Application of a Modified g -Parameter Prior $\left(\frac{k_j}{n^5}\right)$ in Bayesian Model Averaging to Water Pollution in Ibadan

O. B. Akanbi   Saheed A. Afolabi[*]
Department of Statistics, University of Ibadan, Ibadan, Nigeria

Abstract
A special technique that measures the uncertainties embedded in model selection processes is Bayesian Model Averaging (BMA) which depends on the appropriate choices of model and parameter priors. Inspite the importance of the parameter priors' specification in BMA, the existing parameter priors give exitremely low Posterior Model Probability (PMP). Therefore, this paper elicits modified g-parameter priors to improve the performance of the PMP and predictive ability of the model with an application to the Water Pollution of Asejire in Ibadan. The modified g-parameter priors $g_j = \left(\frac{k_j}{n^a}\right), \forall\ a = 3, 4, 5$ established the consistency conditions and asymptotic properties using the models in the literature. The results show that the PMP with the best prior ($g_j = k_j/n^5$) had the least standard deviations (0.0411 at n=100,000 and 0:000 at n=1000) for models 1 & 2 respectively; and had the highest posterior means (0.9577 at n=100,000 and 1.000 at n=1000) for models 1 & 2 respectively. The point and overall predictive performances for the best prior were 2.357 at n=50 and 2.335 at n=100,000 when compared with the BMA Log Predictive Score threshold of 2.335. Applying this best g-parameter prior in modeling the Asejire river, it indicates that the dissolved solids (mg/l) and total solids (mg/l) are the most important pollutants in the river model with their PIP of 6.14% and 6.1% respectively.
Keywords: Posterior Inclusion Probability (PIP), Log-Predictive Score, Model Uncertainty, Dissolved Solids
DOI: 10.7176/JEES/9-11-06
Publication date: November 30th 2019

1.Introduction
Over the years in Nigeria, environmental problem is a great issue especially in the Southern part of the country where oil is spilled into water to cause water pollution. The people of the area are adversely affected with one environmental issue or the other. Previous researches on environment in Nigeria involve the classical approach. To this end, there is prior knowledge about challenges facing the community. I am now motivated to apply Bayesian Analysis through prior elicitation so as to form likelihood in such a way to give a compromise and update of knowledge in pattern of the Posterior using Bayesian Model Averaging (BMA). Bayesian Model Averaging (BMA) is a method that measures the uncertainties embedded in the model selection processes which depends on the appropriate choices of model and parameter priors. By averaging over many different competing models, BMA incorporates model uncertainty into conclusions about parameters and prediction. BMA approach allows the assessment of the predictive skill of a model. Akanbi, (2016) contributed that a composite inference that takes account of model uncertainty can be made in a simple and formally justifiable way. BMA is the method that has been proposed for handling some applications that are very large numbers of models. In BMA, elicitation of priors can be of two forms which are: model and parameter priors. Model priors can be fixed, random, uniform or even custom prior inclusion probability while the parameter priors also knowns as Zellner can also be fixed, empirical Bayes (local) or hyper g prior.

The Zellner g- structure in the parameter prior is expected to be as small as possible such that consistency of the true posterior model probability holds, Zellner, (1986). Fernandez, Ley and Steel (2001a) improved this work based on the priors, Akanbi, (2016) gave an extension in eliciting additional five g-parameter priors. Therefore this research is being undertaken so as to serve as an extension to the literatures on g-parameter prior elicitation in the BMA approach to normal linear regression model based on the increment in prior information with the number of regressors in the model. Hence, the modified parameter prior, $g_j = k_j/n^a$ (a=3, 4 and 5) combined with the uniform model prior is elicited for this study.

2.Bayesian Model Averaging Framework
Suppose a linear model structure of n-independent random samples from a normal regression, with y being the dependent variable, X is the independent variable, ϕ_{0y} a constant, ϕ_y the coefficients and ε a normal iid error term with variance h_j^{-1} $\left[\varepsilon\quad N(0, h_j^{-1})\right]$ with Model j (M_j); j=1,2,3,…,M.

$$M_j : y_j = \phi_j X + \varepsilon \qquad (1)$$

If X contains K potential variables, this means estimating 2^K variable combinations and thus 2^K is given thus;

$$j = 1, 2, ..., M(M = 2^K); 0 < k_j < K \qquad (2)$$

Where k_j is the number of regressors for model j and K is the total number of regressors in the model.

The model weights for this averaging stem from posterior model probabilities (PMP) that arise from Bayes; theorem:

$$P(M_j / y, X) = \frac{P(y / M_j, X)P(M_j)}{P(y / X)} \qquad (3)$$

The integrated likelihood of the model is given thus;

$$P(y / X) = \sum_{b=1}^{2^K} P(y / M_b, X)P(M_b) \qquad (4)$$

The marginal likelihood of the model is given thus;

$$P(y / M_j, X) = \int_0^\infty P(y / \phi_0, \phi_j, h, M_j)P(\phi_0, \phi_j, h / M_j)\, d\phi_0 \, d\phi_j \, dh \qquad (5)$$

Thus, the model weighted posterior distribution for any statistic ϕ.

$$P(\phi_j / y, X) = \sum_{b=1}^{2^K} P(\phi_j / M_j, y, X)P(M_j / X, y) \qquad (6)$$

BMA gets the Posterior Inclusion Probability (PIP) of an explanatory variable by summing the Posterior Model Probabilities across those models that contain the explanatory variable. By comparing two models (js) using Bayes Factor, then we have:

$$B_{js} = \left(\frac{g_j}{g_j+1}\right)^{k_j/2} \left(\frac{g_s}{g_s+1}\right)^{k_s/2} \left[\frac{\frac{1}{g_s+1} y'RX_s y + \frac{g_s}{g_s+1}(y - \overline{y}_{i_n})'(y - \overline{y}_{i_n})}{\frac{1}{g_j+1} y'RX_j y + \frac{g_j}{g_j+1}(y - \overline{y}_{i_n})'(y - \overline{y}_{i_n})}\right]^{(n-1)/2} \quad if k_j; k_s \geq 1 \qquad (7)$$

