A Generalized Piecewise Regression for Transportation Models

Al-Sayed Ahmed Al-Sobky, PhD
Public Works Department,
Faculty of Engineering, Ain Shams University,
Cairo, Egypt

Ibrahim M.I. Ramadan, PhD
Civil Engineering Department,
Faculty of Engineering
at Shoubra, Banha University,
Cairo, Egypt

ABSTRACT
This paper introduces a new piecewise regression methodology that can be used when linear regression fails to represent data. Effort can be saved to determine the best non-linear model shape using this methodology. Therefore, in this paper a nonlinear relationship is introduced using only one independent variable by a simple and direct way. The new approach depends on dividing the data set into several groups and then estimating the best line or segment for each group to perform a continuous broken line. The locations of breakpoints are determined by minimizing the sum of squared errors while the number of segments is determined by maximizing the adjusted coefficient of determination. The proposed approach can be used in many transportation applications such as trip generation models, zonal trip rates, nonlinear correlation coefficient, accident modeling, and traffic characteristics models. The proposed approach was tested against many practical examples and found that it can describe most of the transportation relationships properly and can decrease the number of variables used in the transportation modeling process. The proposed approach can be extended in the future to get the nonlinear relationship using more than one independent variable to cover the rest of transportation applications.

Keywords
Linear regression, segmented regression, nonlinear relationships, trip generation models, accident models.

1. INTRODUCTION
This paper introduces a new piecewise regression methodology that can be used when ordinary linear regression data fails to represent data. It can be helpful is case of non-linear regression model. It can help in determining the best shape of the non-linear function.

Segmented regression, in the proposed format, can be used to perform a nonlinear relationship in the shape of broken line. Applying this approach will help modelers to choose the best set of connected lines and then the best shape of nonlinear function suitable for the irregular data set such as in transportation.

There are three difficulties in the traditional segmented regression; the number of segments, locations of thresholds, and the heights of breakpoints. If there is no knowledge about the number and locations of breakpoints, they are determined by trial and error to minimize the sum of squared errors. The disadvantage of the trial and error approach is the unlimited locations for certain number of thresholds especially in the case of long range of data set. Therefore; the run time can't be expected. On the other hand, the optimum model may not be achieved due to the existence of local optimal solutions. This is typically the same disadvantage in selecting the nonlinear shapes.

Therefore, the objective of this research is to introduce a new approach to get the best locations of thresholds, the best heights of breakpoints, and the best number of segments to perform a simple and accurate nonlinear relationship between the dependent variable and the most effective independent variable. In addition, the nonlinear correlation coefficient for such model can be determined. Transportation modelers can determine the best shape of the nonlinear function that represents that data set using this approach. The proposed approach will be helpful in trip generation modeling, accident modeling, traffic stream modeling, traffic characteristics modeling, and many other types of transportation modeling.

This paper is divided into six main sections. Following the introduction is a literature review including the simple regression and the traditional segmented regression. The proposed approach including the determination of the number and locations of thresholds as well as the heights of breakpoints is presented in Section 3. Illustrative example will be explained in Section 4. Some applications will be presented in Section 5. In addition, the results of the proposed approach are compared with the traditional approaches in this section. Finally, a summary of the main conclusions and suggestions for future work are given in Section 6.

2. LITERATURE REVIEW
The method of least square error is the best fitting methods as the parameters determined by the least square error analysis are normally distributed about the true parameters with the least possible standard deviations [1].

This statement is based upon the assumption that the uncertainties in the data are uncorrelated and normally distributed. When the curve being fitted to the data is a straight line, the term "simple linear regression" is often used while if several independent variables are used, the term "multiple linear regression" is often used [1].

When the dependent variable y is assumed to be continuous, the model that could be used to predict the value of y based on the independent variable (or variables) is called a regression model. When the independent variable is in classes rather than a continuous variable, the problem might require a model that differentiates between these classes. This model is called segmental regression [1].

2.1 Regression Analysis
Regression analysis tries to fit the observations by finding the best line in the case of one independent variable or the best hyper plane in the case of multiple independent variables. It does this by achieving the lowest sum of squared errors for the whole data. The sum of the actual errors will be zero.

When analyzing a relationship between a response, y, and an explanatory variable, x, it may be apparent that for different ranges of x, different linear relationships occur. In these cases, a single linear model may not provide an adequate description for the relationship between x and y. In addition, a nonlinear model may not be recognized easily. Piecewise linear regression, which is called segmented regression, is a form of regression that determines a multiple lines model to fit the data for different ranges of x.

