VOL. 11, NO. 16, AUGUST 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

MEAN, MEDIAN OR SOMETHING ELSE?

Lloyd Ling and Zulkifli Yusop
Centre for Environmental Sustainability and Water Security, Research Institute for Sustainable Environment, Faculty of Civil Engineering Department, Universiti Teknologi Malaysia., Skudai, Johor, Malaysia
E-Mail: lloyd.ling@gmail.com

ABSTRACT

Non parametric inferential statistics was used to guide the numerical optimization study to search for the optimum result in this article. The technique was demonstrated in a case study and a significant improved predictive model was formulated with 4% less residual sum of squares (RSS) than the median model, 20% less than the mean model and 79% less than a benchmarked empirical model. The methodology proposed herewith addressed the selection dilemma between mean and median. It identified an optimum value and formulated a better predictive model than those by either mean or median.

Keywords: bootstrapping, non-parametric inferential statistics, numerical analysis, SPSS.

INTRODUCTION

Researchers often face the decision to choose between mean and median of a dataset as a better collective representation in their study. The decision often leads to a consequential predictive modelling formulation, and therefore it is crucial to be able to make the best selection. The issue spans across different field of studies as a universal dilemma. In hydrological research, some researcher recommended using the median while other recommended the mean value (Schneider and McCuen, 2005), (Hawkins et. al., 2009), (Hawkins, 2014). This study proposed to utilise numerical analysis algorithm (Fattorini, 1999), (Mordecai, 2003), (Jon, 2004), (Jorge and Stephen, 2006), (Ruszczyński, 2006) guided by inferential statistics (Rao, 1997), (Young and Smith, 2005), (Cox, 2006), (Efron, 2010), (Ling and Yusop, 2014, 2014b) for assessment.

In the past few years, the proposed methodology was tested rigorously with rainfall-runoff data and modelling with an aim to hone the runoff prediction result of an empirical model. The methodology was demonstrated through a selected case study which used piecewise function to model the correlation between A and B as:

$$B = \frac{(A-D)^2}{A-D+C} \quad (1)$$

for $A > D$, else $B = 0$

A, B = given observed data
C, D = fitting parameter

C correlates to D via another parameter L in the form of $C = D/L$ (L is a fitting value between 0 and 1). All parameters are either positive integers or real numbers. For illustration, this study used a twenty-two A-B positive real number dataset below 25 (hydrological dataset not shown here) and D was pre-determined to be 0.001 to substitute into equation (1) and simplify (1) into:

$$B = \frac{(A-0.001)^2}{A-0.001+\frac{0.001}{L}} \quad (2)$$

for $A > 0.001$, else $B = 0$

where A, B and L are as defined in previous section. An empirical model assumed that data distribution nature would not affect the model predictability and proposed $L = 0.2$ to further reduce equation (2) into a simple form of:

$$B = \frac{(A-0.001)^2}{A+0.004} \quad (3)$$

for $A > 0.001$, else $B = 0$

where A and B are as defined in previous section.

METHODOLOGY

Given the A-B dataset, twenty-two L values were derived from equation (2). The descriptive statistics of L values was tabulated in Table 1. The study will identify a best collective representation of L value for the dataset and benchmarked against the empirical (model) equation (3) where L was proposed and assumed to be 0.2 despite of its data distribution. Inferential non-parametric statistics was employed for two claim assessments set forth by the empirical model's assumption with two Null hypotheses:

Null Hypothesis 1 (H_{01}): Equation (3) applies for every dataset.
Null Hypothesis 2 (H_{02}): $L = 0.20$ and the value of 0.20 is a constant in equation (3).

