Determination of Optimal Unit Space Data for Taguchi's T-method based on Homogeneity of Output

Z. M. Marlan*, K. R. Jamaludin, F. Ramlie, N. Harudin, N. N. Jaafar

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur

zulkiflimarlahmarlan@gmail.com

Article history

Received:
18 Oct 2019

Received in revised form:
7 Nov 2019

Accepted:
4 Dec 2019

Published online:
25 Dec 2019

*Corresponding author:
zulkiflimarlahmarlan@gmail.com

Abstract

Taguchi's T-method is a prediction model introduced by Genichi Taguchi under the Mahalanobis-Taguchi System to determine the future state or unknown output based on existing or historical data. The prediction model was constructed using normalized signal data involving subtraction of average value of unit space data from signal data. The objective of this research is to determine a group of data having homogeneous characteristics from a densely populated region in a dataset to functioned as a basis for unit space data selection in T-method for predicting an accurate outcome. Histogram was utilized as a tool in representing data in multiple groups and a group with highest data frequency defined as unit space data. Nine different number of bins was used in assessing the effect of unit space data towards prediction accuracy. The result from the experiments on six different datasets indicates that no single number of bin fit for all in offering an optimal result. In addition, the size of unit space data and signal data do not significantly affect the final outcome. However, except for Auto MPG dataset, all different numbers of bin resulted in better prediction accuracy with less MSE and RMSE as compared to conventional T-method.

Keywords: *Taguchi's T-Method; Prediction Model; Histogram; Unit Space; Prediction Accuracy*

1. Introduction

Taguchi's T-method (T-method) is a regression-based predictive model introduced by Genichi Taguchi under Mahalanobis-Taguchi System (MTS) to predict the future state or unknown output based on existing or historical data. Similar to any regression-based model, T-method determined and explained the relationship between response variable and explanatory variable in the form of a mathematical model with the objective to produce an accurate prediction. The distinctive aspect of T-method as compared to other regression-based predictive models is the introduction of unit space concept for normalization and the used of signal-to-noise ratio (SNR) as variable's weightage and evaluation function. T-method is relatively new method and only few research works found in the literature mainly on the application, comparison, and enhancement of the method. Some of the recent work on the enhancement of T-method conducted by Harudin *et al.* [1] in adapting Shamos Bickel and Hodges Lehman estimator into T-Method for normalization, Suguru and Yasushi [2] on multiple output of T-method, Harudin *et.al* [3] on feature selection optimization using Artificial Bee Colony, Harudin *et*

* *Corresponding author: zulkiflimarlahmarlan@gmail.com*

al. [1] in applying robust M-estimator in increasing T-method accuracy, Nakao and Nagata [4] in an analysis involving missing data in T-method, and Nishino and Suzuki [5 on the introduction of median-median line for sample data with outlier in T-method. Earlier research by DasNeogi *et al.* [6], Cudney and Shah [7] and Cudney, Shah and Kestle [8] highlighted that the prediction accuracy of T-method predictive model was significantly affected by the selection of unit space data and suggested that a proper procedure developed specifically in determining unit space data. According to Teshima *et al.* [9], T-method defines the unit space data where the output value is in the medium position and homogeneous (densely populated).

Inoh *et al.* [10] introduced two different versions of unit space as alternatives to conventional T-method known as Ta-method and Tb-method where Ta-method and Tb-method are based on the average value of all output data and maximum SNR, respectively. Compared to Tb-method which involved high computational cost [4] Ta-method has been well accepted by researcher due to its ease of computation and replication. Negishi *et al.* [11], Nishino and Suzuki [5] and Matsushita *et al.* [12] used Ta-method unit space concepts to represent T-method instead of conventional unit space selection technique and used it for the analysis and comparison purposes between T-method and enhancement methods. In Ta-method, no unit space data was defined and all data was used as signal data to construct the predictive model. Normalization of signal data performed by subtracting the average value of each explanatory and response variable from signal data. Issues pertaining to unit space data in T-method lies to the lack of proper procedure in determining a subset of data to be used as unit space. Unlike Ta-method where selection criteria are fixed and dedicated to the average value, T-method existing procedure is subjective by user decision. Subjective in the sense of how the homogeneous data and the location of high dense populated region was defined in a dataset. In addition, upon identification of homogenous and dense populated data, how much data should be selected and whether it would yield optimal prediction accuracy resulted in more arguments. As a result, proper conclusion to the performance of prediction accuracy cannot be drawn.

The objective of this research is to determine a group of data having homogeneous characteristics from dense populated region in a dataset to be used as a basis for unit space data selection in T-method for predicting an accurate outcome. This research will utilize histogram as a tool to materialize the objective due to its specialization in grouping data and represent the shape of distribution, dispersion and central tendency of univariate data [13]. One of the important elements with enormous effect in constructing histogram is the number of bins selected, despite other elements such as the range, bin size and the starting point [14]. Dogan and Dogan [15] highlighted 23 different formulas in determining the number of histogram's bin. Realizing the importance of number of histogram's bin in the representation of data, this research will also assess the effects of different number of bin to T-method prediction accuracy.

