Sample Based Trip Length Distribution Quality based on χ^2 and Mean Absolute Error Values

Hitapriya Suprayitno[1], Vita Ratnasari[2], Nina Saraswati[1], Citto Pacama Fajrinia[1]

***Abstract*—As a vital part of transport modeling, the *trip length distribution* is normally gotten from a sample. In order to get a good model, a method to test the quality of the *sample based trip length distribution* must be available. A method to measure this quality has ever been developed, but it was found that the existing method still can be improved. While still using *goodness of fit statistical test*, the new method proposes two quality measurements. First, to verify whether the observed trip length distribution is conforming to the reference trip length distribution at a certain *confidence level value*, indicated by a value of $\chi^2 < \chi_0^2$. Second, to verify whether the error, *mean absolute error* measured in percentage, $|e\%| = 100 \times |e|_{mean}/x_{i,mean}$ is acceptable. The new method can be used, while still fulfilling the basic principle of sample quality measure, i.e. satisfying *maximum acceptable error at a certain confidence level*.**

Keywords—transport modeling, sample based trip length distribution quality measure.

I. Introduction

Transport Modeling is an important part of a Transportation Planning. Transportation Planning need a future estimate of desired line, traffic flow volume, passenger flow volume and goods flow volume in the network. This important transportation network loading data can only be provided by Transport Model. The desired line is a product of trip distribution modeling steps, in which Trip Length Distribution (TLD) is a capital input for the calculation. Transport model validation is generally done in the end of modeling after the trip assignment step. Therefore, having a good TLD is very important. Besides, TLD is sometimes used to validate the unconventional model [1]–[5].

Special Conventional Model has been recently developed to estimate a New Mass Transit Line in a limited existing network. The model cannot be calibrated and validated in the end after Trip Assignment calculation, since the existing passenger trips are not exist yet. Thus, having accurate samples in each step is a must. In this case having an accurate TLD is capital [6]–[9].

A sample can be considered as good if the sample's observed characteristics is the same as or very similar to the populations observed characteristics. Therefore, the sample quality measure is based on satisfying *maximum acceptable error at a certain confidence level* (MAE at CCL). This quality requirement, for certain cases, can be derived directly into a minimum sample size and its sampling method. For this purpose, in statistical science, several formulas have been developed to determine the minimum sample size to fulfill the sample quality requirement. However, formula to determine the minimum sample size for developing a TLD cannot be found [10]–[17].

There is a general formula to determine sample size for general cases in which the distribution is not known, often called as Slovin Formula. However, its explications, found on several papers or books, are confusing. In certain important detail, they are not harmonious each other [17]–[19]. The author, momentarily, decides that the Slovin Formula is better not to be used for this case, until a correct statistical explication that Slovin Formula can be used can be gotten.

The TLD is normally gotten from a sample. So, a means for measuring its quality must be available. A TLD Quality Measuring method has ever been developed and written [1]. It was found that an improvement still can be made on this method, in order to be better conforming to the general principle of the sample quality.

This paper presents the result of an attempt to improve the previous method for measuring the Sample based TLD quality. To be practical, in this paper, the previous paper cases are used to present the general TLD characteristics and to execute the method trial.

II. Research Method

The research was executed by following these steps: identification of TLD main characteristics, research problem statement, principal of sample quality statement, related statistical inference test, critics on the previous method, a new method concept, method trial, and conclusion.

III. New Method Development

A. TLD Main Characteristics

To be practical, the same example of TLD, used in previous paper, is used here to present the general characteristics of a TLD [1].

In general, a TLD follow a certain general characteristic, the number of long trip is less than the number of short trips, and the number of very short trips is also less. TLD's general distribution pattern, in terms of number of trips on a range of trip length, has a specify form, i.e. started from low number of trips, increase steeply to reach the peak value and then gradually decrease until it reach little number of trips at the

[1]Hitapriya Suprayitno, Nina Saraswati, and Citto Pacama Fajrinia are with Departement of Civil Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, 60111, Indonesia. E-mail: suprayitno.hita@gmail.com.