The rejection of H_{01} implies that the empirical (model) equation (3) is invalid and not applicable for the dataset of this study, while H_{02} rejection indicates that L is not a constant as proposed but a variable. Rejection of both hypotheses will pave way to derive new L value. The selection of a different L value will formulate a new

predictive model using equation (2). Beside the derived mean and median value of *L* dataset, numerical analysis algorithm will search through both mean and median's confidence interval range for an optimum *L* value at alpha = 0.01 level. The goal is to formulate different predictive model using different *L* value for further comparison analyses. In order to include confidence intervals into the discussion, bootstrapping technique, Bias corrected and accelerated (BCa) procedure (2000 samples) was conducted at a stringent 99% confidence level on the *L* dataset (Efron and Tibshirani, 1994), (Davison and Hinkley, 1997) (Rochoxicz, 2011).

MEAN, MEDIAN AND OPTIMUM *L*

The *L* dataset is not normally distributed from its skewness and kurtosis (2.41, 7.23), conventional statistical practice is to adopt the median value as the collective representation for the skewed *L* dataset. Mean, median value and the respective BCa 99% confident interval range are 0.0013 [0.00086, 0.00181] and 0.0011 [0.00079, 0.00131]. BCa biases of mean and median values were at proximate range but mean value has larger BCa standard error than median. Some researchers would recommend selecting median over mean to represent *L* dataset in this case (Wright, 1997), (Howell, 2007), (Hawkins, 2014).

Under the circumstance, the study turned to a different premise for sourcing solution and further assessment. *L* optimization study can be conducted via numerical analyses approach using equation (2). The least square fitting algorithm was set to identify an optimum *L* value by minimizing the residual sum of squares (*RSS*)[1] between predicted *B* and its observed values.

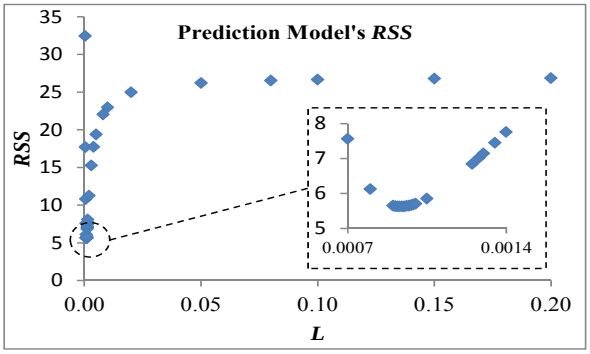

Figure-1. Prediction model's *RSS* of equation (2) as per *L* variations. The dash box shows the magnified view of the lowest *RSS* near to the origin.

The optimization study was conducted twice where first (un-supervised) attempt was based on *L* variation across a wide range values from 0.00001 to 1,000,000 on equation (2). Same algorithm was repeated to (supervised) search for the optimum *L* value but within the BCa confidence interval limits only of both mean and median of the derived *L* dataset in order to confirm the optmized result of the first attempt.

Prediction model's *RSS* varied according to *L* as depicted in Figure-1. The optimization study via numerical analysis identified an optimum *L* value to be 0.00093 where overall predcitive model's RSS is the lowest. This optimum *L* value is neither the mean nor median of the *L* dataset as tabulated in Table-1. Had numerical analysis produced a third choice and compounded the dilemma into a "trilemma" paradox?

MEAN, MEDIAN AND OPTIMUM *L* MODEL

The Mean, median and optimum *L* values were used to formulate three prediction models using equation (2) in order to study the model's prediction efficiency (*E*)[2] respectively and draw further comparisons. Three prediction models were formulated as below. Using mean value, piecewise model (2) becomes:

$$B = \frac{(A-0.001)^2}{A+0.796} \quad (4)$$

for $A > 0.001$, else $B = 0$

Using median value, piecewise model (2) becomes:

$$B = \frac{(A-0.001)^2}{A+0.951} \quad (5)$$

for $A > 0.001$, else $B = 0$

Using optimum *L* value, piecewise model (2) becomes:

$$B = \frac{(A-0.001)^2}{A+1.074} \quad (6)$$

for $A > 0.001$, else $B = 0$

where *A and B* are the same parameters as stated in previous section.