2. Theoretical background

2.1. T-method computation procedure

The computational procedure of T-method as explained by Teshima *et al.* [9] involved three main phases in model development which are a preparation of data, development of mathematical model and evaluation of model as described below:

Phase 1: Sample data consists of *m* number of observation is gathered and organized, comprising of explanatory variables or item (x_{i1}, $x_{i2,}$ $x_{i3\ \ldots}$ x_{ik}: k is number of explanatory variable; i = 1, 2, 3… m) and response variable or output (y_i: i = 1, 2, 3… m). The sample data is then sorted ascending order based on response variable value. A subset of n number of sample data at medium position and homogeneous selected as unit space data and extracted out from sample data as shown in Table 1. The remaining l sample data unselected used as signal data as shown in Table 2. The average unit space data computed for all variables and normalization of signal data performed by subtracting the average value of unit space from signal data. Table 3 shows normalized signal data using equation (1) and (2).

Table 1: Unit space data

No	Item 1	Item 2	..	Item k	Output
1	x_{11}	x_{12}	..	x_{1k}	y_1
2	x_{21}	x_{22}	..	x_{2k}	y_2
…	…	…	..	…	…
n	x_{n1}	x_{n2}	..	x_{nk}	y_n
Average	$\bar{x}_1$	$\bar{x}_2$	..	$\bar{x}_k$	$\bar{y}=M_0$

Table 2: Signal data

No	Item 1	Item 2	..	Item k	Output
1	x'_{11}	x'_{12}	..	x'_{1k}	y'_1
2	x'_{21}	x'_{22}	..	x'_{2k}	y'_2
…	…	…	..	…	…
…	…	…	..	…	…
l	x'_{l1}	x'_{l2}	..	x'_{lk}	y'_l

$$X_{ij}=x'_{ij}-\bar{x}_j \qquad (i=1,2,\ldots,l),(j=1,2,\ldots,k) \tag{1}$$

$$M_i=y'_i-M_0 \qquad (i=1,2,\ldots,l) \tag{2}$$

Table 3: Normalized signal data

No.	Item 1	Item 2	…	Item k	Output value
1	X_{11}	X_{12}	…	X_{1k}	M_1
2	X_{21}	X_{22}	…	X_{2k}	M_2
…	…	…	…	…	…
l	X_{l1}	X_{l2}	…	X_{lk}	M_l

Phase 2: All computation in phase 2 performed using normalized signal data for model construction. Proportional coefficient, β and SNR, η computed using equation (3) and (4), respectively. In T-method, estimation of proportional coefficient, β is based on the ordinary least square method. Table 4 shows a summary of the Proportional coefficient, β and SNR, η for each of the explanatory variables. Integrated estimate value of output computed using equation (10) and the result compared with actual output value as shown in Table 5.

$$\text{Proportional Coefficient, } \beta_j=\frac{M_1X_{1j}+M_2X_{2j}+\ldots+M_lX_{lj}}{r} \quad ;(j=1,2,\ldots,k) \tag{3}$$

$$\text{SNR, } \eta_j=\begin{cases}\dfrac{\frac{1}{r}\left(S_{\beta j}-V_{ej}\right)}{V_{ej}} & \left(\text{when } S_{\beta j}>V_{ej}\right) \quad ;(j=1,2,\ldots,k)\\ =0 & \left(\text{when } S_{\beta j}\le V_{ej}\right)\end{cases} \tag{4}$$

$$\text{Effective Divider, } r=M_1^2+M_2^2+\ldots+M_l^2 \tag{5}$$

$$\text{Total Variation, } S_{Tj}=X_{11}^2+X_{21}^2+\ldots+X_{lj}^2 \tag{6}$$

$$\text{Variation of Proportional Term, } S_{\beta j} = \frac{\left(M_1X_{11}+M_2X_{21}+\ldots+M_lX_{lj}\right)^2}{r} \quad (7)$$

$$\text{Error Variation, } S_{ej} = S_{Tj}-S_{\beta j} \quad (8)$$

$$\text{Error Variance, } V_{ej} = \frac{S_{ej}}{l-1} \quad (9)$$

Table 4. Proportional coefficient, β and SNR, η for each item (explanatory variable)

No.	Item 1	Item 2	Item 3	Item 4	...	Item k
Proportional Coefficient, β	β_1	β_2	β_3	β_4	...	β_k
SNR, η	η_1	η_2	η_3	η_4	...	η_k

$$\text{Integrated Estimate Value, } \widehat{M}_i = \frac{\eta_1\times\frac{X_{i1}}{\beta_1}+\eta_2\times\frac{X_{i2}}{\beta_2}+\ldots+\eta_k\times\frac{X_{ik}}{\beta_k}}{\eta_1+\eta_2+\ldots+\eta_k} \quad ;(i=1,2,\ldots,l) \quad (10)$$

Table 5. Actual Output values and integrated estimate values of output

No.	Actual Output Value	Integrated estimate value
1	M_1	$\widehat{M}1$
2	M_2	$\widehat{M}_2$
…	…	…
l	M_l	$\widehat{M}_l$

Phase 3: The integrated estimate SNR computed using equation (11) used as a single index to assess the linear relationship between actual and estimated output value by taking into consideration three elements, which are sensitivity, slope and variability in evaluating a prediction model.

