# Analysis of CI Behavior due to Independent Variable Value Variation, Case of Simple Linear Zonal Regression Trip Production Model

Hitapriya Suprayitno, Dio Hananda, and Jimi Aditya

*Abstract*— **Good Transportation Planning needs a good and accurate Transport Model. Trip Production Modeling, as the first step, needs to be accurate also. A new Trip Production Accuracy measure, incorporating $R^2$ and Confident Interval values, has been proposed. An experiment to investigate the CI Behavior, due to the variation of Independent Variable Value, needs to be done. The experiment indicates that the CI Value is very useful accuracy measure since it can give the picture of Predicted TP Accuracy. The experiment indicates also that higher Independent Variable Value gives higher CI Value, measured in percentage. A further research to determine accepted CI Value still needs to be executed.**

*Index Terms*— **confidence interval, confidence interval behavior, trip production model, simple linear zonal regression.**

## I. Introduction

Transport Model (TM) is always needed for Transportation Planning. In general, we have 3 TM types : Direct Model (TDM), Conventional Model (TCM), and UnConventional Model (TUCM). The TCM is still good to be used [1-6]. In order to produce a good Transportation Plan, a good and accurate TM is needed. Method or technique for achieving accurate enough TM needs to be built.

The first step of TM is Trip Production (TP) Modeling. TP Model has been developed in three types : the Zonal Regression (ZR), the Catagory Analysis (CA) and the Household Regression (HR) TP Models. For Indonesian practice, the ZR TP Model seems the most appropriate. Researches have developed several types of ZR TP Models, among others are : Simple Linear Zonal Regression (SLZR), Multivariate Linear Zonal Regression (MLZR) and even Non-Linear Zonal Regression (NLZR). The SLZR TP Model is the most practical to be used for professional work and widely used in Indonesia [1-4,7-9].

The most practical TP Model is the Simple Linear Zonal Regression (SLZR). Several experiments in Regression shows that, for certain cases, a better coefficient of determination ($R^2$) does not always produce a more accurate prediction. Good $R^2$ can be easily gotten from a small sample size. Therefore, a better accuracy measure must be formulated. A proposal of accuracy measure has been set. It consists of : a good sampling method, a set of appropriate statistical test, and the use of $R^2$ and Confidence Interval (CI) [8-11]. An investigation of the correlation between Sample Variation, Predicted Values, $R^2$ Values, and CI Values has been done, based on a very small sample. Several important indications have been gotten. The findings strengthen the idea for using $R^2$ and CI as the indicator for modeling accuracy [12]. A more serious experiment based on real Transport Modeling cases needs to be done.

This paper presents an attempt to investigate the effect of the Ratio between Independent Variable Value to the Sample's Maximum Independent Variable (Ratio of $IVV/IVS_{max}$) on CI value. A real case of SLZR Trip Production Modeling was taken.

## II. Research Method

The research was executed by following these method steps : background statement, objective designation, research method, literature review, data processing and analysis, and finalized by conclusions

## III. Literature Review

The Trip Production (TP) as a part of the Trip Generation Model, usually is measured as home base trip production. The TP can be modeled by zonal regression (ZR), category analysis (CA), and household regression (HR). For professional work, the ZR is normally modeled as a Simple Linear Zonal Regression TP Model (SLZR TP Model) [1-4,7].

The SLZR TP Model must be accurate enough. It has been indicated that a model with better $R^2$ does not always give better-predicted value [8]. New accuracy measures have been proposed, which incorporates the $R^2$ and CI values [9].

The Simple Linear Regression formula is presented below [10,11].

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

$$\hat{Y} = a + b\,X \qquad (2)$$

H. Suprayitno is with the Civil Engineering Deparment, Institut Teknologi Sepuluh November(ITS) Surabaya (e-mail: suprayitno.hita@gmail.org)

D. Hananda is with the Alumnee of Civil Engineering Deparment, Institut Teknologi Sepuluh November(ITS) Surabaya.

J. Aditya is with the Alumnee of Civil Engineering Deparment, Institut Teknologi Sepuluh November(ITS) Surabaya.

$$\hat{\beta}_1 = b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad (3)$$

$$\hat{\beta}_0 = a = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (4)$$

where :
Y : the dependent variable
X : the independent variable
$\hat{Y}$ : estimates from the regression equation
$\bar{Y}$ : means of Y
a : slope
b : intercept
n : number of observations

The coefficient of determination ($R^2$) formula is presented below [10,11].