COMPARISON OF PREDICTIVE MODEL'S PREDICTION PERCENTAGE ERROR

Prediction (models) equation (4-6) were benchmarked against empirical (model) equation (3) for further model predictive accuracy assessment. Besides *E* and *RSS*, Bootstrapping BCa procedure (99% confident interval level with 2000 samples) was again employed to analyse model's prediction percentage error pattern for comparison. BCa results generated 99% confidence intervals of the mean, 5% trimmed mean and median of the prediction percentage error from each model. BCa confidence interval spanning across zero indicates high likelihood for the model to yield accurate prediction results (where 0% prediction error cannot be ruled out at

[1] $RSS = \sum_{i=1}^{n} \left(B_{predicted} - B_{observed}\right)^2$

[2] $E = 1 - \frac{RSS}{\sum_{i=1}^{n} \left(B_{predicted} - B_{mean}\right)^2}$

alpha = 0.01). Tabulated BCa results were shown in Table 2.

OVERALL MODEL PREDICTION ERROR AND *RSS*

Based on equation (2), the overall model prediction error (*Err*) can be calculated by the summation of predictive model's residual to indicate the overall model prediction pattern. Zero value indicates a perfect overall model prediction with no error, the negative value indicates the overall model tendency of under-prediction and vice versa. Figure-2 depicted overall predictive model's error tendencies with respect to *L* variation. Figure-3 re-presented the grand schema of this indicator to *RSS* due to *L* variation.

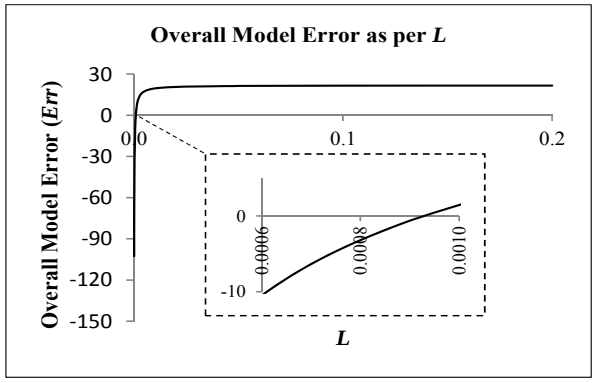

Figure-2. Predictive model error according to *L* variation. The dash box magnified the area near to the origin where the overall model error curve crossed x-axis (model error = 0).

As shown in Figure-2, the optimum *L* value to yield the overall predictive model error of zero lies within the range near to the origin. The correlation between overall model error (*Err*) and *L* variation was modelled by using IBM, PASW (version 18) with the following equation for *L* variation range of [0.0001, 0.2]. (Adjusted R square = 1.000, Standard error = 0.001, $p < 0.001$):

$$Err = e^{4.882 - \frac{0.0002}{L}} - 110 \quad (7)$$

Err = overall model prediction error
L = as defined previously

To study the predictive model's *RSS* according to *L* variation, the correlation between model *RSS* and overall model error (Err) was modelled with the following equation for *L* variation range of [0.0001, 0.2]:

$$RSS = -0.00002Err^3 + 0.046Err^2 - 0.005Err + 5.642 \quad (8)$$

where Err and RSS are as defined previously. Equation (8) has an adjusted R square = 1.000, Standard error < 0.001 and $p < 0.001$.

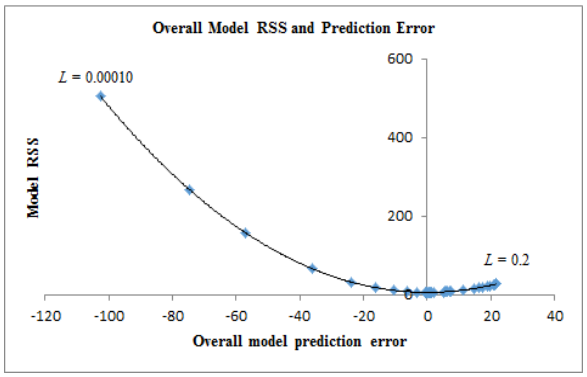

Figure-3. Overall model RSS & prediction error as per L variation graph shows that as L value = 0.0001, the overall model prediction error will be negative (model under-prediction tendency shown on negative x-axis) while overall model RSS also increases to large magnitude (shown on y-axis). Empirical model (L=0.2 as indicated on the lower right area) tends to have positive overall prediction error (model over-prediction tendency shown on positive x-axis).