Nonlinear least square method is used to fit the calibration data set by a nonlinear function based on its known shape [2]. In this case, modelers have to test numerous numbers of nonlinear equations to achieve stronger model. These trials need special software able to differentiate the nonlinear equations or to use numerical analysis to minimize the sum of squared errors. Even using such software, modelers often consume time to select the adequate nonlinear shape.

In the last decades, Adaptive Neural Fuzzy Inference System, ANFIS, is used to build an accurate nonlinear model using any number of independent variables [3]. However, this model can't be easily interpreted and needs special software such as MATLAB [4] during the calibration process as well as during the model usage.

On the other hand, segmented or piecewise regression was used to fit the observations using a two-segment broken line based on one independent variable. This approach may bead equate for the patterns of non-human behavior such as crop production [5] and water discharge [6] as these patterns are not complex. Furthermore, many trials are needed to get the breakpoint that achieve higher coefficient of determination.

2.2 Traditional Segmented Regression

Segmented regression, also known as piecewise regression or broken-stick regression, is a method in regression analysis in which the independent variable is partitioned into intervals and a separate line segment is assigned to each interval. Segmented regression analysis can also be performed on multivariate data by partitioning the various independent variables. Segmented regression is useful when different relationships between the variables are exhibited in different regions. The boundaries between segments are known as thresholds which are important in the modeling and the decision making.

The least square error method is applied separately to each segment, by which the regression segments are made to fit the data set. In this case there is a correlation coefficient for each group and one correlation for the whole model.

In determination of the most suitable trend, statistical tests must be performed to ensure that this trend is reliable (significant). When no significant breakpoints can be detected, the problem becomes complicated. There was an effort that has been done in clustering(grouping) the data to determine the significant breakpoints. These methods for data clustering are K-means clustering [7], fuzzy C-means clustering [8], hierarchical clustering [9], clustering using Gaussian mixture model [10], and clustering using neural network [11]. All of these methods have some weak points. For example, in K-means clustering, initial cluster centers are assumed and then all data points is distributed into these initial clusters based on certain rule such as the Euclidian distance [12]. Many trials are needed to reach the situation in which the distance between every point inside each cluster and the center of this cluster is less than its distances to the centers of other clusters. Unfortunately, the final clusters are affected by the selection of the initial centers.

2.3 Traditional Data Grouping

2.3.1 Two segments with unknown threshold

Most of literatures that have been reviewed by the authors have dealt with only two groups with different method of grouping. One method for calculating the breakpoint, when there is only one breakpoint, at x=c, the model can be written as follows [13]:

$y = a1 + b1x$ for $x \leq c$……………………………...............……(1)

$y = a2 + b2x$ for $x \geq c$………………………………………...…(2)

To ensure a continuous regression function at the breakpoint, the two equations for y need to be equal at the breakpoint (when x = c):

$a1 + b1c = a2 + b2c$…………………………………..……..(3)

Nonlinear least squares regression techniques were used to fit this model to the data using an iterative technique.

2.3.2 Multi segments with unknown thresholds

Küchenhoff, 1996, described the complete algorithm three thresholds. He started with the generalized linear model with two breakpoints in the following parameterization.

$E(Y|X=x) = G\ (\alpha+\beta1x + \beta2(x-\tau1) + \beta3\ (x-\tau2))$………………..(4)

With $\tau1<\tau2$ and $\beta i\neq 0$ for i=2,3. He used the maximum likelihood analysis to estimate his model. Again his model is restricted to three segments but he stated that his model can be extended to the case of more than three segments.

2.3.3 Evaluation of the segmental model

There is no formal test available to evaluate the segmented regression model or to compare it with the nonlinear models. However, results can be evaluated using the model standard error (RMSE), the coefficient of determination (R2), or the adjusted coefficient of determination (Adjusted_R2).

The measures RMSE and Adjusted_R2 represent the goodness of the model fitting without reaching the over fitting case. The best model has smaller value of RMSE and high value of R2 and Adjusted_R2.For reviewing these statistical measures which have been used in estimating and testing the proposed approach, see 1.