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \qquad (5)$$

where :
$R^2$ : coefficient of determination
Y : the dependent variable
$\hat{Y}$ : estimates from the regression equation
$\bar{Y}$ : means of Y

The Confidence Interval (CI) image is presented in Fig. 1 below, while the formula is also presented below [10,11].

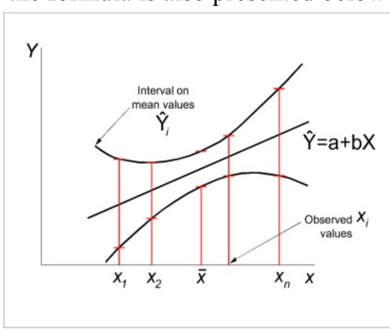


Fig. 1. Confidence Interval of Predicted Value

$$CI\ \hat{Y} = \hat{Y} \pm t_{\frac{\alpha}{2};df=n-2}\ S_{\hat{Y}} \qquad (6)$$

$$S_{\hat{Y}} = s\left\{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right\}^{1/2} \qquad (7)$$

s = □ □$_i^2$ / □□□□□□□$y_i$ - y□$^{□}$□□□n-2
(8)

where :
CI $\hat{Y}$ : confidence interval of predicted Y value (CIoPV)
$Y_i$ : the value of observed $Y_i$
$\hat{Y}_i$ : the value of predicted $Y_i$
$X_i$ : the value of observed $X_i$
n : number samples
$S_{\hat{Y}}$ : standard deviation of $\hat{Y}$
s : standard deviation of $Y_i$
t : student distribution value for a certain confidence value

## IV. CI CALCULATION AND ANALYSIS

### *A. Research Case*

The research case was taken in Gresik City, East Java. A Trip Production Modeling of motorcycle trip during morning peak hour was calculated, for Gresik Urban Area. The data was taken by Household Interview Survey (HIS), with a sample of 800 households. The Modeling Area is divided into 25 zones, based on kelurahan administrative area or a group of kelurahans. The modeling area covers the Gresik District, a part of Kebomas and Manyar Districts. Gresik Urban Are Map, Motorcycle Population Data and HIS Data are presented below in Fig. 2 and in Table I.