$$\text{Integrated Estimate SN Ratio, } \eta_{est} = 10\log\left(\frac{\frac{1}{r}\left(S_\beta-V_e\right)}{V_e}\right) \quad \text{(db)} \quad (11)$$

$$\text{Linear equation, } L = M_1\widehat{M}_1+M_2\widehat{M}_2+\ldots+M_l\widehat{M}_l \quad (12)$$

$$\text{Effective divider, } r = M_1^2+M_2^2+\ldots+M_l^2 \quad (13)$$

$$\text{Total variation, } S_T = \widehat{M}_1^2+\widehat{M}_2^2+\ldots+\widehat{M}_l^2 \quad (14)$$

$$\text{Variation of proportional term, } S_\beta = \frac{L^2}{r} \quad (15)$$

$$\text{Error variation, } S_e = S_T-S_\beta \quad (16)$$

$$\text{Error variance, } V_e = \frac{S_e}{l-1} \quad (17)$$

The pre-normalized integrated estimate value, $\hat{y}_i$ computed using equation (18) to acquire the predicted value without normalization.

$$\hat{y}_i=\widehat{M}_i+M_0 \quad ; (i=1,2,\ldots,l) \quad (18)$$

The estimation of the unknown output value from new obtained dataset requires for normalization of data by subtracting the average value of unit space from newly obtained signal data. The average of unit space data referring to average value obtained from training data. Equation (10) used to determine the integrated estimate value of new normalized signal data. Equation (18) used to compute the new integrated estimate value before normalization.

2.2. Unit space concept

The concept of unit space originated by Genichi Taguchi dated back during the development of the Mahalanobis-Taguchi method (MT method) where unit space is defined as homogeneous group with respect to others. In MT method, unit space is defined as a homogeneous group or population used as a reference point in developing a measurement system. The understanding of a homogeneous group or population also refers to a normal state, a normal group, a state of high density, a state of constant and extended to a state that associated with high frequency or an average. Genichi Taguchi later discovered that in recognition problem, the accuracy of recognition is improved if the unit space is set in the medium position. This has become a basis of unit space definition for T-method where unit space is defined as a subset of data within the sample of observation that located at medium position on top of having homogeneity characteristic and comes from dense region criterion. Figure 1 illustrates the concept of unit space in T-method. Unit space data in T-method is used to normalized signal data before the construction of the mathematical model as explained in section 2.1.

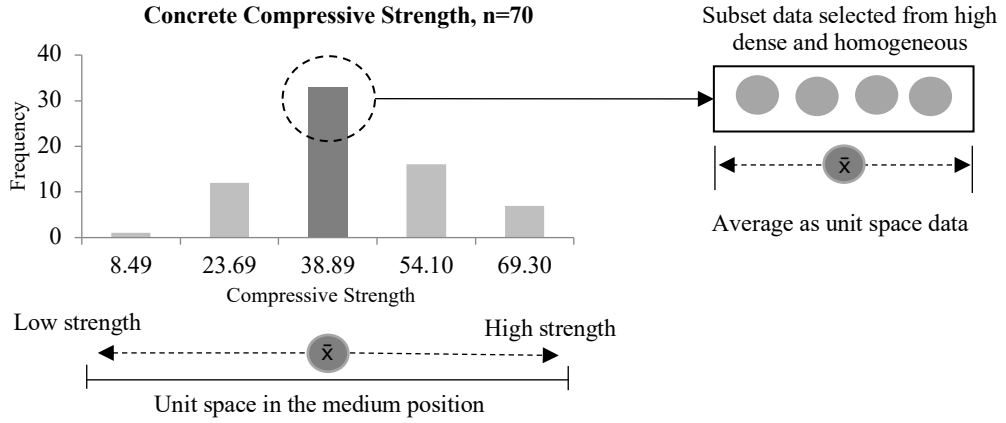

Figure 1: Unit space concept in T-method

2.3. Unit space variants

Inoh *et al.* [10] introduced two versions of unit space which are Ta-method and Tb-method. Ta-method uses an average value of each item and output for normalization of signal data as shown in equation (19) until (22). In Ta-method, signal data consists of all sample data without defining and discarding unit space data. The average value calculated using all sample data in each item and output. Normalization procedure similar to in T-method by subtracting the average value of each item and output value from signal data. The remaining procedure in estimating integrated estimate value is similar to T-method described in section 2.1.

$$\bar{\mathrm{x}}_j = \frac{1}{m}(\mathrm{x}_{1j} + \mathrm{x}_{2j} + \cdots + \mathrm{x}_{mj}); \qquad (i = 1,2,\dots,k) \tag{19}$$

$$\mathrm{M}_0 = \frac{1}{m}(y_1 + y_2 + \cdots + y_m) \tag{20}$$

$$\mathrm{X}_{\mathrm{ij}} = x_{ij} - \bar{x}_j \qquad (i = 1,2,\dots,m), (j = 1,2,\dots,k) \tag{21}$$

$$\mathrm{M}_{\mathrm{i}} = y_i - M_0 \qquad (i = 1,2,\dots,m) \tag{22}$$

Tb-method uses a value in each item and output that maximize SNR for normalization of signal data. Similar to Ta-method, signal data in Tb-method consists of all sample data without defining and discarding unit space data. Start by

considering the first item and the output value, normalized signal data performed using the first sample data of item and output to calculate SNR. The computation repeated for the first item and its output by normalizing the signal data using the second sample data to calculate SNR. In the end, sample data that maximize SNR for the first item selected for final normalization in estimating integrated estimate value. The procedure repeated for each item. The remaining procedure in estimating integrated estimate value similar to T-method as explained in section 2.1.