2.1 Priors in BMA

The model prior $P(M_j)$ is specified by the researcher which should reflect the prior belief about the model. Though there are other model priors such as binomial, beta-binomial and custom prior inclusion probabilities but for this research, the uniform model prior was used such that $P(M_j)\alpha 1$ in the below expression:

$$P(M_j) = \frac{1}{2^K}; P(M_j) > 0 \qquad (8)$$

And

$$\sum_{j=1}^{M} P(M_j) = 1 \qquad (9)$$

Following the rule of thumb as Zellner, (1986) assumed that covariance of the prior should be proportional to covariance expression $(X_j^{*'} X_j^{*})^{-1}$ of the posterior derived from the data, we have:

$$P(\phi_0) \, \alpha \, 1 \qquad (10)$$

The probability for precision is

$$P(h) \, \alpha \, \frac{1}{h} \qquad (11)$$

Thus, the parameter prior is;

$$P(\phi_j / h) \quad N(0_K, h^{-1}[g_j X_j^{*'} X_j^{*}]^{-1}) \qquad (12)$$

Table 1: Summary of all existing Elicited g-Parameter Priors

Prior	Specification	Description	Source
I	UIP	Similar to the Unit Information Prior but with mean zero instead of MLE.	FLS, (2001a)
Ii	$g_j = 1/n$	The prior contains information approximately equal to that contained in a single typical observation. The resulting PMP are closely approximated by the Schwarz Criterion, BIC	Raftey, (1995)
Iii	$g_j = k_j/n$	They assign more information to the prior as regressors increases in the model, i.e. they induce more shrinkage in φ_j (to the prior mean of zero) as the number of regressors grows	FLS, (2001a)
Iv	$g_j = K^{1/k_j}/n$	The prior information decreases with the number of regressors in the model	FLS, (2001a)
V	$g_j = \sqrt{1/n}$	They chose a smaller asymptotic penalty term for large models than in Schwarz criterion	FLS, (2001a)
vi	$g_j = \sqrt{k_j/n}$	They induced more shrinkage as the number of regressors grows.	FLS, (2001a)
vii	$g_j = 1/(\max[n, K^2])$	They preferred prior of Fernandez, Ley an Steel (2001), a mix of Priors	FLS, (2001a)
viii	$g_j = 1/(\ln n)^3$	They choose this to mimic the Hannan-Quinn criterion with $C_{HQ} = 3$ as n becomes large.	Hannan-Quinn, (1979)
Ix	$g_j = \ln(k_j + 1)/\ln n$	This decreases slower with sample size to have asymptotic convergence of InBjs to the Hannan-Quinn criterion with $C_{HQ} = 1$	Hannan-Quinn, (1979)
X	$g_j = \dfrac{\delta_\gamma^{1/k_j}}{\left(1 - \delta_\gamma^{1/k_j}\right)}$	This was suggested by Laud and Ibrahim (1996) by using a natural conjugate prior structure, subjectively elicited through predictive implications.	Laud and Ibrahim, (1996)
xi	$g_j = 1/k^2$	This prior was suggested by the Risk Inflation Criterion (RIC) of Foster and George (1994)	Foster and George, (1994)
xii	$\beta \quad N(\mu, \sigma^2 V)$	Data dependent prior, $\varphi = 2.85$, V =2.58, λ = 0.28; if the R^2 of the full model is less than 0.9, and φ=9.2, V =0.2, λ=0.1684 if the R^2 if the full model is greater than 0.9.	Raftey *et al*., (1997)
xiii	$g_j = 1/n^2$	Prior to capture information for fast increasing sample sizes.	Olubusoye & Akanbi (2015)
Xiv	$g_j = \sqrt{k_j}/n$	Prior to capture information for reducing number of regressors in a model compared to the sample size.	Akanbi (2016)
Xv	$g_j = k_j/n^2$	Prior to capture information for fast increasing sample sizes compared to the number of regressors in a model.	Akanbi (2016)
Xvi	$g_j = k_j^2/n$	Prior to capture information for fast increasing number of regressors in a model compared to the sample size.	Akanbi (2016)
Xvii	$g_j = 3/(\ln n)^3$	Prior to capture reduction of information by reducing the sample sizes but with a higher value of the numerator compared with the FLS. Its asymptotic convergence is Hannan-Quinn Criterion with level CHQ = 3	Akanbi (2016)

Source: Akanbi (2016)

3.A Modified g-Parameter Prior

g-class priors elicitation in BMA needs some basic conditions to follow such as non-negativity, Consistency and Asymptotic properties. The g specification should as well meet certain criteria for consistency of posterior model probabilities and the convergence of the Bayes factor as stated in Fernandez et al (2001a) (FLS). Though, this research is to improve the modified g- parameter priors (g = k_j/n and k_j/n^2) by FLS, (2001a) and Akanbi, (2016) respectively. The model Ms $\in$ M generates the sample 'y', the data throughout this section.

$$y_s = \phi_{0i_n} + X_s \phi_s + \varepsilon_s; \varepsilon_s \quad N(0, h_s^{-1}) \tag{13}$$

Thus:

$$P\lim_{n\to\infty} p(M_s / y, X) = 1 \text{ and } P\lim_{n\to\infty} p(M_j / y, X) = 0; \forall M_j \neq M_s \tag{14}$$

Nothing that, the first probability limit is with the respect to the true model Ms.

The g-parameter prior takes the functional form of:

$$g_j = \frac{t_1(k_j)}{t_2(n)} \text{ with } \lim_{n\to\infty} t_2(n) = \infty \tag{15}$$

Where, $t_1(k_j)$ is the numerator function, in most cases a constant or number of regressors in the model, $t_2(n)$ is the denominator function, usually the sample size used for simulation procedure and $t_2'(n)$ is the first order derivative of the function $t_2(n)$

Given, the assumption that Ms generates the Data, then if the following conditions

(a) $\lim_{n\to\infty} \dfrac{t_2'(n)}{t_2(n)} = 0$

(b) $\lim_{n\to\infty} \dfrac{n}{t_2(n)} \in [0, \infty)$

(c) $t_1(-)$ is an increasing function

Now, we examine the conditions mentioned above with regard to our modified g_j prior;

$$g_j = \frac{t_1(k_j)}{t_2(n)} = \frac{k_j}{n^a} \forall a = 3, 4, 5 \tag{16}$$

Then, the conditions are satisfied as established below.

(a) $\lim_{n\to\infty} \dfrac{t_2'(n)}{t_2(n)} = \lim_{n\to\infty} \dfrac{an^{a-1}}{n^a} = \lim_{n\to\infty} \dfrac{an^a n^{-1}}{n^a} = \lim_{n\to\infty} \dfrac{a}{n} = 0$

(b) $\lim_{n\to\infty} \dfrac{n}{t_2(n)} = \lim_{n\to\infty} \dfrac{n}{n^a} = 0 \in [0, \infty)$

(c) $t_1(-) = k_j$ (constant) is a non decreasing function.