3. THE PROPOSED APPROACH

In this paper, the proposed approach will be divided into three main steps; the first step explains how to get the approximate location of a new threshold. The second step depicts how to calculate the accurate location of the new threshold as well as the best heights of breakpoints in case of the number and the approximate locations of thresholds are known. The third step demonstrates how to determine the best number of thresholds. The proposed approach starts the first step to get the approximate location of the first threshold. The accurate location of the first threshold as well as the best height of its breakpoint will be calculated along with the second step. The significance of the first threshold will be examined using the third step. If the first threshold is significant, the approximate location of second threshold will be obtained by the first step again. The second step is used again to calculate the accurate locations of the two threshold as well as the best heights of their breakpoints. The third step is repeated to examine the significance of the second threshold. If the second threshold is significant, the above cycle will be repeated until reaching the insignificant threshold. The best number of thresholds is the maximum number of significant thresholds. In the following, each step will be separately explained.

3.1 Approximate Location of a New Threshold

Before getting the approximate location of the first threshold, a simple regression should be conducted to get the best one-line model and to calculate the residuals. The one-line model will be the best model if no residuals exist and then no need to add new thresholds. If the observations are not collinear, different residuals will be exist and then there is a need to add a new threshold to reduce the residuals and consequently to improve the goodness of fit. The location of the first threshold is assumed to be beside the maximum residual. This location will decrease the sum of squared errors rather than other locations.

Therefore, the approximate location of the first threshold will be near to the location of the maximum residual with respect to the best one-line model. The observations are divided into two approximate groups and the accurate location may be just right or left of the approximate location.

To get the approximate location of the second threshold, the best two-line model should be obtained as will be explained below. The approximate location of the second threshold will be near to the maximum residual with respect to the best two-line model and so on.

The new threshold will divide the observations into groups. Therefore, each group should have enough number of observations to be used to get the best line representing this group. Therefore, if the maximum residual location will create a group with insufficient observations, the following highest residual will be considered and so on until reaching the highest residual location which creates groups with sufficient number of observations.

Generally, each group should have at least two observations in addition to the observation at the maximum residual. Modeler can increase this limit to three or more if the observations are well distributed over the whole range. If there are three observations at the same location, the minimum number of observation in every group should be four or more to avoid the vertical lines in the model.

3.2 Breakpoints Determination

After determining the approximate locations of certain number of thresholds which divide the observations into approximate groups, least square error method is used to get the accurate locations and heights of the breakpoints in two stages. The first stage aims to improve the grouping system by determining which groups contain the observations at the approximate location. The second stage is to get the accurate locations of the thresholds as well as the heights of breakpoints.

The two stages are identically except in the observations at the approximate locations of threshold. They will be considered in the right and left groups in the first stage. However, they will be considered only in one group in the second stage.

Generally, the model parameters to be determined in the two stages will be the slope Si of every line-segment (i), the accurate threshold location Ti for every threshold (i) as well as the initial height H0 of the model at the minimum value X0.

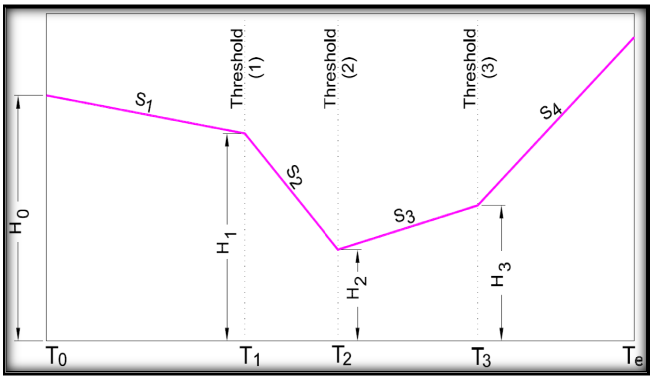

Figure 1: The Model Parameters off our-line Model

Figure 1 explains the case of four-line model which has three thresholds at locations T1, T2, and T3, three breakpoints at heights H1, H2, and H3, and four slopes S1, S2, S3,and S4 for the four line segments.

The model parameters will be determined using linear least square method as follow:

