derailment, ZIPd

(y_i) = Derailment accident and incident cases

λ_i = Independent variable of "weather influence", "track problem", "train problem", and "signal error", and "maintenance error".

a) Poisson model log link

$$\lambda_i = \exp [-1.028563 + 0.259199(\text{weather influence}) + 0.775256(\text{track problem}) + 0.184611(\text{train problem}) + 0.473201(\text{signal error}) + (-0.009169(\text{maintenance error})] \qquad (18)$$

b) Zero-inflation model with logit link

$$P_i = \frac{1}{1 + \exp(-(z_i \gamma))}$$

Where ;

$$Z_i \gamma = -0.3610 + [-12.8908(\text{weather influence})] + [-18.0759(\text{track problem})] + [-20.3536(\text{train problem})] + [-0.3965(\text{signal error})] + [-11.6307(\text{maintenance error})] \qquad (19)$$

3. Signal Pass at Danger (SPAD), Ps

(y_i) = SPAD accident and incident cases

λ_i = Independent variable of "weather influence", "train driver mistake", "communication error", and "other's factor".

a) Poisson model log link

λ_i = exp [-1.4692+ 0.3586(train driver mistake)] + [-0.3231(communication error) + 0.7895(other's factor)]   (20)

b) Zero-inflation model with logit link

$$P_i = \frac{1}{1+\exp(-(z_i\gamma))}$$

Where ;

$Z_i\gamma$ = 2.789 + [-14.501(train driver mistake)] + [-13.140(communication error)] + [-2.211(other's factor)]   (21)

Above model of ZIP has two predictors had tested for collision, derailment and SPAD. Such model tested have also implemented for road geometric infrastructure [9]. In Equation 16, Equation 18 and Equation 20 is a result of ZIP using count model, while Equation 17, Equation 19 and Equation 21 shows result of ZIP using logit model. Most of count model give similar result in pattern with Poisson model, but smaller in number of estimated coefficient value for all variable since intercept point is change. Those result also reflect to the significant result of variable. Most of variable for count model shows increment of expected log count, however not for maintenance error in equation 18, which is their expected log count of the response variable is decreased by 0.009169. It also same for communication error in equation 20 show decrements expected log count by 0.3231. For the logit model, result shows entire prediction model has negative value of estimated coefficient. That mean, the inflated coefficient of the variable (in Equation 17, Equation 19 and Equation 21 is decrease in log odd of an inflated zero. Since ZIP model is measure sample data with excess zeros, quantity number of sample data can influence the prediction result. Huge number of sample data is much better for this model.

3.6 Vuong Test

Table 5 shows result for Vuong test. After comparing both model, from the Vuong z-statistic result, it shows that zero-inflated Poisson model (ZIP) is more fitted rather than Poisson model (P) for all type of accident with p-value of ZIPc, ZIPd and ZIPs is 0.19695481, 0.1301056 and 0.0689108 respectively. Event p-value result is less significant; however, it shows ZIP model is better than Poisson model. For that, Zero-inflated Poisson model was selected as prediction model in this study which is also proved by several studies for highways [10 - 11] and it was based on the current technique of choosing the suitable routes/locations [12 - 13].

Table 5 - Vuong test result.

Category	Vuong z-statistic	H_A	p-value
Raw Collision (Pc, ZIPc)	-0.8525487	model2 > model1	0.19695481
Raw Derailment (Pd, ZIPd)	-1.125892	model2 > model1	0.1301056
Raw SPAD (Ps, ZIPs)	-1.4838898	model2 > model1	0.06891908

4. Summary

Safety is always the top priority in any system or organization and since the railway transportation has become the daily transportation mode among the people, high safety features is required to ensure the effectiveness and efficiency of the railway transportation. This model approach is significant to recognize the problem of accident and incident cases deep to their root of causes. The process using Ishikawa diagram found that there are 10 contributory factor need to be focusing. Furthermore, entire contributory factor had tested to find their correlation between type of accident and incident to the contributory factor. Result of correlation coefficient shows there are four (4) factor has a positive relation to the collision, five (5) factor has positive relation to the derailment, and 3 factor has positive relation to the SPAD (Table 3).

Poisson model is tested to know whether the data is under dispersion or over dispersion. Thus, result shows that entire data for collision, derailment and SPAD is under dispersion with value < 1. The result is 0.7070064, 0.6242038, and 0.8726115 respectively. That mean there are only two possible prediction model can be used either Poisson model or Zero-inflated Poisson model can be fitted. For that, Zero-inflated model is applied as predictor before comparing both models using Vuong test. This Vuong test shows that Zero-inflated model has greater P-value rather than Poisson

model for "collisions", "derailment" and "SPAD" (Table 5). Because of that, Zero-inflated Poisson is more suitable to use to predict accident and incident cases of railways for collision, derailment and SPAD.

Acknowledgement

References

[1]   The Office of Transport Safety Investigations, Signal passed at danger by track machine consist 8M71, No. 04439, (2009),

[2]   Hughes R. G., Section VI: Tools for quality improvement and patient safety, *Patient Saf. Qual. An Evid. -Based Handb. Nurses*, (2008), pp. 1–39.

[3]   Vorley, G. Mini guide to root cause analysis, *Quality Management and Quality*, (2008).

[4]   Chen, W.-K. *Linear Networks and Systems*, Belmont, CA: Wadsworth, (1993), pp. 123–135.

[5]   Smith, B. "An approach to graphs of linear forms (Unpublished work style)," unpublished.

[6]   Schultz, T. G, Evaluation of safety at railroad - highway grade crossing. *Joint Transportation Research Program*, *US,* No. FHWA/IN/JHRP-65/29 (1965).

[7]   Austin, R. D. and Carson, J. L., An alternative accident prediction model for highway-rail interfaces. *Accident Analysis & Prevention,* Volume 34, (2000), pp. 31–42.

[8]   Oh, J., Simon P. W., and Doohee N, Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention*, 2006, Volume 38, (2006), pp. 346–356.

[9]   Prasetijo, J., and Musa, W. Z., Modeling zero-inflated regression of road accidents at Johor federal road F001. *MATEC Web of Conferences*, Volume 47, No. 03001, (2016).

[10]   Prasetijo, J. and Zainal, W. Z., Development of continuous speed profile using GPS at Johor federal roads F0050. *MATEC Web of Conferences*, Volume 47, No. 03001, (2016).

[11]   Prasetijo, J., Musa, W. Z. and Zainal, W. Z., Road fatality model based on over-dispersion data along federal route F0050. *MATEC Web of Conferences*, Volume 103, No. 080122, (2017).

[12]   Omar, N., Prasetijo, J., David Daniel, B. and Abdullah, M. A. E., Accident analysis and highway safety. *MATEC Web of Conferences*, Volume 103, No. 08002, (2017).

[13]   Hosseinpour, M., Yahaya, A.S., Ghadiri, S.M., Prasetijo, J., Application of adaptive neuro-fuzzy inference system for road accident prediction. *KSCE Journal of Civil Engineering*, Volume 17, (2013), pp. 1761–1772.