[2]Vita Ratnasari is with Departement of Statistics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, 60111, Indonesia. E-mail: vitaratna70@gmail.com.

maximum trip length. It must be noted that it deals with a distribution, it means a series of a parameter values, and not a single parameter value [1], [4], [5]. As an illustration, an Example of a TLD is presented in Table 1 and Figure 1.

B. Basic Sample Quality Measure and Sample Size

A sample can be considered as a good sample if and only if the sample's observed characteristics value is the same as or very similar to the populations observed characteristics value. To achieve this condition, two requirements must be accomplished: minimum sample size and the sample data collection method. Thus, quality measure for sample quality basically is "Maximum Acceptable Error at a Certain Confidence Level" (MAE at CCL). The error should be the error of characteristics value, so the statistical test must be used can be different from one case to the other. It depends very much on the nature of the observed characteristic [10].

Normally, a TLD is presented as a discrete curve, so it can be considered as a discrete distribution pattern. Therefore, statistical Goodness of Fit test should be appropriate to test the TLD Quality [1].

C. Goodness of Fit Test Principle

Goodness of Fit test is designated to verify whether an observed distribution is the same as the reference distribution. Depends on the distribution type, one of two statistical tests, i.e. the χ^2 test or the Kolmogorov-Smirnoff test, must be chosen to be used in Goodness of Fit test [10]. The TLD is a discrete curve, so the χ^2 test must be used [1]. The general hypothesis form is presented as follows.

H_0: the observed distribution is the same as the reference distribution

H_1: the observed distribution is not the same as the reference distribution

H_0 is accepted if $\chi^2 < \chi_0^2$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - R_i)^2}{R_i} \quad (1)$$

$$\upsilon = k - r - l \quad (2)$$

where:

χ^2 = calculated χ^2 value
χ_0^2 = the χ^2 value at υ and at certain confidence level
O_i = observed distribution value
R_i = reference distribution value
υ = degree of freedom
k = number of distribution value
r = number of parameter

D. Previous Method for Measuring Sample Based TLD Quality

A Method to Measure the TLD Sample Quality has been developed. The method is based on Goodness of Fit statistical inference by using a χ^2 Test. The Sample Quality is expressed by the Confidence Level [1].

Confidence Level = $P(\boldsymbol{\chi^2}, \boldsymbol{\upsilon})$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

$$\upsilon = k - r - l \quad (2)$$

where:

χ^2 = calculated χ^2 value
O_i = observed value
E_i = expected value
υ = degree of freedom
k = number of values of distribution
r = number of parameter

E. Critics on the Previous Method

The main critics on the previous method is based on the principle that the confidence level is normally used only to express the confidence level degree, and the error is used to express the accuracy degree. Only the combination of those two which is capable to well express the sample quality. The previous method, by using only the Goodness of Fit technique, measures only the Confidence Level without considering the existing error. This is not conforming to the principle of sample quality: MAE at CCL. The previous method, measure the confidence level value based on calculated χ^2 and its α value, can make the result a bit confusing. Basically, estimating at a higher confidence level value causes a wider range of estimation value, thus a higher error. Meanwhile, the Goodness of Fit test is not capable to calculate the error.

Therefore, the sample based TLD quality measurement, by still using the Goodness of Fit test technique, should include two results as follow.

- Conformity verification of whether the sample distribution is the same as the reference distribution.
- Acceptability verification of the error value.

F. Method Improvement

Developing a Method Improvement is basically implementing the sample based TLD quality measurement principle formulated in the above Critics on the Previous Method.

Before all, this is about comparing the distribution pattern between the sample distributions against the reference distribution. The real distribution values of those two are different significantly. So, comparing directly the two distributions, in terms of existing distribution values difference, is not correct. Hence, both sample's and reference's distribution values must be expressed in percentage.