- The height H_i of every breakpoint (i) should be calculated in terms of its location T_i, the height and location of the previous breakpoint T_{i-1} and H_{i-1}respectively, as well as the slope S_i of the line-segment ended by the breakpoint (i) as follows:
- $H_i = H_{i-1} + S_i \cdot (T_i - T_{i-1})$ ……………………….(5)
- The equation of every line-segment can be written as follows:
- $Y_p^e = H_{i-1} + S_i \cdot (X_p - T_{i-1})$…………………..……………(6)
- Where X_p and Y_p^e are the observed input and the estimated output values for the point (p) respectively, H_{i-1}, S_i and T_{i-1} are as defined above. The point (p) should belong to the group (i) which is just before the threshold (i).
- Sum of squared errors SSE_i for every group (i) can be calculated for all data points belong to the group (i).
- Sum of squared errors SSE for the whole model is the summation of SSE_i for all groups. It should be noticed that the locations and heights of all breakpoints, equations of all line segments, sum of squared errors for all groups, and sum of squared errors for the whole model were computed in terms of H_0, the slopes $S_1, S_2, \ldots, S_{M+1}$as well as the locations of thresholds $T_1, T_2, \ldots, T_M$ where M is the number of thresholds.
- Using the concept of least square method, SSE should be partially differentiated with respect to all model parameters; $H_0, S_1, S_2, \ldots S_M, T_1, T_2, \ldots, T_{M-1}$and then (2M) non-linear equations in (2M) unknowns will be obtained. Solving these equations, the model parameters will be determined.
- After obtaining the new locations of thresholds in the first stage, the groups containing the observations at the approximate locations of thresholds become known and then the accurate grouping system should be used in the second stage to get the accurate locations and heights of breakpoints by repeating the same steps of stage one.

- After obtaining the accurate model parameters in the second stage, all heights of breakpoints as well as all equations of line segments can be calculated by substituting in the equations 5 and 6.
- The model can be validated using the calibration data set by computing the validation measures R^2, RMSE and Adjusted_R^2 to get the goodness of fitting.

The importance of the Adjusted_R2 measure is to take into consideration the total number of model parameters including the thresholds, initial height, and the slopes. Therefore, the statistic measure Adjusted_R2 will be used to judge the importance and significance of a certain threshold as will be explained below.

3.3 Number of Segments

Section 3.1 explains how to get the approximate location of the first threshold while Section 3.2 explains how to get the accurate location of the first threshold, the height of its breakpoint as well as the Adjusted_R2. The significance of the first threshold is determined as follows:

- If the value of the Adjusted_R^2 for the two-line model using the first threshold is not better than the value of the Adjusted_R^2 of the one-line model, then the first threshold is not significant and no need to add it. Consequently, the one-line model is better than the segmented model in representing the calibration data set.
- If the value of the Adjusted_$R^2$for the two-line model using the first threshold is better than the value of the Adjusted_R^2 of the one-line model, then the first threshold is significant. Consequently, the obtained two-line model is better than the linear model in representing the calibration data set and a second threshold should be examined.

Generally, the best number of thresholds (m) occurs when the value of the Adjusted_R2 for the segmented model having (m+1)-line model is better than the value of the Adjusted_R2 for the segmented model having (m+2) lines.