Thus, the seven Asymptotic properties of the modified g- parameter prior are now derived as follows:

Case i: Distribution of the Modified Parameter Prior

$$P(\phi_j / h) \quad f^{k_j} N\left(0, h^{-1}\left[\frac{k_j}{n^a} X_j^{*'} X_j^*\right]^{-1}\right) \tag{17}$$

Case ii: Posterior Probability of the Parameter using the Modified g_j

$$P(\phi_j / y, M_j) \quad f^{k_j}\left(a_3, \overline{\phi}_{j3}, V_3(\phi_j / y, M_j)\right) \tag{18}$$

where

$$Mean = \overline{\phi}_{j3} = \left[\left(1 + \frac{k_j}{n^a}\right) X_j^{*'} X_j^*\right]^{-1} X_j^{*'} y; 3 \leq a \leq 5 \tag{19}$$

and

$$V_3(\phi_j / y, M_j) = \frac{\left[\left(1 + \frac{k_j}{n^a}\right)^{-1} y' R X_j y + \frac{k_j}{n^a + k_j}(y - \overline{y}_{i_n})'(y - \overline{y}_{i_n})\right]\left[\left(1 + \frac{k_j}{n^a}\right) X_j^{*'} X_j^*\right]^{-1}}{a_3 - 2}; a_3 > 2 \tag{20}$$

Case iii: Marginal Likelihood of Model j using the Modified g_j

$$P(M_j / y, X) = \frac{\left(\frac{k_j}{n^a + k_j}\right)^{k_j/2} \left[\left(1 + \frac{k_j}{n^a}\right)^{-1} y'RX_j y + \frac{k_j}{n^a + k_j}(y - \bar{y}_{i_n})'(y - \bar{y}_{i_n})\right]^{-(n-1)/2}}{\sum_{j=1}^{M} P(y, X / M_j)}$$ (21)

Case iv: Bayes Factor for Models (j,s) using the Modified g_j

$$B_{js} = \left(\frac{k_j}{n^a + k_j}\right)^{k_j/2} \left(\frac{k_s}{n^a + k_s}\right)^{k_s/2} \left[\frac{\left(1 + \frac{k_s}{n^a}\right)^{-1} y'RX_s y + \frac{k_s}{n^a + k_s}(y - \bar{y}_{i_n})'(y - \bar{y}_{i_n})}{\left(1 + \frac{k_j}{n^a}\right)^{-1} y'RX_j y + \frac{k_j}{n^a + k_j}(y - \bar{y}_{i_n})'(y - \bar{y}_{i_n})}\right]^{(n-1)/2}$$ (22)

Case v: Posterior Model Probability
The Mean and Variance-covariance Matrix are given thus;
Mean

$$E(\phi_j / y, M_j) \equiv \bar{\phi}_j = \bar{V}_j X_j^{*'} y$$ (23)

Covariance

$$V(\phi_j / y, M_j) = \frac{d \bar{s}_j^2}{d - 2} \bar{V}_j; d > 2$$ (24)

Where

$$\bar{V}_j = \left[\left(1 + \frac{k_j}{n^a}\right) X_j^{*'} X_j^*\right]^{-1}$$ (25)

And

$$\bar{s}_j^2 = \frac{\left(1 + \frac{k_j}{n^a}\right)^{-1} y'RX_j y + \frac{k_j / n^a}{k_j / n^a + 1}(y - \bar{y}_{i_n})'(y - \bar{y}_{i_n})}{d}$$ (26)

Where $RX_j = I_n - X_j^*(X_j^{*'} X_j^*)^{-1} X_j^{*'}$

Case vi: Relationship of the Modified g_j to Information Criteria
Since $t_2(n) = n^a$, then we have:

$$P\lim \frac{\ln B_{js}}{\frac{n}{2} \ln\left(\frac{y'RX_s y}{y'RX_j y}\right) + \frac{k_s - k_j}{2} \ln(n^a)}$$ (27)

$$P\lim \frac{\ln B_{js}}{S_{js}}$$ (28)

Case vii: Predictive Distribution for Model using the Modified g_j

$$P(y_w / X_w^*, D^E, X_j^*) =$$

$$\sum_{j=1}^{M} f_s \left[y_w / n - 1, \bar{y} + \left(1 + \frac{k_j}{n^a}\right)^{-1} X_{w,j}^* \phi_j^*, \frac{n-1}{q_j^*}\left(1 + \frac{1}{n} + \left(1 + \frac{k_j}{n^a} X_{w,j}^{*'}(X_j^{*'} X_j^*)^{-1} X_{w,j}^*\right)^{-1}\right)\right] P(M_j / D^E, X)$$ (29)

Thus, the Log predictive score (LPS) is

$$\text{LPS}(X_w^*, D^Q, X_j^*) = -\frac{1}{u} \sum_{j=1}^{u} \ln P(y_{wj} / X_w^*, D^Q, X_j^*); D^E = u$$ (30)

4.Simulation and Analysis

The concept of simulation experiment here are borrowed from the literature of Bayesian Model Averaging like Raftery, Madigan and Hoeting (1997); Fernadez et al., (2001a); Lee and Steel (2007a); Eicher et al., (2009) and Akanbi, (2016). According to their performed simulations, a design matrix Z for the regressors is an n×K, K = 15 is a fixed number of regressors for a sample size n, such that (z(1), z (2), · · · , z (10)) are drawn from N(0,1) and the subsequent five columns (z(11), · · · , z (15)) are built standard from;

$$(z_1, ..., z_5) \, (0.3 \; 0.5 \; 0.7 \; 0.9 \; 1.1)' \, (1 \; 1 \; 1 \; 1 \; 1) + \varepsilon \; ; \; \varepsilon \quad N(0,1) \qquad (31)$$

Leading to matrix $X^* = (X_1^*, ..., X_{15}^*)$ which fulfills $I'X^* = 0$ using the models:

Model 1

$$y = 4 + 2X_{(1)}^* - X_{(5)}^* + 1.5X_{(7)}^* + X_{(11)}^* + 0.5X_{(13)}^* + \eta \qquad (32)$$

Model 2 (Null Model)

$$y = 1 + \eta, \eta \quad N(0, \sigma^2 \equiv 6.25) \qquad (33)$$

It is indicated that Model 1 is explained by a more or less realistic situation where one third of the regressors intervene, while Model 2 is an extreme case without any relationship between regressors and response. In this analysis, a uniform prior is used over the model space M using MCMC of 50,000 recorded drawings after a burn-in 20,000 drawings and sample sizes of n=50,100,1000,10000,100000 with the model prior;

$$P(M_j) = \frac{1}{2^K}; j = 1, 2, 3, ..., K \qquad (34)$$

5.Results and Discussions

5.1Convergence and Implementation

To examine the convergence of the chain, the empirical (MCMC) and the exact (Bayes factor) are compared. Though, the results are reported based on Bayes factors, the chain is run long enough to have PMP almost equal to those exact results. An important tool to assess this convergence is the correlation coefficient between these two components (Bayes Factors and Empirical relative frequencies of model visited).