2.4. Histogram

The term histogram was first introduced by a famous statistician named Karl Pearson back in 1891 which refers to a common form of graphical representation [16]. A histogram is obtained by dividing the range of the univariate data into equal size of bin to display the distribution of data in the vertical bar form. This can be done by first identifying the width of bin or specifying the number of histogram's bin before the frequency of data lie in each bin counted. Being one of the oldest tools for graphical display of data, histogram continued to be relevant and opted by many researchers from various field in analyzing of data. For instance, Tan *et al.* [17] use histogram to investigation of static analog-to-digital converter nonlinearity measurement and Aflaifel *et al.* [18] uses histogram to assess the efficacy of uterotonic treatment for post-partum haemorrhage. One of the prevailing concerns in constructing a histogram is the suitable number of bin. Too many bins will result in a jagged histogram while too little of bin resulted in a loss of valuable information due to a single block histogram formed. He and Meeden [19] introduced a loss function that incorporates the concepts of rougher densities requires more bin than smooth densities using stepwise Bayes procedure based on Bayesian bootstrap in determining number of histogram's bin. Lahoka [14] highlighted recommended number of bin from various researchers which claimed to be optimal and few formulas to determine the number of bin. Dogan and Dogan [15] compiled 23 formulas developed by past researchers on how to determine the number of bin.

3. Methodology

3.1. Experiment data

The data used in this research was obtained from UC Irvine Machine Learning Repository [20] involving six different datasets from various fields which are airfoil self-noise, auto mpg, concrete compressive strength, energy efficiency (cooling load), energy efficiency (heating load) and yacht hydrodynamics as shown in Table 6. Each dataset contains 70 training data for model construction and 30 testing data for validation, randomly selected. This research focused on predicting the value of univariate response variable involving multiple explanatory variables.

Table 6. Details of experiment's datasets

No	Datasets (p)	Explanatory Variable (x_k)	Response Variable (y)
1	Airfoil Self Noise	Frequency (Hz), Angle of Attack (^{0}), Chord Length (m), Free-Stream Velocity (m/s), Suction Side Displacement Thickness (m)	Scaled Sound Pressure Level (dB)
2	Auto MPG	Cylinder, Displacement, Horsepower, Weight, Acceleration	Miles per Gallon (MPG)

3	Concrete Compressive Strength	Cement (kg/m^3), Blast Furnace Slag (kg/m^3), Fly Ash (kg/m^3), Water (kg/m^3), Superplasticizer (kg/m^3), Coarse Aggregate (kg/m^3), Fine Aggregate (kg/m^3), Age (day)	Concrete Compressive Strength (MPa)
4	Energy Efficiency	Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, Glazing Area Distribution	Cooling Load
5	Energy Efficiency	Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, Glazing Area Distribution	Heating Load
6	Yacht Hydrodynamics	Longitudinal Position of the Center of Buoyancy, Prismatic Coefficient, Length-Displacement Ratio, Beam-Draught Ratio, Length-Beam Ratio, Froude Number	Residuary Resistance per Unit Weight of Displacement

3.2. Bin estimation

In determining unit space data in T-method, a group of data having homogeneous characteristic located in a dense populated region need to be identified before a subset of data can be selected as unit space data from that respective group. This research utilized a histogram as a tool to determine a group of homogeneous data and dense populated in a dataset and defined unit space data as data within highest frequency bin. Nine cases denoting nine different numbers of bins derived from various formulas compiled by Dogan and Dogan [15] were used in constructing the histogram to determine a group of data from a homogeneous and dense populated region in dataset. Table 7 shows nine cases with different number of bins based on 70 training data derived from various formulas. The number of bins computed is a rounded up basis to the nearest integer whole number.

Table 7. Formulas and estimated number of bins [15]

No.	Case (*q*)	Number of Bin	Formula
1	a	4	Cohran = $\sqrt{n/5}$
2	b	5	Cencov = $\sqrt[3]{n}$
3	c	6	Larson = $1+\lceil 2.2\times\log_{10}(n)\rceil$
			Terrel and Scott = $\sqrt[3]{2n}$
4	d	7	Anonymous2 = $2^k \geq n$
5	e	8	Sturges = $1+\lceil 3.3\times\log_{10}(n)\rceil$
			Ishikawa = $6+(n/50)$
			Anonymous1 = $2.5\times\sqrt[4]{n}$
6	f	9	Rice = $2\times\sqrt[3]{n}$
7	g	10	Mosteller and Tukey = $10\times\log_{10}\sqrt{n}$
8	h	11	Bendat and Piersol = $1.87\times(n-1)^{0.4}$
9	i	17	Velleman = $2\times\sqrt{n}$

3.3. Experiment design

In this research, for each dataset, nine experiments conducted by differentiating the number of bin in constructing a histogram using response variable value as illustrated in Figure 2. In every experiment, the bin with the highest frequency of response variable data selected as a basis for unit space data selection for explanatory variables. The minimum and maximum value in the respective bin constitute a range of unit space data. The average value of unit space data used for normalization of signal data by subtracting the average value from signal data for each explanatory and response variable. In these experiments, unit space is

discarded from signal data for model construction as in conventional T-method. The result obtained from the experiments will be compared to result of T-method and Ta-method. The result from T-method is based on 5 samples of unit space data obtained from the medium position of response variable.