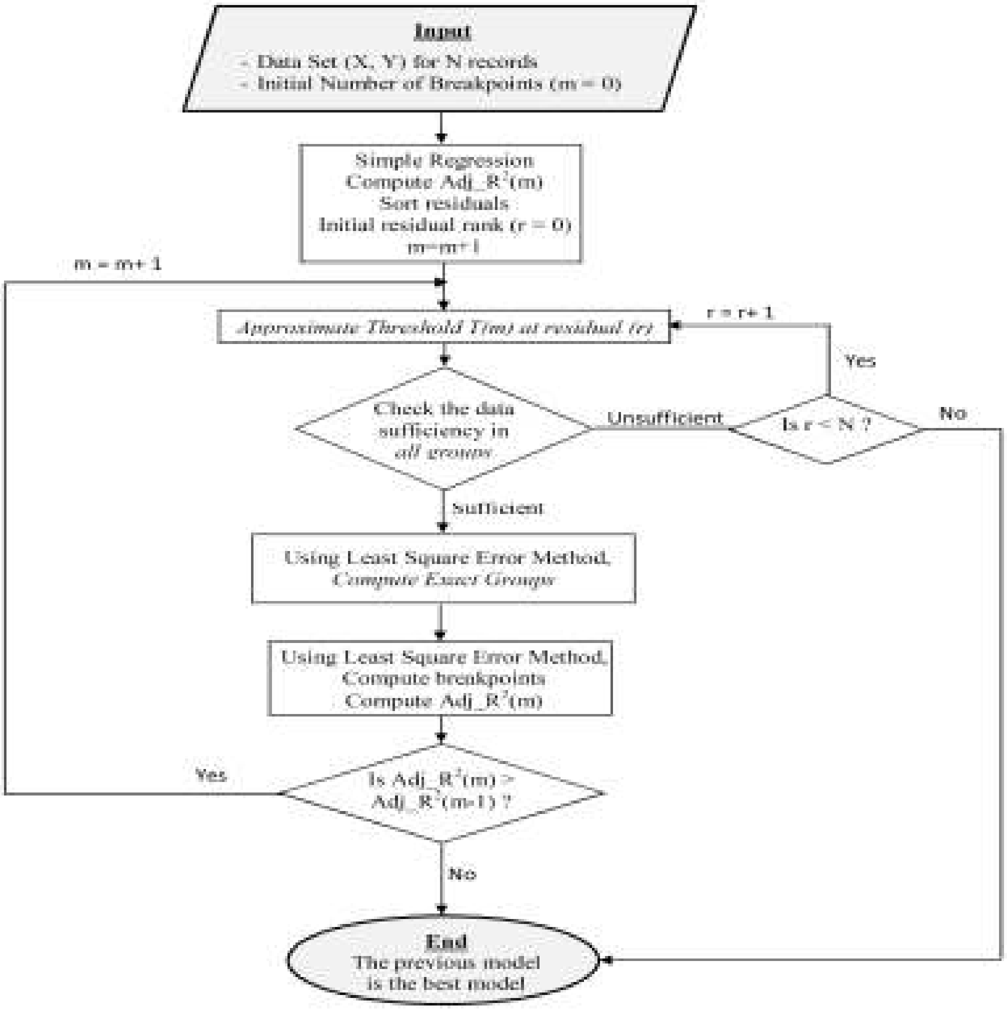

Figure 2: Flowchart for the Proposed Approach

It is worth mentioning that if the value of the Adjusted_R2 of the best number of segments is considered low, then adding a new threshold will not improve the model and in this case the modelers have to add new independent variable not to use other non-linear models. Therefore, the value of the Adjusted_R2 for the best segmented model can express the goodness of non-linear correlation. Figure 2 shows a flowchart for the proposed approach.

The proposed approach along with its three steps will be illustrated numerically by the following example.

4. NUMERICAL EXAMPLE

The calibration data used in this synthetic example is depicted in Table 1.

Table 1: Data used in the Numerical Example

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	1.9	8.9	18.3	26.8	30.0	27.3	23.6	18.1	18.8	16.7	14.9	11.1	9.1	8.3	7.6

4.1 Approximate Location of the First Threshold

The one-line model is obtained using the simple regression as illustrated in Figure 3.

Ype = 20.202 – 0.514 X …….(7)

Where $R^2 = 0.076$, and Adjusted_$R^2 = 0.005$

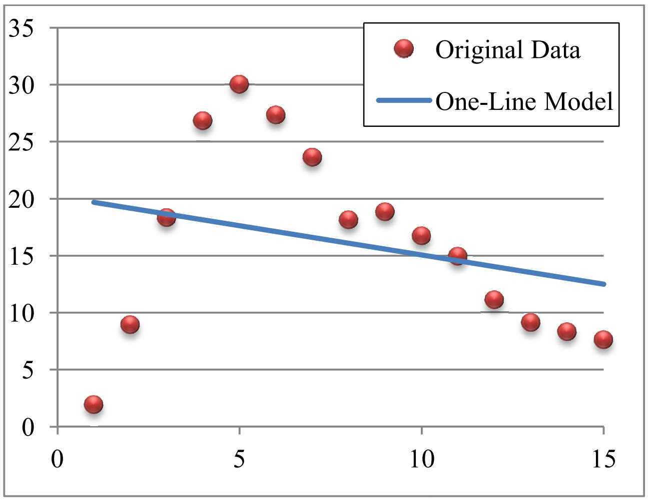

Figure 3: Graphical Representation of the Best One-Line Model

The maximum residual occurs at x=1; however, this locations can't be consider an approximate location for the first threshold as it divides the observation into a right group having all points and a left group having no points. Therefore, the approximate location of the first threshold will be at the second highest residual (at x = 5) which divides the observation into a right group having 9 points and a left group having 4 points which are enough to estimate the two-line model.

4.2 First Breakpoint Determination

The best two-line model was obtained using the procedures stated in Section 3.2 and found as illustrated in Figure4.

$$\text{Ype}=\begin{cases} 1.36 + 8.41(X-1) & X \le 4.363 \\ 29.644 - 2.267(X - 4.363) & X \ge 4.363 \end{cases} \quad \ldots\ldots\ldots\ldots(8)$$

Where $R^2 = 0.975$, and Adjusted_$R^2 = 0.968$.