5.2Posterior Model Inference (PMI)

The Posterior Probability assigned to the model that generated the data is one of the main indicators of performance of the Bayesian Methodology. It is expected that the true model should be high for small or moderate values of *n* that are likely to occur in practice. Generally, the posterior probability of this true model converges to 1 for large samples. The motive of any model used is to visit the only the true model which is one (1), meaning that; the smaller the visited model, the better it is. The Quartiles of the ratio between the posterior probability of the correct model and the highest posterior probability of the next model, in most cases this ratio tends to be far above unity to confirm the certainty of the true model.

Table 2: Posterior Probability for Model 1 using the Modified **g**-Parameter Priors

PMI (1)	Priors	n=50		n=100		n=1000		n=10000		n=100000	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PMP	k_j / n^3	0.5378	0.1994	0.7309	0.1985	0.8206	0.1645	0.8089	0.2002	0.9343	0.0631
	k_j / n^4	0.6095	0.2167	0.8205	0.1733	0.7958	0.1994	0.7432	0.1984	0.9460	0.0521
	k_j / n^5	0.6819	0.2150	0.8145	0.1836	0.7678	0.2116	0.8734	0.1657	0.9577	0.0411
Model Visited	k_j / n^3	4658.2	2214.7	2480.8	1840.8	1119.6	1289.1	1432.53	1634.8	321.61	524.14
	k_j / n^4	3420.6	2150.8	1388.5	1544.9	1618.7	1674.9	2060.61	1787.3	201.14	182.42
	k_j / n^5	2804.5	2130.0	1424.7	1499.1	2017.0	1874.0	1186.93	1595.9	46.31	132.18
		Q₁	Q₃	Q₁	Q₃	Q₁	Q₃	Q₁	Q₃	Q₁	Q₃
Quartile Models Ratio	k_j / n^3	0.45	0.87	0.61	0.96	0.74	0.95	0.7	0.98	0.75	0.99
	k_j / n^4	0.45	0.81	0.67	0.97	0.63	0.97	0.57	0.93	0.77	1.00
	k_j / n^5	0.49	0.89	0.65	0.97	0.57	0.96	0.77	0.99	0.8	6.3

It can be affirmed from the table above that as sample size n increases, posterior probability of this true model

converges to 1 whereby the best modified g-parameter prior ($g_j = k_j / n^5$) was concluded to be the best for the Model 1 with the estimated value of $0.9577 \cong 1$. From the records means and standard deviations of the number of visited models in the model 1 with $50 \leq n \leq 100,000$ of sample sizes, it can be deduced that the g-parameter prior ($g_j = k_j / n^5$) gives the best result for the Model 1 with the least value of 46.31 when n = 100, 000 (large). The Quartiles of ratio of the true model 1 posterior probability established the best prior with Q_3 value of 6.3.

Table 3: Posterior Probability for Model 2 using the Modified **g**-Parameter Priors

PMI (1)	Priors	n=50		n=100		n=1000		n=10000		n=100000	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PMP	k_j / n^3	0.8342	0.1086	0.9331	0.0467	0.9988	0.0016	0.9999	0.0003	1.000	0.000
	k_j / n^4	0.9688	0.0438	0.9896	0.0363	0.9999	0.000	1.000	0.000	1.000	0.000
	k_j / n^5	0.9963	0.0081	0.9991	0.0048	1.000	0.000	1.000	0.000	1.000	0.000
Model Visited	k_j / n^3	1720.8	1149.1	671.58	471.43	32.93	15.37	13.69	3.94	12.68	2.93
	k_j / n^4	316.58	419.85	121.01	340.47	13.43	1.99	12.73	1.48	11.28	1.62
	k_j / n^5	60.01	83.1	24.72	43.97	12.81	1.52	12.71	1.46	10.41	1.32
		Q_1	Q_3	Q_1	Q_3	Q_1	Q_3	Q_1	Q_3	Q_1	Q_3
Quartile Models Ratio	k_j / n^3	2.23	8.50	2.9	19.1	1.4	23.2	12.5	236.2	13.1	235.8
	k_j / n^4	11.4	56.5	13.2	173.4	1.4	23.2	12.5	234.7	13.8	240.2
	k_j / n^5	64.7	488	7.59	183.3	1.4	23.2	12.5	234.7	14.0	241

It is indicated from the table above that as sample size n increases, posterior probability of this true model converges to 1 whereby the best g-parameter prior ($g_j = k_j / n^5$) was concluded to be the best for the Model 2 with the estimated value of exactly 1 from when n = 1, 000 to n = 100, 000. The means and the standard deviations of the number of visited models in the model 2 with $50 \leq n \leq 100,000$ of sample sizes established that the g-parameter prior ($g_j = k_j / n^5$) gives the best result for the Model 2 with the least value of 10.41 when n = 100, 000. From the quartiles of the ratio between the posterior probability of the correct model and the highest posterior probability of the next model in the Model 2, it is highly shown that all the g-parameter priors ($g_j = k_j / n^a$; $\forall a = 3, 4, 5$) ascertained the true model 2 with the highest values range from Q3 = 8.5 to Q3 = 241 when n = 50 to n = 100, 000 as it far above unity.

5.2.1 Posterior Inclusion Probability (PIP)

This section presents the means and standard deviations of the posterior probabilities of including each of the regressors (1, 5, 7, 11 and 13) as indicated in the above equation of model 1. It is expected that as sample size (n) increases, those means of these regressors also tend to 1. It gives the degree of errors when the posterior model probability is allocated to the wrong sampling model.