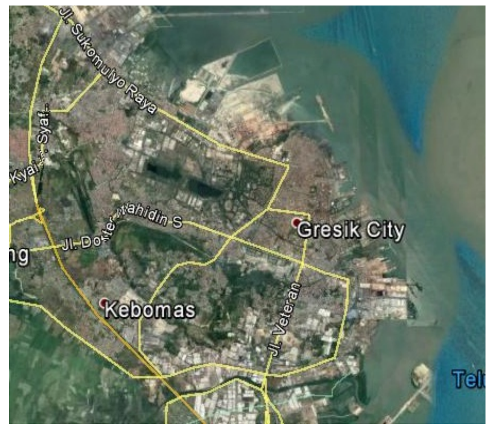


Fig. 2 Map of Gresik City

TABLE I
ZONAL POPULATION AND HIS DATA

| No. | Kecamatan | Zone | Population | | HIS Data | |
|---|---|---|---|---|---|---|
| | | | Population | Motorcycle | Motorcycle | Trip |
| | | | *person* | *motorcycle* | *motorcycle* | *trip/hr* |
| 1 | Gresik | Ngipik | 1,694 | 935 | 12 | 9 |
| 2 | | Tlogopojok | 7,858 | 6,592 | 26 | 23 |
| 3 | | Sidorukun | 5,182 | 3,021 | 3 | 3 |
| 4 | | Kramatinggil | 2,802 | 1,268 | 2 | 2 |
| 5 | | Karang | 5,397 | 4,120 | 31 | 39 |
| 6 | | Lumpur dsk | 19,526 | 11,102 | 59 | 55 |
| 7 | | Pulopancikan dsk | 7,761 | 5,191 | 8 | 7 |
| 8 | | Trate dsk | 10,974 | 6,986 | 32 | 24 |
| 9 | | Sukorame dsk | 7,687 | 5,144 | 41 | 34 |
| 10 | | Sidokumpul dsk | 18,053 | 11,614 | 159 | 141 |
| 11 | Kebomas | Dahanrejo | 5,533 | 4,210 | 11 | 9 |
| 12 | | Kembangan | 9,104 | 7,013 | 63 | 54 |
| 13 | | Kedanyang | 8,034 | 6,215 | 4 | 2 |
| 14 | | Indro | 7,440 | 3,377 | 4 | 2 |
| 15 | | Giri dsk | 14,047 | 6,405 | 80 | 63 |
| 16 | | Randuagung dsk | 24,245 | 17,375 | 251 | 218 |
| 17 | | Prambangan dsk | 7,205 | 2,774 | 17 | 11 |
| 18 | | Sidomoro | 9,676 | 5,682 | 113 | 79 |
| 19 | | Segoromadu dsk | 4,202 | 1,692 | 13 | 9 |
| 20 | | Singosari dsk | 14,206 | 9,209 | 58 | 54 |
| 21 | Manyar | Suci | 16,514 | 10,267 | 244 | 253 |
| 22 | | Yosowilangun | 12,712 | 7,590 | 151 | 105 |
| 23 | | Pongangan | 9,288 | 6,032 | 85 | 67 |
| 24 | | Romoo | 5,681 | 3,126 | 14 | 7 |
| 25 | | Sukomulyo | 9,842 | 5,260 | 53 | 36 |

### *B. Trip Production Model*

An SLZR Trip Production Model has been built by using HIS Data. The Independent Variable (IV) is the number of motorcycles (MC) in each zone, while the Dependent Variable (DV) is the TP for each zone. The SLZR TP Model is presented below. While the Regression Image is presented in Fig. 3.

$$TP^{MC}_i = 0.9075\ MC_i - 3.4412 \quad (9)$$

$$R^2 = 0.9713 \quad (10)$$

Where :

$TP^{MC}_i$ : morning peak motor cycle trip production value, zone i (*mc-trip/hr*)
$MC_i$ : motorcycle population, zone i (*motorcycle*)
$R^2$ : coefficient of determination

TP Predicted Values for the 25 zones have been calculated. The CIs were calculated only for two zones, one with the smallest and another with the biggest number of motorcycles. The calculation result is presented in Table II below. While the CI Image is presented in Fig. 4.

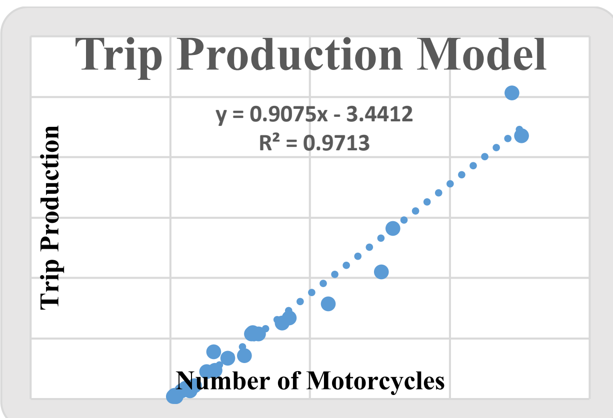


Fig. 3. Calculation of Simple Linear Zonal Regression TP Model

TABLE II
PREDICTED TP VALUES AND CI VALUES

| No | Zone | MC Population | Trip Production | Confidence Interval | |
|---|---|---|---|---|---|
| | | | | Min | Max |
| | | *motorcycle* | *MC-trip/hr* | *MC-trip/hr* | *MC-trip/hr* |
| 1 | Ngipik | 935 | 845 | -3,789 | 5,478 |
| 2 | Randuagung | 17375 | 15,756 | -1,664,238 | 1,695,749 |