The first threshold is significant because it increases the value of Adjusted_R^2 from 0.005 into 0.968.

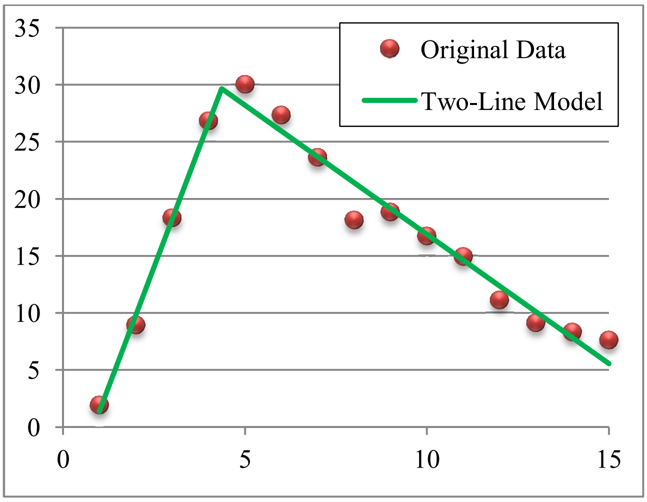

Figure 4: Graphical Representation of the Best Two-Line Model

4.3 Approximate Location of the Second Threshold

The maximum residual occurs at x=8. Therefore, the approximate location of the second threshold will be at x = 8 as it divides the observations into a right group having at least 7 points and a left group having at least 3 points which are enough to estimate the three-line model.

4.4 Two Breakpoints Determination

The best three-line model was obtained as stated in Section 3.2 and found to be as follow:

$$\text{Ype} = \begin{cases} 1.372 + 8.452(X-1) & X \le 4.588 \\ 31.693 - 3.53 * (X - 4.588) & 4.588 \le X \le 8.028 \ldots(9) \\ 19.549 - 1.846 * (X - 8.028) & X \ge 8.028 \end{cases}$$

Where $R^2 = 0.989$, and Adjusted_$R^2 = 0.983$.The second threshold is significant because it increase the value of Adjusted_R^2 form 0.968 into 0.983. Figure 5shows the graphical representation of the best three- line model.

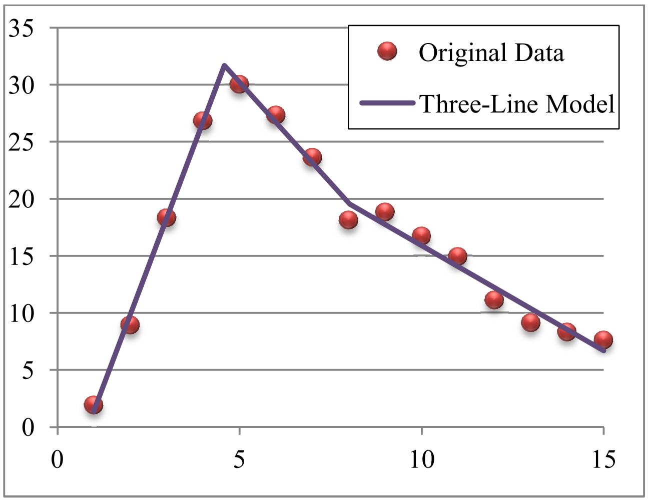

Figure 5 Graphical representation of the best three line model.

The third threshold is examined and found to be not significant as the value of Adjusted_R2 of the best four-line model is 0.981 which is less than that of the three-line model. Therefore, the two thresholds model is the best one.

5. APPLICATIONS

The proposed segmented model could be used in many applications of the transportation modeling; the following applications are some examples:

5.1 Trip Generation Rate

The average trip ends (T) of hotels versus employees (X)at P.M. peak hour was determined in ITE 7 page 564. The average rate is 0.9 and the best fitted curve is as follow:

$\ln(T) = 0.63 \ln(X)+1.89$ where R2 is 0.72.

The proposed model was applied for the data extracted from ITE and the best segmental model was a two-line model with one threshold at X = 310 as follows (see Figure 6) where R2 = 0.829and Adjusted_R2=0.772:

T=

$$\begin{cases} 53.092 + 0.765X & X \leq 310 \\ 290.27 + 0.013(X - 310) & X > 310 \end{cases} \text{………………..……….. (10)}$$

The average rates in the proposed model are 1.108and 0.609for the left and right groups respectively. The proposed model can be considered better than ITE model especially for the hotels having high number of employees. On the other hand, using ITE average rate for all observations leads to negative value of R2 i.e. the average trips (184) for all observations is better than using the average rate. While using the proposed average rates leads to R2 = 0.59.