The new method of Sample Based TLD Quality Measurement procedure can be presented as follow. First, set an accepted error value in percentage and a confidence level value. Second, check whether the sample distribution is conforming to the reference distribution. The verification is based on χ^2 test. Third, check whether the error value conform to the requirement. The verification is based on the mean absolute error value, measured in percentage. All of the mathematical formulas needed to conduct the quality measurement are presented below. The statistical test is presented below.

- To verify whether the sample distribution is part of the reference distribution.
  If $\chi^2 < \chi^2_{0,CL}$, the observed distribution is conform to the reference distribution at a certain CL.
- To verify whether the existing error is less than the requirement.
  If $|e\%|_{mean} < e_0|$, the sample quality is acceptable.

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - x_i)^2}{x_{0i}} \quad (4)$$

$$\upsilon = k - r - l \quad (2)$$

$$|e|_{mean} = \sum_{i=1}^{n} \frac{(|x_i - x_{0i}|)}{k} \quad (5)$$

$$x_{i,mean} = \sum_{i=1}^{n} \frac{x_i}{k} \quad (6)$$

$$|e\%|_{mean} = 100 \times \frac{|e|_{mean}}{x_{i,mean}} \quad (7)$$

where:

CL	: confidence level
χ_0^2	: required χ^2 value for the confidence level
χ^2	: case χ^2 value
x_{0i}	: sample distribution value
x_i	: reference distribution value
υ	: degree of freedom for χ^2 distribution
k	: number of distribution points
r	: number of distribution parameter
e_0	: accepted e value, as a requirement
$\|e\|_{mean}$	: mean of absolute error
$\|e\%\|_{mean}$	: mean absolute error, measured in percentage
$x_{i,mean}$	: mean of distribution value

IV. NEW METHOD TRIAL

A. Trial Case

To be practical, two cases used in the previous method paper are used in this paper as trial cases to examine the proposed method [1]. The two cases are as follows.

- BRI Kertajaya Office - working trip
- SMA 9 Wijaya Kusuma - schooling trip

The working trip length and the schooling trip length are the distance from employee home to the office and the distance from students home to the school. A number of employee's and student's addresses were noted. The distance from home to the cases addresses were measured. These can be considered as morning working and schooling trips.

For each case, a sample of minimum 50 trips were taken and considered as a reference, from which an 80% samples were taken randomly as the sample of the reference population. A number of 50s individus has been taken in order to have more than 30 individus. The number of 30 is considered as a number at which a certain group starts to have a clear statistical distribution pattern. In order that the two TLD, the reference and the sample, can be comparable, the TLD is measured in percentage rather than in number of trip. For these two trials, the Maximum Acceptable Error is set at 10% and the Confidence Level is set at 95%.

B. Trial Case 1 – BRI Kertajaya Office

BRI Kertajaya Office is a branch office of the Bank Rakyat Indonesia (BRI), a state owned bank. The office main data is as follow.

Name	: BRI Kertajaya Office
Status	: A branch office of a state owned bank.
Address	: Jl. Kertajaya 78, Surabaya
Number of Staff	: 78
Reference Sample	: 50
Sample of Reference	: 40 (80%)

Afterward, a Trip Length Distribution was constructed for the Reference Sample and for the Sample of Reference. The statistical Goodness of Fit test gave the following result.

- $\chi^2 < \chi^2_{0;[2,(1-95\%)]} \sim (1.376 < 5.99)$
  At a confidence level of 95%, the sample distribution can be considered as the same as the reference distribution.
- The mean absolute error is equal to 9% < 10%.
  The mean absolute error is less than the maximum accepted error. The sample is accepted.

The statistical test calculation and the distribution graphs are presented in Table 2 and Figure 2.

C. Trial Case 2 – SMA 9 Wijaya Kusuma

The SMA 9 Wijaya Kusuma is one of favorite High Schools in Surabaya. The high school main data are as follow:

Name	: SMA 9 Wijaya Kusuma
Status	: A state owned favorite High School
Address	: Jl. Wijaya Kusuma 48, Surabaya
Number of Student	: 879
Reference Sample	: 54
Sample of Reference	: 44 (81,5%)

Afterward, a Trip Length Distribution was constructed for the Reference Sample and for the Sample of Reference. The statistical test gave the following result.