Table 4: Means and S.Ds of the Posterior Probabilities of Model 1 Regressors with n = 50, 1000 and 100,000

Priors		$g_j = k_j / n^3$		$g_j = k_j / n^4$		$g_j = k_j / n^5$	
	Regressors	Mean	S.D	Mean	S.D	Mean	S.D
n=50	*1	0.933	0.164	0.809	0.295	0.637	0.397
	2	0.017	0.164	0.003	0.295	0.001	0.397
	3	0.038	0.164	0.011	0.295	0.003	0.397
	4	0.060	0.164	0.024	0.295	0.010	0.397
	*5	0.07	0.164	0.012	0.295	0.001	0.397
	6	0.014	0.164	0.002	0.295	0.000	0.397
	*7	0.660	0.164	0.419	0.295	0.216	0.397
	8	0.013	0.164	0.002	0.295	0.000	0.397
	9	0.012	0.164	0.002	0.295	0.000	0.397
	10	0.021	0.164	0.003	0.295	0.000	0.397
	*11	0.637	0.164	0.554	0.295	0.426	0.397
	12	0.051	0.164	0.036	0.295	0.035	0.397
	*13	0.200	0.164	0.167	0.295	0.123	0.397
	14	0.040	0.164	0.027	0.295	0.024	0.397
	15	0.046	0.164	0.027	0.295	0.017	0.397
n=1,000	Regressors	Mean	S.D	Mean	S.D	Mean	S.D
	*1	0.992	0.045	0.936	0.183	0.743	0.354
	2	0.000	0.045	0.000	0.183	0.000	0.354
	3	0.001	0.045	0.000	0.183	0.000	0.354
	4	0.007	0.045	0.003	0.183	0.000	0.354
	*5	1.000	0.005	1.000	0.003	1.000	0.354
	6	0.000	0.045	0.000	0.183	0.000	0.354
	*7	0.771	0.005	1.000	0.003	1.000	0.354
	8	0.000	0.045	0.000	0.183	0.000	0.354
	9	0.000	0.045	0.000	0.183	0.000	0.354
	10	0.000	0.045	0.000	0.183	0.000	0.354
	*11	0.850	0.045	0.801	0.183	0.636	0.354
	12	0.008	0.045	0.011	0.183	0.004	0.354
	*13	1.000	0.005	1.000	0.003	0.093	0.354
	14	0.012	0.045	0.007	0.183	0.002	0.354
	15	0.007	0.045	0.017	0.183	0.002	0.354
n=100,000	Regressors	Mean	S.D	Mean	S.D	Mean	S.D
	*1	1.000	0.000	1.000	0.000	1.000	0.000
	2	0.000	0.000	0.000	0.000	0.000	0.001
	3	0.000	0.000	0.000	0.000	0.000	0.001
	4	0.000	0.000	0.000	0.000	0.000	0.001
	*5	1.000	0.000	1.000	0.000	1.000	0.000
	6	0.000	0.000	0.000	0.000	0.000	0.001
	*7	1.000	0.000	1.000	0.000	1.000	0.000
	8	0.000	0.000	0.000	0.000	0.000	0.001
	9	0.000	0.000	0.000	0.000	0.000	0.001
	10	0.000	0.000	0.000	0.000	0.000	0.001
	*11	1.000	0.000	1.000	0.000	1.000	0.000
	12	0.000	0.000	0.000	0.000	0.000	0.001
	*13	1.000	0.000	1.000	0.000	1.000	0.000
	14	0.000	0.000	0.000	0.000	0.000	0.001
	15	0.000	0.000	0.000	0.000	0.000	0.001

It is indicated from the Table above that regressors 1, 7 and 11 are close to 1 while other regressors 5 and 13 misbehaved with sample size n = 50 for the Model 1 of the three g-parameter priors ($g_j = k_j / n^a$; $\forall a = 3, 4, 5$); with sample size n = 1,000 for the Model 1, all regressors (1, 5, 7, 11 and 13) are close to 1 but regressors (5, 7 and 13) are equally 1 in terms of mean of the three g-parameter priors ($g_j = k_j / n^a$; $\forall a = 3, 4, 5$) and for n=100,000 it is shown that all the regressors (1, 5, 7, 11 and 13) are equal to 1 in terms of mean for the Model 1 of the three

g-parameter priors ($g_j = k_j / n^a$; $\forall a = 3, 4, 5$).

This establishes that highest sample size yields the best result in this case and hence, the best modified g - parameter prior is $g_j = k_j / n^5$.

5.3 Predictive Inference (PI)

This section deals with predictive inference via the Log predictive Score (LPS) in terms of point and overall predictions for some samples based on the values of the regressors X_w^*; for model 1, w=19 different vectors of the K = 15 regressors. The below Table depicts the predictions via log predictive score (LPS) for model 1 via the 100 samples (y, X*).

Table 5: Medians of LPS (X_w^*, y, X^*): Point and Overall Predictions using the modified g-parameter priors ($g_j = k_j / n^a$; $\forall a = 3, 4, 5$)

		n=50	n=100	n=1000	n=10,000	n=100,000
	Priors	X_{min}^*	X_{min}^*	X_{min}^*	X_{min}^*	X_{min}^*
Point Prediction	$g_j = k_j / n^3$	2.30213	2.40094	2.24929	2.39517	2.42331
	$g_j = k_j / n^4$	2.35721	2.40094	2.24929	2.39517	2.42331
	$g_j = k_j / n^5$	2.35721	2.40094	2.38751	2.40113	2.42331
Overall Prediction	$g_j = k_j / n^3$	2.213	2.521	2.187	2.433	2.331
	$g_j = k_j / n^4$	2.236	2.513	2.156	2.439	2.331
	$g_j = k_j / n^5$	2.247	2.522	2.166	2.411	2.335