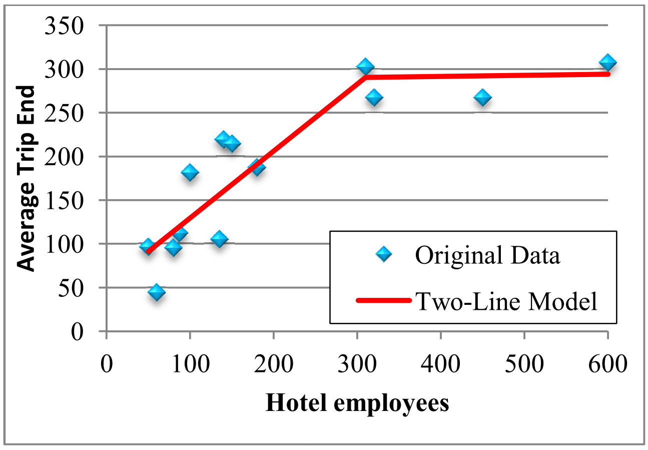

Figure 6: Trip Rate Example

5.2 Accident Model

Boakye Agyemang et al, 2013 have proved that there is a strong correlation between the accidents (Y) in Ghana and population (X). The model was a simple regression model as follows:

Y = -263 + 0.000532X where the values of R2 and Adjusted_R2 were 0.729 and 0.713 respectively.

Applying the proposed model, the thresholds will be at population of 16.1056 and 17.939569 Millions capita and the proposed model will be as follows (see Figure 7):

Y=

$$\begin{cases} 20{,}336 - 0.00085\,X & X \leq 16{,}056{,}000 \\ 6{,}75 + 0.002 * (X - 16{,}056{,}000)\text{: } 16{,}065{,}000 < X \leq 17{,}939{,}569 & \\ 10{,}556 + 0.00022 * (X - 17{,}939{,}569) & X > 17{,}939{,}569 \end{cases}$$

……………….(11)

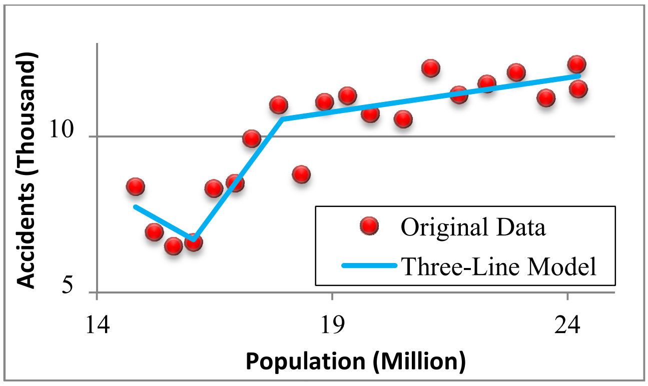

Figure 7: Accident model example

The values of R2 and Adjusted_R2 are 0.883 and 0.841 respectively which ensue that there is a strong relationship between the accident in Ghana and the population. However; to get better accident model other variables should be used beside population.

5.3 Traffic Flow Relationship

The proposed model could be used to describe the traffic flow characteristics based on the density-speed relationship. The observations were imported from an Egyptian road (Ismailia desert road) and the best traditional model was:

Speed = 349.3*Density -0.29 - 79.7…………………...……..(12)

Where R2= 0.969 and Adjusted_R2=0.967

The density jam is 163.27 and the free flow speed is undefined

The proposed model is as follows (See Figure 8):

Speed=

$$\begin{cases} 108.40 - 1.85\ \text{Density} & 0 \leq \text{Density} \leq 23.08 \\ 65.8 - 2.38 * (\text{Density} - 23.08): & 23.08 < Density \leq 33.8 \\ 40.28 - 0.32 * (\text{Density} - 33.82): & \text{Density} > 33.82 \end{cases}$$

……………..….(13)

Where R2= 0.983 and Adjusted_R2=0.980

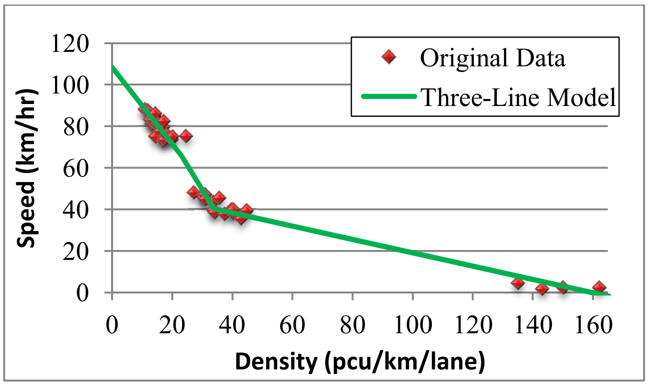

Figure 8: Density-Speed Relationship Example

Based on the proposed model, the jam density is 160.59 and the free flow speed is 108.40 which are more logical than that in the traditional model.