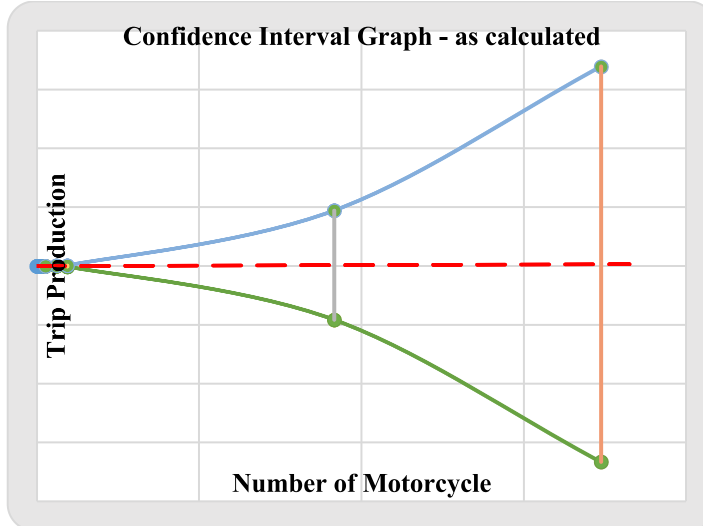


Fig. 4 Confidence Interval Image

*C. Confidence Interval (CI) Calculation for Analyse*

To investigate the CI Value Behavior due to IV Value Variation, six Zonal MC Values were taken for calculation. The calculation result is presented in Table III below.

*D. Analysis of Confidence Interval (CI) Behavior*

For easiness and clearness of CI Behavior investigation, the CI Values are measured in percentage while the IV Values are measured in Ratio between The IV Value and The IV Base Value. The IV Base Value is the Maximum MC Value in Regression Calculation, means the maximum MC Value gotten by HIS. The calculation is presented in Table IV.

TABLE III
CONFIDENCE INTERVAL CALCULATION RESULTS OF SIX IVS

| No | Zone | Trip Production | Confidence Interval | |
|---|---|---|---|---|
| | | | min | max |
| | | *trip/hr* | *trip/hr* | *trip/hr* |
| 1 | HIS-max | 224 | -453 | 902 |
| 2 | in between value | 534 | -2,161 | 3,230 |
| 3 | Ngipik | 845 | -3,789 | 5,478 |
| 4 | Sidorukun | 5,150 | -233,911 | 244,211 |
| 5 | in between value | 8,300 | -455,420 | 472,020 |
| 6 | Randuagung | 15,756 | -1,664,238 | 1,695,749 |

TABLE IV
CALCULATION OF CONFIDENCE INTERVAL AND RATIO

| No | Zone | Trip Production | Confidence interval | ⊥ CI | % CI | Ratio IV |
|---|---|---|---|---|---|---|
| | | trip/hr | trip/hr | trip/hr | % | |
| 1 | HIS-max | 224 | 902 | 678 | 302% | 1.00 |
| 2 | in between value | 534 | 3,230 | 2,696 | 504% | 2.38 |
| 3 | Ngipik | 845 | 5,478 | 4,633 | 549% | 3.77 |
| 4 | Sidorukun | 5,150 | 244,211 | 239,061 | 4642% | 22.97 |
| 5 | in between value | 8,300 | 472,020 | 463,720 | 5587% | 37.02 |
| 6 | Randuagung | 15,756 | 1,695,749 | 1,679,994 | 10663% | 70.27 |

The Ratio Values and CI Values are tabulated and draw graphically. It can be seen easily that bigger Ratio Value give bigger CI Value. It means that the HIS must be executed with a big number of HIS Samples for certain zone.

The Ratio Values and the corresponding CI Values are presented in Table V below. While the correlation graph is presented in Fig. 5 as follows.

TABLE V
IVV/$IVS_{MAX}$ RATIO VALUE AND CI VALUE (%)

| No | Ratio | Confidence Interval |
|---|---|---|
| | | % |
| 1 | 1.00 | 302% |
| 2 | 2.38 | 504% |
| 3 | 3.77 | 549% |
| 4 | 22.97 | 4642% |
| 5 | 37.02 | 5587% |
| 6 | 70.27 | 10663% |

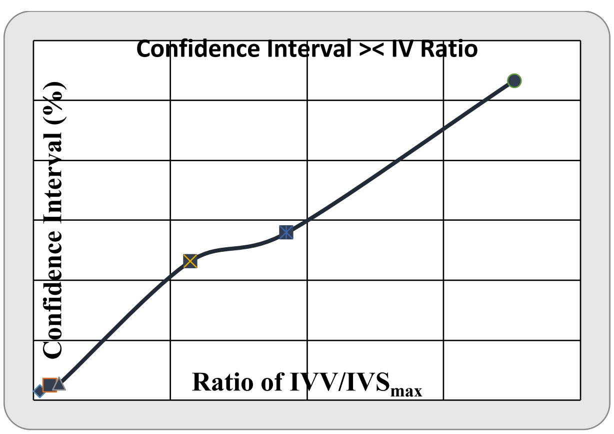


Fig. 5. The graph correlation between the Ratio Value and the CI Value

## V. CONCLUSIONS

The research has been finished satisfactorily, the research objective has been fulfilled. Main conclusions are written below.