- $\chi^2 < \chi^2_{0;[2,(1-95\%)]} \sim (1.6 < 11.07)$
  At a confidence level of 95%, the sample distribution can be considered as the same as the reference distribution.
- The mean absolute error is equal to 10.3% > 10%.
  The mean absolute error is slightly higher than the maximum accepted error. The sample is not accepted.

The statistical test calculation and the distribution graphs are presented in Table 3 and Figure 3.

D. Remarks

The previous method for measuring *sample based TLD* quality has been improved. It is clear that the new method can be used easily. The new method produce a conformity verification of the observed distribution to the reference distribution and acceptability verification of the error value measured in percentage. So the improved method is already conforming to the sample quality measurement principle: MAE at CCL. But by itself, the method is not capable to give a *minimum sample size formula* directly.

V. CONCLUSION

The research objective has been successfully attained. Main conclusion can be drawn and presented as follows.

- A new method, conform to the principle of sample quality: "Maximum Accepted Error at a Certain Confidence Level", has been developed.
- The new method is easy to be used.
- The Goodness of Fit statistical test is still used.

- The method produce two main components as result:
  - Conformity verification of the sample distribution relation to the reference distribution at a certain confidence level value.
  - Acceptability verification of the error, the error is expressed in mean absolute error, measured in percentage.
- Regarding the nature of the test method, a direct formula to determine the minimum sample size cannot be derived based on this method.

The paper discussions trigger other curiosities : how does the sample based TLD quality varies in relation to the sample size variation, is the Slovin formula appropriate to be used for this case, what is the range of problem the Slovin formula could cover, how is the behaviour of TLD Pattern due to the variation of Trip Length Interval used, what is the appropriate deterrence function, how is the detterence function varies due to the TLD Pattern variation, how is the appropriate method to develop deterrence function.

ACKNOWLEDGEMENT

This small research is a part of the main research on trying to find the Trip Length Distribution Quality Variation across Different Sample Size. The data are collected by Nina Saraswati and Citto Pacama Fajrinia as part of their thesis work.

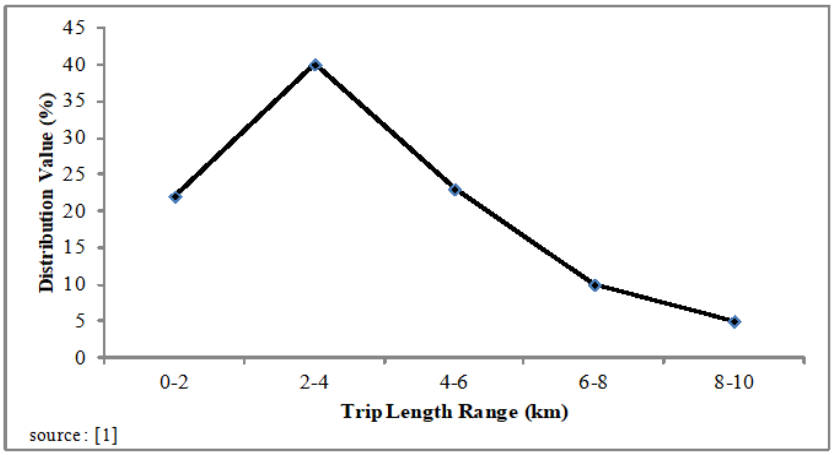

Figure 1. Example of a Trip Length Distribution Graph[1]