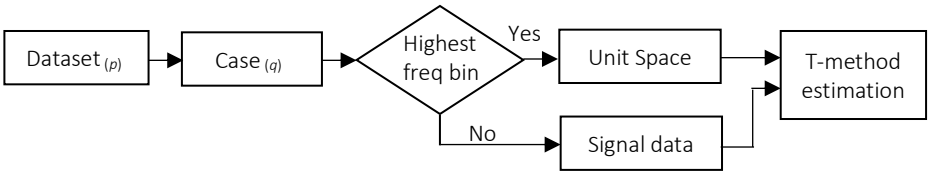

Figure 2. Framework of experiment design

3.4. Performance criteria

The performance of T-method prediction accuracy evaluated using Mean Square Error (MSE) and Root Mean Square Error (RMSE) as the error metrics and Coefficient of Determination (R^2) as a measure of goodness-of-fit. The calculation of the error metrics and goodness-of-fit as shown in equation (23), (24) and (25). The prediction model constructed using training data assessed using both error metrics and measure of goodness-of-fit while the predicted value using testing data evaluated using error metrics only.

$$\text{Mean Square Error, MSE} = \frac{1}{l}\sum_{i=1}^{l}(M_i-\hat{M}_i)^2 \quad (23)$$

$$\text{Root Mean Square Error, RMSE} = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(M_i-\hat{M}_i)^2} \quad (24)$$

$$\text{Coefficient of Determination, } R^2 = 1-\frac{\sum_{i=1}^{l}(M_i-\hat{M}_i)^2}{\sum_{i=1}^{l}(M_i-\bar{M}_i)^2} \quad (25)$$

4. Result and discussion

4.1. Effect of different number of histogram's bin

The experiments using a different number of histogram's bin resulted to a different number of unit space data selected as the range of unit space data changed with respect to the highest frequency bin. A small number of unit space data selected as a result of more number of histogram's bin used resulted in more number of signal data left for model construction. However, the result from the experiments indicates that the size of unit space data and signal data is not a significant factor in achieving the optimal prediction accuracy. The ability to group a homogeneous and dense populated data has more influent to the final prediction outcome.

4.2. Analysis of case study

4.2.1 Airfoil noise dataset

The result in Table 8 shows the prediction performance based on train and test data in predicting the scaled sound pressure level using airfoil noise dataset. The R squared value for case (a), (c) and (d) shows that the model constructed using train data fit better as compared to conventional T-method by 7%, 5% and 10%, respectively. Case (c) recorded the lowest of MSE and RMSE for test data despite case (d) recorded the lowest for train data. All cases surpassed the conventional T-method prediction's accuracy for the test data in terms of MSE and RMSE at a range of 15% to 56% and 8% to 34% improvement, respectively. However, as compared to Ta-method, only case (a), (c), (e), (f), (h) and (i) recorded a better result with less error.

Table 8. Train and test result for airfoil noise dataset

Train:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
No. of Unit space Data	22	22	17	16	16	14	12	14	11	5	1
No. of Signal Data	48	48	53	54	54	56	58	56	59	65	70
R squared	69.88%	57.78%	68.64%	71.79%	58.80%	59.12%	53.36%	64.95%	63.12%	65.38%	65.23%
MSE	37.01	50.43	39.96	30.26	32.76	56.02	61.73	43.98	44.20	53.50	33.55
RMSE	6.08	7.10	6.32	5.50	5.72	7.48	7.86	6.63	6.65	7.31	5.79
Test:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
R squared	38.74%	36.60%	35.48%	37.60%	34.90%	38.65%	38.10%	38.90%	40.33%	34.13%	37.91%
MSE	65.12	106.79	54.95	90.03	65.26	57.83	80.98	56.23	72.43	126.01	77.30
RMSE	8.07	10.33	7.41	9.49	8.08	7.60	9.00	7.50	8.51	11.23	8.79

4.2.2 Auto mpg dataset

The result in Table 9 shows the prediction performance based on train and test data in predicting the vehicle's fuel consumption in miles per gallon using auto-mpg dataset. The R squared value for case (a), (c), (d), (e), (f), (g), (h) and (i) shows that the model constructed using train data fit better as compared to conventional T-method and Ta-method. Case (b) recorded the lowest of MSE and RMSE for both test data and train data. As compared to conventional T-method for test data, only case (b), (f) and (g) recorded a better MSE and RMSE result by 33%, 21%, 21% and 18%, 11%, 11% improvement, respectively. Overall, case (b) performed better than both conventional T-method and Ta-method for test data at 33% and 11% MSE improvement and 18% and 6% RMSE improvement, respectively.