- The CI is an important measure, it can give a picture of the Predicted Value (TP) accuracy.
- The resulted CI Line is conforming with the theoretical CI Line form, presented in Fig. 1.
- The higher IV Value produces bigger CI Value, thus less degree of accuracy.
- Getting the less CI Value needs smaller IV Ratio Value.
- The sampling size in term of Number of HIS Samples and Number of Observed Zones need to be investigated, based on further deeper experiment.
- The accepted CI Value need to be defined, based on further deeper experiment.

This research should be further developed by investigating the HIS Sample characteristics, by investigating the mathematical behavior of CI Formula for Simple Linear Regression, by investigating the CI Value Behavior on the regression line, by investigating the influence of sampling variation to $R^2$ and CI Values, by investigating the influence on the predicted value error caused by sampling variation, by investigating the CI Formula for other Regression Models. Briefly, achieving a sound conclusion, of how the sampling must be designed to reach a good SLZR TP accuracy, still needs several serious findings.

## ACKNOWLEDGEMENT

This paper is part of the main research to develop a good and accurate Transport Conventional Model Trip Distribution Modeling Method. The data were collected by Dio Hananda Ziantono and Jimi Aditya for their Magister's Thesis work.

## REFERENCES

[1] O.Z. Tamin, "*Perencanaan, Pemodelan dan Rekayasa Transportasi,*" Penerbit ITB, Bandung, 2008.

[2] H. Suprayitno, "Metoda Penilaian Kualitas Jaringan Jalan Utama di Wilayah Kabupten," Disertasi Doktorat, Jurusan Teknik Sipil, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, 2014.

[3] H. Suprayitno, "A Basic Framework for Regarding a Transportation System,*" Prosiding Seminar Nasional Teknik Sipil XII*, 2016.

[4] H. Suprayitno and V.A. Upa, "Special Conventional Transport Model for a New BRT Line Passenger Demand Prediction," *Journal of Technology and Social Science*, Vol.1, No. 3, pp. 10-18, 2017

[5] H. Suprayitno, "Metoda Pemodelan Angkutan Barang Perkotaan – Kasus Kotamadya Surabaya," *Simposium FSTPT II*, Surabaya, 1999.

[6] H. Suprayitno, "Penyusunan Metoda Perhitungan Model Distribusi Perjalanan Berbasis Data Volume Lalu Lintas pada Kasus Pembebanan All-or-Nothing," *Seminar Nasional Aplikasi Teknik Prasarana Wilayah*, Surabaya, 2015.

[7] N.D.A. Praditya, and H. Suprayitno, "Model Bangkitan Perjalanan Kota Samarinda Berbasis Zona Kecamatan*," Gema Aktualita*, Vol. 7, No. 1, (9), Juni 2018.

[8] M.S.D. Cahyono and H. Suprayitno, "Model Pertumbuhan dan Perbandingan Model Prediksi Penduduk Kota Surabaya Periode Tahun 1994-2013," *Prosiding Seminar Nasional Teknik Sipil XI,* Surabaya, *329-339*, 2015.

[9] H. Suprayitno, and V. Ratnasari, "Reflexion on linear regression trip production modelling method for ensuring good model quality," In *AIP Conference Proceedings*, vol. 1903, no. 1, p. 060013. AIP Publishing, 2017.

[10] L. Blank, *Statistical Procedures for Engineering, Management and Science*. McGraw-Hill International Book Company, pp. 487-528, 1982.

[11] N.R. Drapper and H. Smith, "*Applied Regression Analysis*," Third Edition, Wiley Seris in Probabilistics and Statistics, Canada, 1998.

[12] H. Suprayitno, "Investigating the Behavior of Predicted Value Confidence Interval due to Sample Variation for Linear Regression Trip Production Model," *Presented in ISTSDC 2017 Makassar and is still being reviewed to be accepted in IJASEIT,* 2017.