It can be established from the above Table that the vector of regressors that lead to the minimum value for the mean (100 replication) of the sampling model 1 for the modified priors ($g_j = k_j / n^a$; $\forall a = 3, 4, 5$) are all close to the threshold of 2.335 as specified for BMA models, especially when n = 50. In the same vein, the above Table presents the overall predictive performance via the LPS (X_w^*, y, X^*) for the 19 different values of X_w^* and the 100 replications of (y, X*). Obviously, all the elicited g-parameter priors showed well predictive behaviour for n = 100, 000, but the best of all is the modified prior ($g_j = k_j / n^5$) with the exact value of threshold i.e. 2.335 as specified for BMA models.
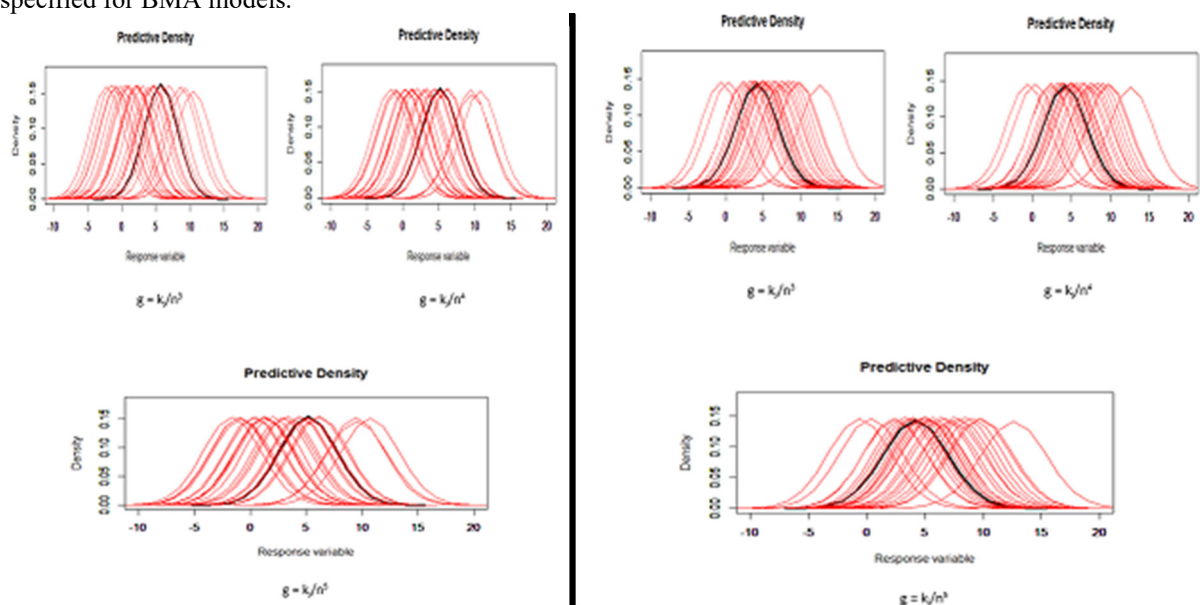

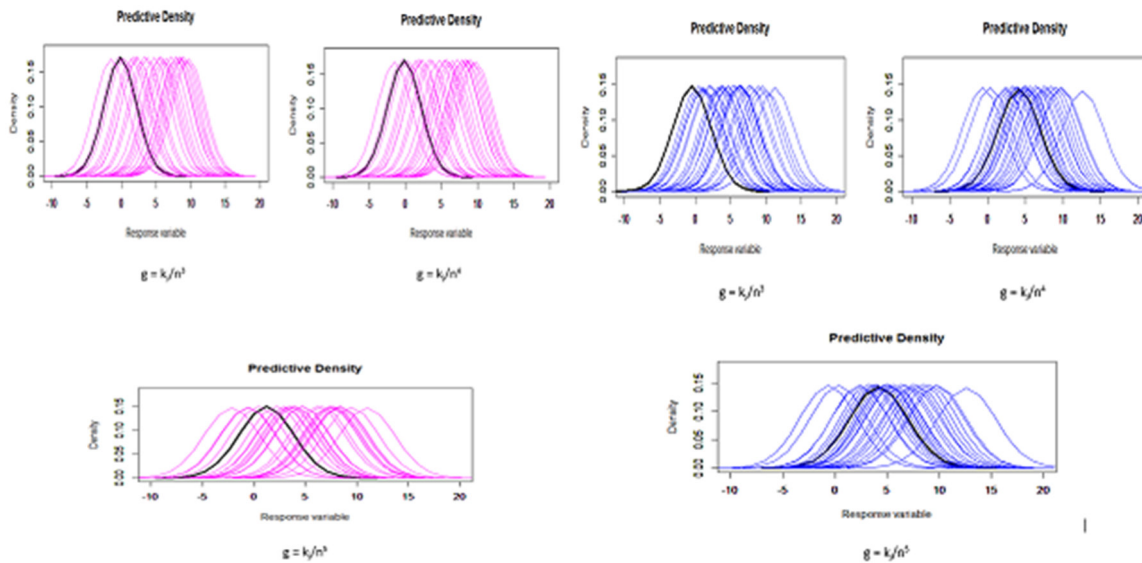Figure 1: Point Prediction with the Modified g-priors for n=50 and n=100

Figure 2: Point Prediction with the Modified g-priors for n=1,000 and n=10,000
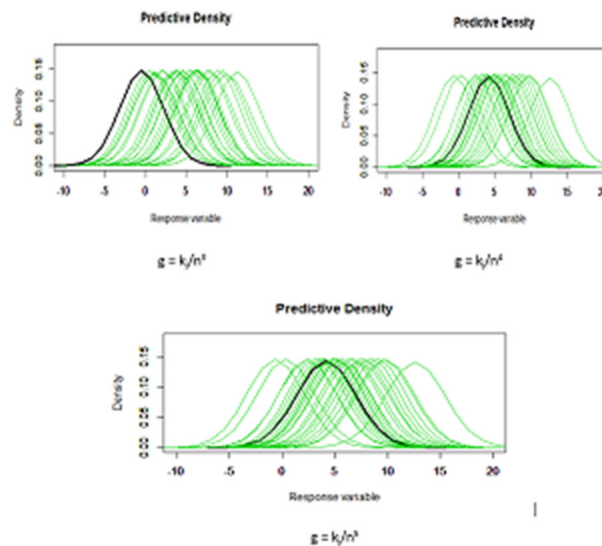

Figure 3: Point Prediction with the Modified g-priors for n=100,000

6. Application of BMA with the Best Modified g-Parameter Prior ($g_j = k_j / n^5$) to Water Pollution in Ibadan

Water pollution is the contamination of water bodies, usually as a result of human activities. Water is considered polluted when unwanted materials with potentials to threaten human and other natural systems find their ways into water sources or reserved fresh water in homes or industries. Therefore, the BMA method is applied to the water pollutants and its pollution level to account for the uncertainties embedded in both the parameters and model using the best modified g-parameter prior, $g_j = k_j / n^5$ with the water pollution level model given below:

$$WPL = \phi_0 + \phi_1 DO + \phi_2 TUR + \phi_3 COL + \phi_4 pH + \phi_5 ALK + \phi_6 TH + \phi_7 CAH + \phi_8 CL + \phi_9 FE + \phi_{10} SI +$$
$$\phi_{11} SOL + \phi_{12} DS + \phi_{13} SS + \phi_{14} COD + \varepsilon \qquad (35)$$

where ε is a stochastic error term, independently and identically distributed as $N(0, \sigma^2)$ with the variables Water Pollution level (WPL)as the regressand, Dissolved Oxygen (DO), Turbidity (TUR), Colour (COL), PH, Alkalinity (ALK), Total Hardness (TH), Calcium Hardness (CAH), Chloride (CL), Iron (FE), Silica (SI), Total Solids (SOL), Dissolved Solids (DS), Total Suspended Solids (SS) and Chemical Oxygen Demand (COD).