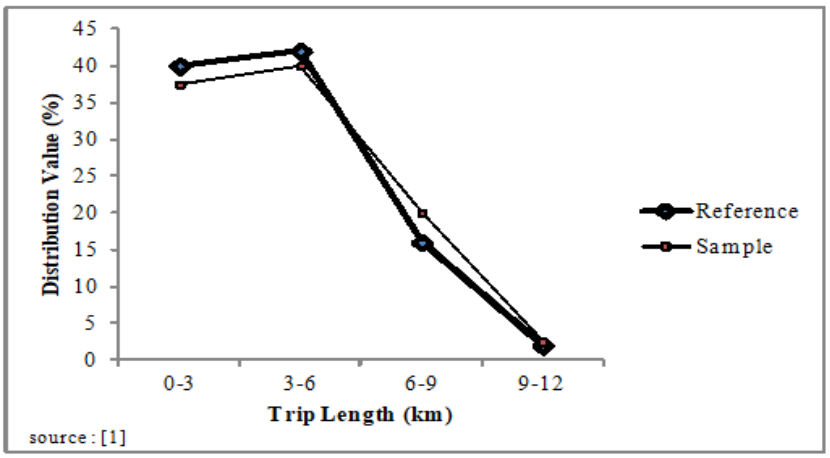

Figure 2. Case 1 – Trip Length Distribution Graph[1]

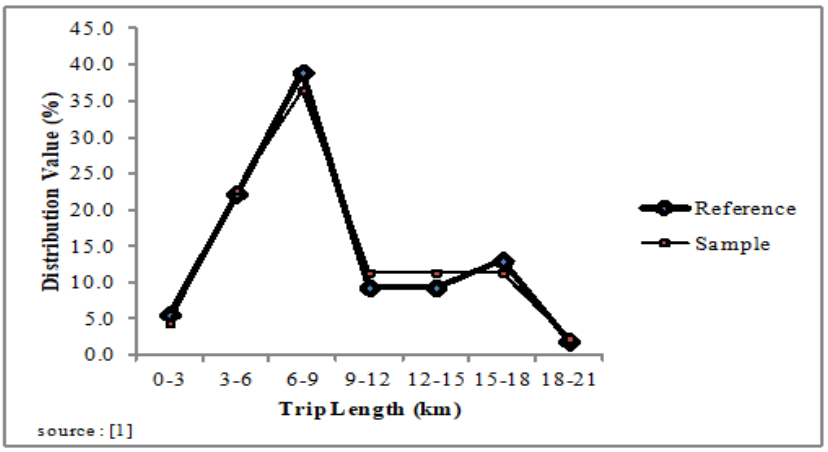

Figure 3. Case 2 – Trip Length Distribution Graph[1]

TABLE 1.
EXAMPLE OF A TRIP LENGTH DISTRIBUTION DATA[1]

No	Trip Length	Distribution	
	km	*trip*	*%*
1	0-2	28	23
2	2-4	48	40
3	4-6	26	22
4	6-8	12	10
5	8-10	6	5
Total		120	100

TABLE 2.
CASE 1 – SAMPLE QUALITY CALCULATION

No	Trip Length	Number		%		χ^2_i	$\lvert\bar{e}\rvert$
	km	**Reference**	**Sample**	**Reference**	**Sample**		
1	0-3	20	15	40.0	37.5	0.156	2.5
2	3-6	21	16	42.0	40.0	0.095	2.0
3	6-9	8	8	16.0	20.0	1.000	4.0
4	9-12	1	1	2.0	2.5	0.125	0.5
Total		50	40	100	100		9.0
	υ	2		**Total χ^2**		1.376	
	$\chi^2_{0(95\%)}$	5.99		**Mean Absolute Error (%)**			9%
	Notes	**CL**	95%				
		χ^2	$\chi^2 < \chi^2_0$	**the observed TLD = the reference TLD**			
		MAE(%)	9% < 10%	**the error is accepted**			