Equation (7) was modelled to represent Figure-2 while equation (8) describes Figure-3. As shown in Figure-3, optimum *L* value will yield lowest *RSS* and *Err* around the region near to the origin. Figure-4 is the close up view of Figure-3 near to the origin.

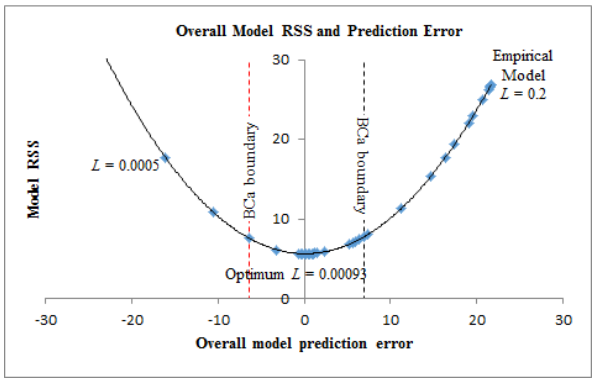

Figure-4. Zoom in view of Figure-3 near to the origin. 3 different L values were indicated on the curve for illustration purpose. Optimum L = 0.00093 yields lowest model RSS (lowest point on y-axis) and overall model prediction error (near to zero on x-axis). The BCa boundary lines (in dash lines) are the 99% BCa confidence interval of the L dataset of this study [0.00079, 0.00181].

RESULTS AND DISCUSSION

The dilemma to select between mean and median as a better collective representation of a dataset was addressed in this study through the proposed optimization study via numerical analysis approach guided by non-parametric inferential statistics. The ultimate goal of this study is to identify the best collective representation value

from a dataset for the formulation of a good predictive model without the limitation to choose between mean and median only.

The common pitfall in the least square fitting algorithm is to wrongly identify local minima or maxima as optimum solution thus producing inconsistent wrong results. The initial guess point for least square fitting algorithm to commence an optimization search often played an influential role to end results. Researchers often started the initial guess point with a wild guess which could lead to a wrong conclusion (Hansen, 1992), (Horst and Tuy, 1996), (Fattorini, 1999), (Mordecai, 2003), (Jon, 2004), (Jorge and Stephen, 2006), (Ruszczyński, 2006). Un-supervised least square fitting algorithm started to produce inconsistent optimum *L* results when the initial guess point was > 1000 in this study. However, optimum *L* value remained the same only if the initial guess point was set at any value below 1000. On the other hand, repetitive optimum *L* search within BCa confidence interval ranges consistently converged to the same optimum *L* value. The Inferential statistics can be an effective guide to narrow the optimum *L* search and identify a statistical significant optimum *L* solution in swift and precise manner. This study assessed the validity of the optimization study methodology (numerical analysis algorithm in specific) and identified the optimum *L* value with the guide from inferential statistics. The optimum *L* result was also verified against the BCa results to assure that least square algorithm did not wrongly identify local minima as global minima.

Based on proximate comparison results in Table 1, the median *L* value would be selected as a better choice over mean. As the *L* dataset is not so normally distributed, median *L* value has always been recommended as a better choice. The selection dilemma appears to be addressed but is the median *L* value really the best collective representation of the *L* dataset to formulate the best predictive model? Was the decision based upon proximate range comparison between mean and median *L* too subjective? The proposed methodology in this study identified an optimum *L* value other than mean and median which complicated the decision making to choose between mean and median only. In order to further analyse the difficult deadlocked situation, all three *L* values were adopted for the predictive model formulation for further evaluations.