Table 6: Posterior Probabilities of Including each of the Regressors (PIP) in the Water Pollution Level

Regressors	PIP	Post Mean	Post S. D.	Cond. P. Sign	Index
DS	0.0614	0.0476	0.1895	1	12
Sol	0.0610	0.0475	0.1918	1	11
Alk	0.0380	0.0220	0.1125	1	5
CaH	0.0031	0.0048	0.0947	1	7
Cl	0.0005	-0.0006	0.0887	0	8
pH	0.0005	29.0408	7.9941	1	4
TH	0.0003	0.9765	1.2543	1	6
DO	0.0000	0.0000	0.0000	NA	1
TUR	0.0000	0.0000	0.0000	NA	2
Col	0.0000	0.0000	0.0000	NA	3
Fe	0.0000	0.0000	0.0000	NA	9
Si	0.0000	0.0000	0.0000	NA	10
SS	0.0000	0.278	0.0000	NA	13
COD	0.0000	0.000	0.0000	NA	14

Table 6 presents the means and standard deviations of the posterior inclusion probabilities (PIP) of each of the regressors in the water pollution level. It is indicated that the dissolved solids (DS) with PIP of 6.14% is very important if modelling water pollution of Asejire River in Ibadan.

Table 7: The MCMC and the Exact Posterior probabilities for the First Best 5 Models

Models	PMP (Exact)	PMP (MCMC)	Predictors
0001	0.0704614	0.0705800	DS
0008	0.0608948	0.0513800	SOL
0204	0.0350687	0.0426200	ALK and DS
0004	0.0240914	0.0250200	DS
0284	0.001588	0.0025200	DS, CaH, ALK

It is shown from the table above that the best model Dissolved Solid (mg/l) has PMP of 7.0% among the 1186 models visited.
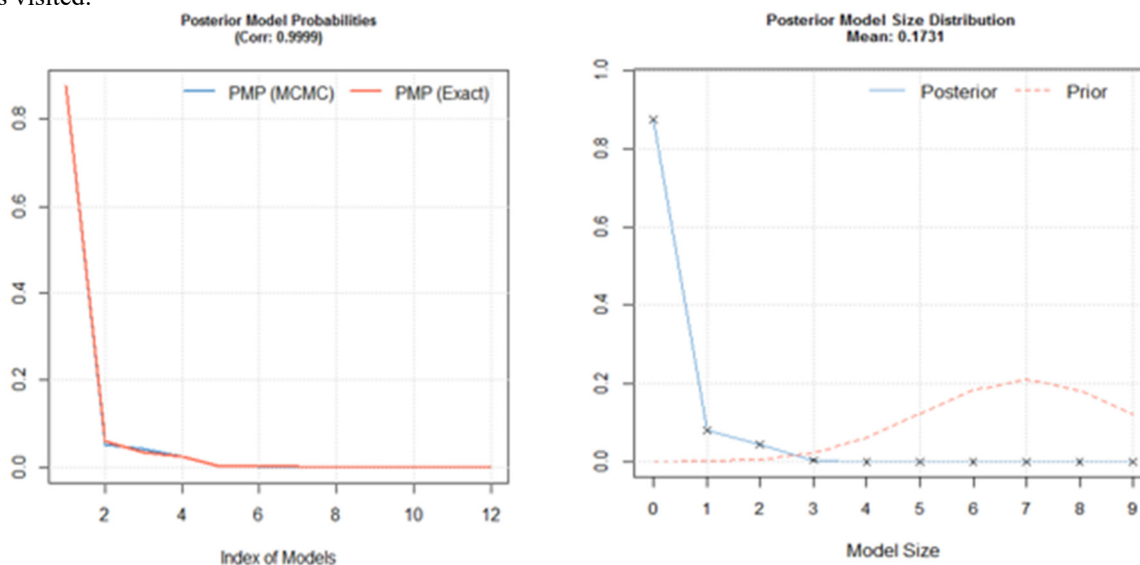

Figure 4: Posterior Model Probabilities and Model Size

It can be observed from the above figure that PMP (Exact) is closed to PMP (MCMC) due to the statistics of shrinkage factor which is exactly 1.
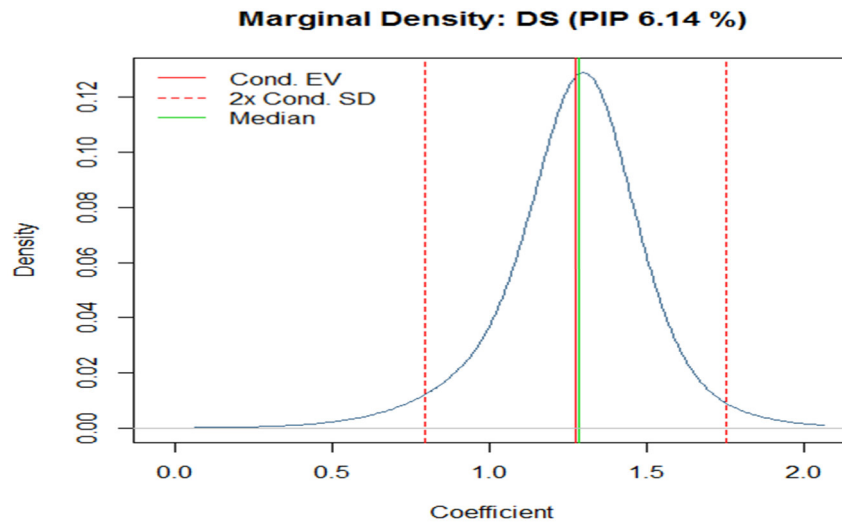
Figure 5: Marginal Density for Dissolved Solid

From the above figure, DS appeared as the most important pollutant in the water pollution model with PIP of 6.14%.
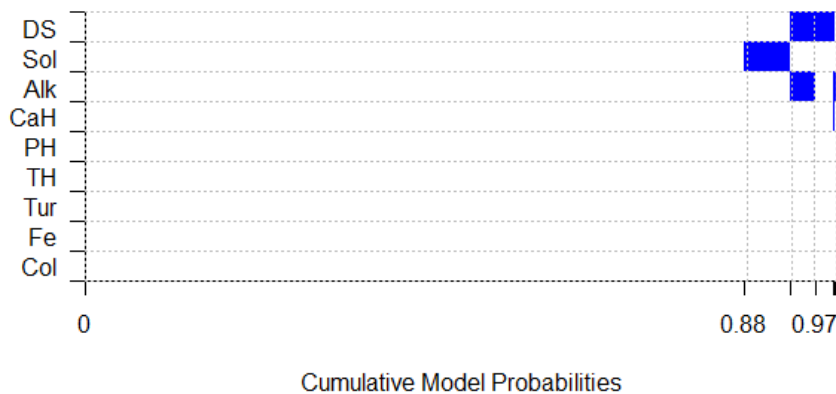

Figure 6: Cumulative Model Probabilities with the Signs of their Regressors

Figure 6 shows the cumulative model inclusion probabilities based on best 14 models. It also depicts the inclusion of a regressor with its sign in the model selection process. This image plot is based on the Bayes factor of the MC^3 simulator. The blue colour means a positive sign. It is confirmed that the selected best model with PMP of 97% includes only the dissolved solids (DS).