In addition, the traffic condition for this road can be approximately classified into three categories; low traffic up to density of 23veh/km or speed more than 66km/h, medium traffic for density between 23-34 veh/km or speed between 40-66 km/hr, and high traffic after density of 34 veh/hr or speed less than 40 km/hr.

5.4 Estimation of Nonlinear Correlation

If the relationship between the dependent and independent variables is nonlinear, as in most of transportation patterns, a segmented model can be calibrated and then the value of Adjusted_R2 can be used as an indicator for the nonlinear correlation between the two variables. In case of weak correlation, it's recommended to use another input variable not to use another non-linear model. Therefore, the value of the Adjusted_R2 for the best segmented model can express the so called non-linear correlation.

On the other hand, the most effective independent variable is the input achieving higher value of Adjusted_R2 using the segmented model.

6. CONCLUSIONS AND RECOMMENDATIONS

This paper presented a new simple approach for using the piecewise regression analysis to emulate the transportation patterns using one independent variable. In addition, a simple procedure was suggested for segmenting the calibration data set and then determining the best broken line model. The proposed single-input model was applied to estimate the nonlinear correlation coefficient, select the effective independent variable, determine the best number of segments, and the best equation for each line segment. The proposed approach was validated using case studies. Results are compared with previous models. The comparison confirms that the proposed model is simple, accurate, reliable, economic and programmable. The proposed approach can be used in many transportation applications such as trip generation, accident modeling, and traffic stream modeling to achieve an accepted accuracy using only one independent variable. Further research is needed in the area of using more than one variable to get more accurate modes.

7. REFERENCES

[1] Lyman ott, 1988: "An Introduction to statistical methods and data analysis" , PWS- Kent publishing company, Boston , USA.

[2] Greene, W, 2000,"Econometric Analysis, Upper Saddle River": Prentice-Hall.

[3] Jang, J.S.R., 1993, " ANFIS: Adaptive-network-based fuzzy inference system", IEEE Trans. Syst., Man, Cybern. 23, 665-685.

[4] MathworksInc, 2014, MATLAB 2014a User's Guide. .

[5] R.J. Oosterbaan, D.P.Sharma, K.N.Singh and K.V.G.K.Rao, 1990, Crop production and soil salinity: evaluation of field data from India by segmented linear regression. In: Proceedings of the Symposium on Land Drainage for Salinity Control in Arid and Semi-Arid Regions, February 25th to March 2nd, 1990, Cairo, Egypt, Vol. 3, Session V, p. 373 - 383.

[6] Ryan, Sandra E.; Porth, Laurie S, 2007, "A tutorial on the piecewise regression approach applied to bed load transport data". Gen. Tech. Rep. RMRS-GTR-189. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 41 p.

[7] Tapas Kanungo, et al. An Efficient k-Means, 2002,"Clustering Algorithm: Analysis and Implementation". IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 7.

[8] Subhagata Chattopadhyay, et al, 2011, "A COMPARATIVE STUDY OF FUZZY C-MEANS

[9] Gunnar Carlsson ,FacundoM´emoli , 2010, "Characterization, Stability and Convergence of Hierarchical Clustering Methods", Journal of Machine Learning Research 11 (2010) 1425-1470

[10] Charles A. Bouman, et al, 2005, "CLUSTER: An Unsupervised Algorithm for Modeling Gaussian Mixtures", http://www.ece.purdue.edu/~bouman

[11] Harald Hruschka, and Martin Natter, 1992, " Using Neural network for clustering based market segmentation", Forschungsbericht/ Research memorandum No 307.

[12] Hartigan, J., & Wang, M. A K-means clustering algorithm, 1979, Applied Statistics, 28, 100–108.

[13] Sandra E. Ryan, and Laurie S. Porth, 2007, " A Tutorial on the Piecewise Regression Approach Applied to Bedload Transport Data", General Technical Report RMRS-GTR-189