Under the optimization convergence and divergence study, predictive model's *RSS* was found to decrease as *L* varied away from mean toward the optimum *L* value (Figure-1). Predictive model comparison study (table 2) also showed that equation (6) has highest *E* (0.991), 20% less *RSS* than equation (4) and 4% less than equation (5). Equation (6) emerged as the better prediction model followed by equation (5) and then equation (4).

Significance of the empirical model's assumption to further simplify equation (2) into equation (3) was challenged as well. From Table (1), neither BCa confidence intervals of mean nor median of the *L* value includes 0.2, and therefore the empirical model's simplification proposal (H_{01}) can be rejected at alpha = 0.01 level. Table 1 also showed that BCa mean confidence interval range housed 110% possible *L* variation within its upper and lower limit while median harboured 67%, these wide *L* variation possibilities implied that *L* is impossible to be a constant value (0.2) but a variable thus H_{02} can be rejected at alpha = 0.01 level as well. These *L* variations indicated that any *L* value within those ranges would be a significant fit at the same alpha level thus the optimum *L* can also be justified with BCa results as the best exemplification of *L* dataset (at alpha = 0.01) because it falls within both BCa confidence interval ranges.

Empirical (model) equation (3) consistently over-predicted *B* values, the model prediction percentage error's mean, 5% trimmed mean and median had positive confident interval range (Table 2). Its standard deviation of the model prediction percentage error fluctuated at positive intervals with highest positive model prediction error (over-prediction) range compared to mean (model) equation (4), median (model) equation (5) and the optimum *L* (model) equation (6). Contrary, optimum *L* (model) equation (6) had model prediction percentage error confidence interval ranges which spanned from negative to positive values. These interval span indicated the likelihood (at alpha = 0.01) of having zero percentage prediction error (Table 2). Model equation (6) also has higher *E* value and 79% less *RSS* than (model) equation (3). The empirical model's assumption to simplify equation (2) into equation (3) induced more prediction errors and over-predicted *B* consistently in this study.

BCa results comparison (Table 2) between (model) equation (4), (5) and (6) showed that predictive model equation (4) has the highest range of prediction percentage error with higher error fluctuation than its counterpart models. On average, model equation (4) over-predicted *B* values while the other two models showed model prediction percentage error with confidence interval ranges which spanned across zero. Those spanning pattern indicated that equation (5) and equation (6) are likely (at alpha = 0.01) to produce accurate *B* predictions with zero percentage error. Model equation (5) and equation (6) are therefore better predictive models than (model) equation (4). BCa results comparison between (model) equation (5) and equation (6) showed that (model) equation (6) managed to predict *B* values with lower percentage error and less error fluctuations than (model) equation (5) thus further demarcated (model) equation (6) as the best predictive model among the three.

Optimization study approach was based on the minimization of *RSS* to determine the optimum result but neither *RSS* nor its conjugate *E* can offer any insight about the predictive model's prediction pattern. Besides BCa results (Table 2), model prediction error (*Err*) was used to analyse predictive model's prediction patterns (model's overall over-prediction or under-prediction tendency). Equation (7) and (8) were modelled to provide an overview of the *RSS*, *E* and the prediction tendency (*Err*) of a predictive model due to *L* variation.

BCa boundaries (as shown in Figure-4) identified a range where the predictive model's over-prediction tendency crossed over to under-prediction pattern at the optimum point which was correctly identified by numerical analysis algorithm within the indicated BCa confidence interval range. The optimum *L* value yielded an overall model prediction error near to zero. Further reduction in *L* value beyond the optimum point will produce predictive model with higher *RSS* and under-prediction tendency for *B* values (Figure-4).

CONCLUSIONS

Inferential statistics narrowed the optimum search band while optimization study pin pointed an optimum *L* value within the BCa confidence interval range; both methods supplemented each other in this regard. The optimum *L* value (0.00093) from this study can also be verified by taking the second derivative of equation (8), where a local minimum was found to have an overall model prediction error (*Err*) of 0.049. The corresponding *L* value with the overall model prediction error can be solved with equation (7) to be 0.00093. This result verified that numerical analyses optimisation algorithm had correctly identified a statistical significant optimum *L* value.