7.Conclusion

In this paper, the elicited modified g priors need only the choice of one scalar hyper parameter known g-class. The consistencies conditions and asymptotic properties for the modified g-parameter priors were derived. The empirical results on both posterior model and predictive inferences indicate that the modified prior $g_j = k_j / n^5$ was the best out of the three g modified parameter priors considered in the BMA technique. This implies that, the higher the power of the sample size (n), the more efficient and the g parameter prior. The application of the best g prior to modelling Asejire River shows that the effect of dissolved solids (mg/l) and total solids (mg/l) as water pollutants in Asejire River, Ibadan, Oyo State are very important. Thus, the two water pollutants are recommended in modelling Asejire River and also to used the elicited modified parameter prior, $g_j = k_j / n^5$ combined with a uniform model prior for model selection or Bayesian model averaging in Asojire River model whenever informative prior is not available for both small and large samples.

References

Albert.J. 2007. Bayesian Computation with R. Springer. *ISBN 9780-387-71384-7.*
Clyde, M. and E.I. George. (2004). "Model Uncertainty," Statistical Science 19, 81-94.
Eicher, T.S., C. Papageorgiou and A.E. Raftery (2009) "Determining Growth Determinants:

Default Priors and Predictive Performance in Bayesian Model Averaging", *Working Paper No. 76, Center for Statistics and the Social Sciences, University of Washington, Seattle.*

Fernndez C., E. Ley and M.F.J. Steel. (2001a). "Benchmark Priors for Bayesian Model Averaging", *Journal of Econometrics 100, 381-427.*

Fernndez C., E. Ley and M.F.J. Steel. (2001b). Model Uncertainty in CrossCountry Growth Regressions", *Journal of Applied Econometrics 16, 563-576.*

Foster, D. P. and George, E. I. (1994). The Risk Inflation Criterion for Multiple Regression. *The Annual of Statistics 22, 1947-1975.*

George, E.I. and R.E. McCulloch. (1993). "Variable Selection via Gibbs Sampling", *Journal of the American Statistical Association 88, 881-889.*

Hannan, E. J. and Quinn, B. G. (1979): The Determination of the Order of an Auto-regression. *Journal of the Royal Statistical Society, Series B. 14, 107-114.*

Hoeting, J.A., D. Madigan, A.E. Raftery and C.T. Volinsky. (1999). "Bayesian Model Averaging: A Tutorial", *Statistical Science 14, 382-417.*

Laud and Ibrahim, (1996). A Natural Conjugate Prior Structure, Subjectively Elicited through Predictive Implication.

Ley, E. and Steel, M. F. J., (2011). Mixture of g-Priors for Bayesian Model Averaging with Economic Applications. *Journal of Applied Econometrics. 26(3), 628-647.*

Madigan, D. and A.E. Raftery. (1994). "Model Selection and Accounting for Model Uncertainty in Graphical Models using Occams Window", Journal of the American Statistical Association 89, 1535-1546.

Madigan, D., Gavrin, J. and Raftery, A.E. (1995), "Eliciting Prior Information to Enhance the Predictive Performance of Bayesian Graphical Models", Communications in Statistics, Theory and Methods, 24, 2271-2292.

Olubusoye, O.E and Okewole, D.M (2009). Prior Sensitivity in Bayesian Linear regression Model. *Int. Journal (Sciences) Vol. 3, No. 1, pp21-29.*

Olubusoye, O. E. and Akanbi, O. B. (2014). Bayesian and Classical Estimation Methods in Linear Regression Model: A Monte Carlo Simulation Analysis. *Journal of Science Research 13, 73-83.*

Olubusoye, O. E. and Akanbi, O. B. (2015). On g-Prior Elicitation in Bayesian Model Averaging Approach to Normal Linear Regression Model. Conference Procedings of Mathematics Department, University of Ibadan, Nigeria. In honour of S. A. Iloris 70th Birthday.

Olubusoye, O. E., Akanbi, O. B. and Kazeem, B. A. (2015). Modelling the Nigerian Inflation Process using the Bayesian Model Averaging (BMA) Approach. Proceedings of the 19th Annual Conference of the Africa Region Chapter of the Econometric Society, Ethiopia, III, 117-140.

Biographies

Dr. O. B. Akanbi: He lives in Ibadan, Nigeria. He is a lecturer in Statistics Department, University of Ibadan, Nigeria. His areas of research are Econometrics, Bayesian Inference or Modelling (Bayesian Model Averaging) and Operations research. He has Professional diploma (Statistics), B.Sc. (Statistics), M.Sc. (Statistics) and Ph.D. (Statistics) from the University of Ibadan, Nigeria. Dr. Akanbi has to his credit a number of refereed journal articles and conference papers within and outside Nigeria. He is a member of some statistical societies; International Statistical Institute (ISI); International Biometric Society (IBS) Group Nigeria (GNi); Statistics & Probability African society (SPAS); Nigerian Statistical Association (NSA); and Nigerian Young Statisticians Group (NYSG).

Saheed A. Afolabi: He was born on 23rd December, 1990 in Osogbo, Osun State of Nigeria. He finished his primary and secondary education in 2001 and 2007 respectively, he attended Federal Polytechnic Ede, Osun State between 2009 to 2011 to obtain his National Diploma in Statistics which he was the overall best student of the session, he did a Direct entry to the University of Ibadan, Ibadan, Oyo State, Nigeria to obtain BSc in Statistics, (2014) and MSc in Statistics, (2019) with a PhD grade. He is a member of Nigeria Statistical Association (2017), Professional Statistical Society (2017), International Biometric Society (2018), Royal Statistical Society (2019) and American Statistical Association (2019). The Author's major field is Environmental Statistics but he has published papers in different fields.