Table 9. Train and test result for auto MPG dataset

Train:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
No. of Unit space Data	26	22	19	20	17	15	15	11	11	5	1
No. of Signal Data	44	48	51	50	53	55	55	59	59	65	70
R squared	77.51%	44.37%	73.97%	71.19%	74.85%	67.88%	67.88%	71.66%	71.37%	67.81%	67.61%
MSE	45.74	20.95	45.98	38.72	50.68	24.31	24.31	40.10	46.91	30.25	23.49
RMSE	6.76	4.58	6.78	6.22	7.12	4.93	4.93	6.33	6.85	5.50	4.85
Test:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
R squared	70.32%	70.23%	69.56%	69.25%	70.11%	69.60%	69.60%	70.64%	69.74%	70.67%	69.99%
MSE	39.65	25.11	45.65	38.89	49.26	29.24	29.24	42.86	48.79	37.24	28.30
RMSE	6.30	5.01	6.76	6.24	7.02	5.41	5.41	6.55	6.98	6.10	5.32

4.2.3. Concrete compressive strength dataset

The result in Table 10 shows the prediction performance based on train and test data in predicting the concrete compressive strength using concrete compressive strength dataset. The R squared value for all cases except case (c) shows that the model constructed using train data fit better than the conventional T-method and Ta-method. Case (i) recorded a better performance in terms of MSE and RMSE for both train and test data and surpassed the performance recorded using conventional T-method by 41% and 23% improvement, respectively and Ta-method by 11% and 6% improvement, respectively. In general, all cases performed better with less MSE and RMSE recorded for test data as compared to T-method at a range of 2% to 41% and 1% to 23% improvement, respectively but only case (e), (h) and (i) recorded better result as compared to Ta-Method by 6%, 3% and 14% less MSE and 3%, 1% and 7% less RMSE, respectively.

Table 10. Train and test result for concrete compressive strength dataset

Train:

Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
No. of Unit space Data	33	22	20	21	19	17	14	15	11	5	1
No. of Signal Data	37	48	50	49	51	53	56	55	59	65	70
R squared	51.63%	52.04%	45.37%	48.30%	53.90%	50.02%	49.70%	56.53%	52.72%	46.24%	46.51%
MSE	319.34	285.17	354.22	282.89	230.35	286.01	284.88	239.92	210.01	280.68	247.58
RMSE	17.87	16.89	18.82	16.82	15.18	16.91	16.88	15.49	14.49	16.75	15.73

Test:

Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
R squared	63.31%	67.59%	59.49%	63.84%	70.44%	59.88%	66.12%	70.17%	71.69%	66.02%	67.40%
MSE	266.85	242.58	313.38	280.01	202.78	278.73	268.86	210.65	186.93	318.21	216.22
RMSE	16.34	15.58	17.70	16.73	14.24	16.70	16.40	14.51	13.67	17.84	14.70

4.2.4. Energy efficiency (cooling load) dataset

The result in Table 11 shows the prediction performance based on train and test data in predicting the cooling load using energy efficiency dataset. The R squared value recorded for all cases shows that the model constructed using train data fit weaker than the conventional T-method and Ta-method. Case (d) recorded the lowest of MSE and RMSE for test data despite case (g) recorded the lowest for train data. All cases recorded a better MSE and RMSE performance for test data as compared to conventional T-method at a range of 3% to 21% improvement and 1% to 11% improvement, respectively but only case (a), (b), (c), (d), (e), (f) and (i) better than Ta-Method for test data.

Table 11. Train and test result for energy efficiency (cooling load) dataset

Train:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
No. of Unit space Data	33	31	30	26	22	21	15	17	16	5	1
No. of Signal Data	37	39	40	44	48	49	55	53	54	65	70
R squared	22.87%	39.27%	43.58%	55.69%	62.64%	63.80%	68.90%	67.41%	68.17%	76.01%	75.83%
MSE	45.16	44.35	43.43	39.98	37.81	37.23	34.41	35.61	34.63	36.30	28.37
RMSE	6.72	6.66	6.59	6.32	6.15	6.10	5.87	5.97	5.88	6.03	5.33
Test:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
R squared	80.78%	81.23%	81.69%	82.41%	82.48%	82.55%	78.31%	78.22%	80.74%	80.71%	80.46%
MSE	23.13	22.98	22.61	22.26	22.76	22.81	27.52	27.50	24.59	28.33	25.00
RMSE	4.81	4.79	4.75	4.72	4.77	4.78	5.25	5.24	4.96	5.32	5.00

4.2.5. Energy efficiency (heating load) dataset

The result in Table 12 shows the prediction performance based on train and test data in predicting the heating load using energy efficiency dataset. Most of the cases except case (c) recorded lesser R squared value as compared to conventional T-method indicating weaker fit of actual value of training data to prediction model. Case (a) recorded the lowest of MSE and RMSE for test data despite case (d) recorded the lowest for train data. All cases recorded a better prediction performance in term of MSE and RMSE for test data as compared to conventional T-method at a range of `44% to 50% and 25% to 29% improvement, respectively and Ta-method at a range of 2% to 12% and 1% to 6% improvement, respectively.