TABLE 3.
CASE 2 – SAMPLE QUALITY CALCULATION

No	Trip Length	Number		%		χ^2_i	$\lvert\bar{e}\rvert$
	km	**Reference**	**Sample**	**Reference**	**Sample**		
1	0-3	3	2	5.6	4.5	0.2	1.0
2	3-6	12	10	22.2	22.7	0.0	0.5
3	6-9	21	16	38.9	36.4	0.2	2.5
4	9-12	5	5	9.3	11.4	0.5	2.1
5	12-15	5	5	9.3	11.4	0.5	2.1
6	15-18	7	5	13.0	11.4	0.2	1.6
7	18-21	1	1	1.9	2.3	0.1	0.4
Total		54	44	100	100		10.3
	υ	5		**Total χ^2**		1.6	
	$\chi^2_{0(95\%)}$	11.07		**Mean Absolute Error (%)**			10.3%
	Notes	**CL**	95%				
		χ^2	$\chi^2 < \chi^2_0$	**the observed TLD = the reference TLD**			
		MAE(%)	10.3% > 10%	**the error is not accepted**			

REFERENCES

[1] H. Suprayitno, N. Saraswati, and C. P. Fajrinia, "Developing a method for measuring the quality of a sample based trip length distribution for urban trip," *Rekayasa Tek. Sipil - REKATS*, vol. 3, no. 3, pp. 252–258, 2016.

[2] H. Suprayitno, "Penyusunan Metoda Perhitungan Model Distribusi Perjalanan Berbasis Data Volume Lalu Lintas pada Kasus Pembebanan All-or-Nothing," in *Seminar Nasional Aplikasi Teknik Prasarana Wilayah*, 2015, p. D-181.

[3] H. Suprayitno, "Manual Validation and Calibration Method for All-or-Nothing Traffic Assignment," in *The 2nd Internasional Seminar on Science and Technology (ISST) 2016*, 2017, vol. 0, no. 2, pp. 29–36.

[4] O. Z. Tamin, *Perencanaan, pemodelan, dan rekayasa transportasi*. ITB, 2008.

[5] J. de D. Ortúzar and L. G. Willumsen, *Modelling transport*. Chichester: John Wiley & Sons Ltd, 2004.

[6] H. Suprayitno and V. Ananda Upa, "Mamminasata BRT User

Trip Characteristics for the Design of Demand Modelling Method for a New BRT Line," *IPTEK J. Technol. Sci.*, vol. 27, no. 3, pp. 47–52, Jan. 2017.

[7] H. Suprayitno and V. A. Upa, "Special Conventional Transport Model for a New BRT Line Passenger Demand Prediction (The General Modeling Method)," *J. Technol. Soc. Sci.*, vol. 1, no. 3, pp. 10–18, 2017.

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

[9] H. Suprayitno, V. Ratnasari, and N. Saraswati, "Experiment Design for Determining the Minimum Sample Size for Developing Sample Based Trip Length Distribution," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 267, no. 1, p. 012029.

[10] L. Blank, *Statistical procedures for engineering, management, and science*. New York: McGraw-Hill, 1982.

[11] E. Burmeister and L. M. Aitken, "Sample size: How many is enough?," *Aust. Crit. Care*, vol. 25, no. 4, pp. 271–274, Nov. 2012.

[12] D. A. Freedman, "Sampling," in *The SAGE Encyclopedia of Social Science Research Methods*, California: Sage Publication, Inc., 2004, pp. 986–989.

[13] B. Gerstman, "Sample Size, Precision, and Power." San Jose State University, San Jose, Washinton, 2003.

[14] V. Ratnasari, S. Sunaryo, and Setiawan, "Goodness of Fit pada Binary Logit dan Probit," *J. Mat. dan Ilmu Pengetah. Alam*, vol. 11, no. 2, pp. 31–35, 2008.

[15] S. Rose, N. Spinks, and A. I. Canhoto, *Management research : applying the principles*. Abingdon, Oxon: Rotledge, 2015.

[16] S. Siegel, *Nonparametric statistics : for the behavioral sciences*. Tokyo: McGraw-Hill Kogakusha, 1956.

[17] G. D. Israel, "Determining Sample Size," 1992.

[18] N. Mahmudah, "Pemodelan Bangkitan Perjalanan Pelajar di Kabupaten Sleman," *J. Tek. Sipil*, vol. 13, no. 4, p. 301, Feb. 2017.

[19] C. G. Sevilla, *Research methods*. Manila: Rex Book Store, 1992.