The rejection of both null hypotheses concluded that data distribution plays an influential role. The assumption that $L = 0.2$ is invalid and cannot be treated as a constant to simplify equation (2) into the empirical (model) equation (3) which becomes obsolete and not applicable for the dataset in this study. The use of the empirical (model) equation (3) commits type II error.

The choice of the optimum *L* value as the collective representation for *L* dataset formulated the best predictive model equation (6) in this study. The proposed methodology identified an optimum *L* value and formulated a significantly better predictive model than those by either mean or median. It also addressed the common selection dilemma faced by many researchers.

ACKNOWLEDGMENTS

The author would like to thank Universiti Teknologi Malaysia, Centre for Environmental Sustainability and Water Security, Research Institute for Sustainable Environment of UTM, vote no. Q.J130000.2509.07H23 and R.J130000.3009.00M41 for its financial support in this study. This study was also supported by the Asian Core Program of the Japanese Society for the Promotion of Science (JSPS) and the Ministry of Higher Education (MOHE) Malaysia.

REFERENCES

[1] Cox, D.R. (2006). Principles of Statistical Inference, Cambridge University Press. ISBN 0-521-68567-2.

[2] Davison, A.C. and Hinkley, D.V. (1997). Bootstrap methods and their application. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN 0-521-57391-2.

[3] Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs/Cambridge University Press. ISBN 9780521192491.

[4] Efron, B. and Tibshirani, R. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC. ISBN 978-0-412-04231-7.

[5] Fattorini, H.O. (1999). Infinite Dimensional Optimization and Control Theory, Cambridge Univ. Press, 1999.

[6] Hansen, E.R. (1992), Global Optimization using Interval Analysis, Marcel Dekker, New York.

[7] Hawkins, R. H., Ward, T., Woodward, D. E. and Van Mullem J. (2009): Curve Number Hydrology: State of the Practice. Reston, Virginia, ASCE.

[8] Hawkins, R. H. (2014 e-mail communication).

[9] Horst, R. and Tuy, H. (1996) Global Optimization: Deterministic Approaches, Springer.

[10] Howell, D.C. (2007). Statistical methods for psychology (6th Ed.). CA: Thomson Wadsworth, Belmont.

[11] Jon, L. (2004). A First Course in Combinatorial Optimization; Cambridge University Press. ISBN 0-521-01012-8.

[12] Jorge, N. and Stephen J.W. (2006). Numerical Optimization. Springer. ISBN 0-387-30303-0.

[13] Ling, L. and Yusop, Z. (2014). Inferential Statistics of Claim assessment. AIP Conference Proceedings: ISBN: 978-0-7354-1274-3; http://dx.doi.org/10.1063/1.4903675.805.

[14] Ling, L. and Yusop, Z. (2014b). Inferential Statistics Modelling and Claim re-assessment. ICCEMS Conference Proceedings: ISBN: 978-967-11414-7-2; 835-840. http://www.iccems.com/2014/ICCEMSProcAll.pdf.

[15] Mordecai, A. (2003). Nonlinear Programming: Analysis and Methods. Dover Publishing. ISBN 0-486-43227-0.

[16] Rao, C.R. (1997) Statistics and Truth: Putting Chance to Work, World Scientific. ISBN 981-02-3111-3.

[17] Rochoxicz, John A. Jr. (2011). Bootstrapping Analysis, Inferential Statistics and EXCEL. Spreadsheets in Education (eJSiE). 4(3).

[18] Ruszczyński, A. (2006). Nonlinear Optimization. Princeton, NJ: Princeton University Press. ISBN 978-0691119151.

[19] Schneider L. and McCuen, R.H. (2005). Statistical Guidelines for Curve Number Generation. J. Irrig. Drain. Eng,131, pp.282--290.

[20] Wright, D.B. (1997). Understandng statistics: an introduction for the social sciences. London, Sage.

[21] Young, G.A., Smith, R.L. (2005). Essentials of Statistical Inference, CUP. ISBN 0-521-83971-8.