Table 12. Train and test result for energy efficiency (heating load) dataset

Train:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
No. of Unit space Data	34	23	22	23	25	22	22	22	15	5	1
No. of Signal Data	36	47	48	47	45	48	48	48	55	65	70
R squared	38.25%	68.18%	78.81%	76.66%	66.63%	69.34%	69.27%	69.27%	73.38%	78.81%	78.92%
MSE	41.27	34.41	30.38	29.00	39.13	34.83	33.92	33.92	32.03	36.19	27.37
RMSE	6.42	5.87	5.51	5.39	6.26	5.90	5.82	5.82	5.66	6.02	5.23
Test:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
R squared	82.14%	82.28%	79.71%	80.65%	80.64%	81.62%	82.21%	82.21%	81.00%	81.09%	81.42%
MSE	19.24	20.48	21.37	20.83	20.31	20.00	20.24	20.24	21.12	38.30	21.85
RMSE	4.39	4.53	4.62	4.56	4.51	4.47	4.50	4.50	4.60	6.19	4.67

4.2.6. Yacht hydrodynamics dataset

The result in Table 13 shows the R squared value for all cases which constructed using train data fit better than the conventional T-method and Ta-method. Case (i) recorded the lowest of MSE and RMSE for the test data although case (a) recorded the lowest for train data. All cases recorded a better performance in term of MSE and RMSE as compared to result obtained through conventional T-method for test data at a range of 52% to 66% and 31% to 41% improvement, respectively. However, only case (e), (f), (g), (h) and (i) shows better MSE and RMSE result as

compared to Ta-method by 5%, 7%, 8%, 10%, 15% and 2%, 35, 4%, 5% 8% improvement, respectively.

Table 13. Train and test result for yacht hydrodynamics dataset

Train:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
No. of Unit space Data	56	54	53	51	49	48	47	46	41	5	1
No. of Signal Data	14	16	17	19	21	22	23	24	29	65	70
R squared	96.78%	97.44%	96.44%	96.20%	96.44%	94.95%	94.20%	93.87%	90.36%	66.40%	66.51%
MSE	112.99	136.49	143.16	158.25	173.92	175.81	178.55	182.34	192.75	418.43	125.12
RMSE	10.63	11.68	11.96	12.58	13.19	13.26	13.36	13.50	13.88	20.46	11.19
Test:											
Case	a	b	c	d	e	f	g	h	i	T-method	Ta-method
R squared	65.28%	65.28%	65.28%	65.28%	65.28%	65.28%	65.28%	65.28%	65.28%	65.28%	65.28%
MSE	210.62	194.38	188.99	178.89	169.67	166.54	163.56	160.78	150.88	440.57	178.35
RMSE	14.51	13.94	13.75	13.37	13.03	12.91	12.79	12.68	12.28	20.99	13.35

5. Conclusion

In this research, a histogram's bin with highest frequency of data was defined as a group of homogeneous with dense population and the width of the respective bin constitutes a range for unit space data selection. The result from nine experiments referring to nine cases of different number of histogram's bin for six different datasets indicates that for different datasets having different data's pattern and distribution, no single number of histogram's bin fit to all datasets in offering an optimal prediction accuracy. Except for Auto MPG dataset, all nine cases in all datasets recorded better prediction accuracy for test data as compared to conventional T-method. Table 14 summarized the result of the experiments where bold font indicates improvement over Ta-method. Result of case (b) shows enhancement in all dataset as compared to T-method while case (e) and (i) shows improvement over T-method and Ta-method except for auto mpg dataset. This research concludes that histogram is a useful tool in determining a group of homogeneous data in dense populated region to be used as unit space data in T-method. In addition, a selection of number of histogram's bin is imperative to yield optimal prediction accuracy. Last but not least, histogram through bin's width offers an interval or range of data to be used as a guideline for the selection of unit space data from overall sample data.

In future research, the effect of prediction accuracy based on the inclusiveness of unit space data determined using histogram's bin classification into signal data to be assessed and analyzed. The concept of unit space in conventional T-method is to separate and discard the unit space data from signal data for model construction and prediction. Contradicted to Ta-method and Tb-method which includes all data in model formulation and prediction. The inclusiveness of unit space data is important in making a prediction, especially when dealing with small sample size in dataset.

Table 14. Summary of test result for all datasets with improvement

Dataset	Case	a	b	c	d	e	f	g	h	i	T Method-1	**Ta-method**
Airfoil Self Noise	MSE	**65.12**	106.79	**54.95**	90.03	**65.26**	**57.83**	80.98	**56.23**	**72.43**	126.01	**77.30**
	RMSE	**8.07**	10.33	**7.41**	9.49	**8.08**	**7.60**	9.00	**7.50**	**8.51**	11.23	**8.79**
Auto MPG	MSE		**25.11**				29.24	29.24			37.24	**28.30**
	RMSE		**5.01**				5.41	5.41			6.10	**5.32**
Concrete Comp Strength	MSE	266.85	242.58	313.38	280.01	**202.78**	278.73	268.86	**210.65**	**186.93**	318.21	**216.22**
	RMSE	16.34	15.58	17.70	16.73	**14.24**	16.70	16.40	**14.51**	**13.67**	17.84	**14.70**
Energy Efficiency (Cooling)	MSE	**23.13**	**22.98**	**22.61**	**22.26**	**22.76**	**22.81**	27.52	27.50	**24.59**	28.33	**25.00**
	RMSE	**4.81**	**4.79**	**4.75**	**4.72**	**4.77**	**4.78**	5.25	5.24	**4.96**	5.32	**5.00**
Energy Efficiency (Heating)	MSE	**19.24**	**20.48**	**21.37**	**20.83**	**20.31**	**20.00**	**20.24**	**20.24**	**21.12**	38.30	**21.85**
	RMSE	**4.39**	**4.53**	**4.62**	**4.56**	**4.51**	**4.47**	**4.50**	**4.50**	**4.60**	6.19	**4.67**
Yacht Hydrodynamics	MSE	210.62	194.38	188.99	178.89	**169.67**	**166.54**	**163.56**	**160.78**	**150.88**	440.57	**178.35**
	RMSE	14.51	13.94	13.75	13.37	**13.03**	**12.91**	**12.79**	**12.68**	**12.28**	20.99	**13.35**

Acknowledgement

This work was supported by the Potential Academic Staff (PAS) Grant Scheme awarded by the Universiti Teknologi Malaysia (Grant No. Q.K130000.2756.03K33).

6. References

[1] Harudin N, Jamaludin K R, Muhtazaruddin M Nabil, Ramlie F, Wan Muhamad Wan Zuki Azman, "A Feasibility Study in Adapting Shamos Bickel and Hodges Lehman Estimator into T-Method for Normalization", IOP Conference Series: Materials Science and Engineering, Vol. 319, (2018), pp.1-7

[2] Suguru Sekine, Yasushi Nagata, "A Study of Multivariate Taguchi's T Method", Asian Network for Quality, (2018), pp.250-262

[3] Harudin N, Jamaludin, K R, Muhtazaruddin M Nabil, Ramlie F, Wan Muhamad Wan Zuki Azman, Jaafar NN, "Artificial Bee Colony for Features Selection Optimization in Increasing T-Method Accuracy", International Journal of Engineering & Technology, Vol. 7, No. 4.35, (2018), pp. 885-891

[4] Nakao Yuto, Nagata Yasushi, "Analysis of Data including missing values in the Taguchi's T Method", Total Quality Science, Vol.4, No.2, (2018), pp: 53-6

[5] Nishino Keisuke, Suzuki Arata, "Taguchi's T-Method using Median-Median Line for Small Sample with Outliers", IEEJ Transactions on Industry Applications, Vol.138, No.7, (2018), pp:598–604.

[6] DasNeogi Protyusha, Cudney Elizabeth A., "Comparing the Predictive Ability of T-Method and Cobb-Douglas Production Function for Warranty Data", Proceedings of the ASME 2009 International Mechanical Engineering Congress & Exposition, (2009), pp:1-6, https://doi.org/10.1115/IMECE2009-12668

[7] Cudney Elizabeth A., Shah Parthiv, "Predicting Annual Precipitation Using the T-Method", Proceeding of the 2010 Industrial reearch Conference, (2010)

[8] Cudney Elizabeth A., Shah Parthiv A., Kestle Rodney, "Predicting Vehicle Cost using the T-Method", International Journal of Product Development, Vol.12, No.3/4, pp.311-323, https://doi.org/10.1504/IJPD.2010.036393

[9] Teshima Shoichi, Hasegawa Yoshiko, Tatebayashi Kazuo, Quality Recognition and Prediction: Smarter Pattern Technology with the Mahalanobis-Taguchi System, Momentum Press, (2012)

[10] Inoh Junki, Nagata Yasushi, Horita Keisuke, Mori Arisa, "Prediction Accuracies of lmproved Taguchi's T Methods Compared to those of Multiple Regression Analysis", Journal of the Japanese Society for Quality Control, Vol.42, No.2, (2012), pp.103-115

[11] Negishi Shintaro, Morimoto Yusuke, Takayama Satoshi, Ishigame Atsushi, "Daily Peak Load Forecasting by Taguchi's T Method", Electrical Engineering in Japan, Vol. 201, No. 1, (2017), pp. 57-65

[12] Matsushita Makoto, Ogino Kohtaroh, Morioka Hiroki, "Improving the Accuracy of T-Method-based Item Selection by Incorporating AI (GA)", Paper presented to International Conference on Robust Quality Engineering, (2018), Kuala Lumpur, August 1-5, Unpublished

[13] Fowlkes William Y., Creveling Clyde M., Engineering Methods for Robust Product Design: Using Taguchi Methods in Technology and Product Development, Addison-Wesley Publishing Company, 1995

[14] Lohaka Hippolyte O., Making a Grouped-Data Frequency Table: Development and Examination of the Iteration Algorithm, College of Education of Ohio University, 2007

[15] Dogan Nurhan, Dogan Ismet, "Determination of the Number of Bins/Classes used in Histograms and Frequency Tables: A Short Bibliography", Journal of Statistical Research, Vol. 7, No. 2(4), (2010), pp. 77-86

[16] Ioannidis Yannis, "The History of Histograms (abridged)", Proceedings of the 29th International Conference on Very Large Data Bases, Vol. 29, (2003), pp.19-30

[17] Tan Kong Yew, Cheng Jason Qi Quan, Chuah Joon Huang, "Investigation of Static Analog-to-Digital Converter Nonlinearity Measurement Using Histogram and Servo-Loop Method", Proceeding of the 2018 IEEE 5th International Conference on Smart Instrumentation, Measurement and Applications, (2018)

[18] Aflaifel Nasreen B., Chandhiok Nomita, Fawole Bukola, Geller Stacie E., Weeks Andrew D., "Use of Histograms to Assess the Efficacy of Uterotonic Treatment for Post-Partum Haemorrhage: A Feasibility Study", Best Practice & Research Clinical Obstetrics and Gynaecology, (2019), pp.1-13

[19] He Kun, Meeden Glen, "Selecting the Number of Bins in a Histogram: A Decision Theoretic Approach", Journal of Statistical Planning and Inference, Vol. 61, (1997), pp.49-59

[20] UC Irvine Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php.