Using criterion-based model averaging in two-input multiple response surface methodology problems

Domingo Pavolo* and Delson Chikobvu

Department of Mathematical Statistics and Actuarial Sciences,
School of Natural and Agricultural Sciences,
University of the Free State,
Bloemfontein, South Africa
Email: 2017476603@ufs4life.ac.za
Email: Chikobvu@ufs.ac.za
*Corresponding author

Abstract: Experimental designs in multiple response surface methodology (MRSM) often result in small sample size datasets with associated modelling and model selection problems. Classical model selection criteria are inefficient when using small sample size datasets while the model selection process has inherent uncertainties. Modelling of small sample size datasets below ($10 + k$), where k is the maximum number of regressors inclusive of the intercept, suffers from credibility problems. In this empirical paper, criterion-based frequentist model-averaging (CBFMA) is proposed as a solution to the small sample size problems of modelling MRSM datasets. We also compare the goodness of fit and prediction accuracy of using CBFMA models versus ordinary least squares (OLS) candidate models. Findings suggest that CBFMA models have good fitness to data and predictive accuracy. Also, the small sample size model selection criteria bias problem is improved on. However, in the MRSM context, CBFMA does not directly solve both criterion and response surface uncertainties, and averaged model estimators have mean squared errors that are greater than the best OLS candidate models.

Keywords: multiple response surface methodology; MRSM; experimental design; all possible regression models; frequentist criterion-based model averaging; small sample size datasets; process optimisation.

Reference to this paper should be made as follows: Pavolo, D. and Chikobvu, D. (2022) 'Using criterion-based model averaging in two-input multiple response surface methodology problems', *Int. J. Operational Research*, Vol. 44, No. 1, pp.80–101.

Biographical notes: Domingo Pavolo is a PhD student in the Department of Mathematical Statistics and Actuarial Sciences, School of Natural and Agricultural Sciences, University of the Free State, Bloemfontein South Africa. He received his MEng in Manufacturing Systems and Operations Management in 2005 and MSc in Operations Research in 2003 at the National University of Science and Technology, Zimbabwe. His research interests are in applied industrial statistics and productivity, response surface methodology and statistical quality control.

Delson Chikobvu is a Senior Lecturer in the Department of Mathematical Statistics and Actuarial Sciences, School of Natural and Agricultural Sciences, University of the Free State, Bloemfontein South Africa. His research interests include statistical quality control, response surface methodology and statistical modelling.

1 Introduction

Myers et al. 2016 summarise multiple response surface methodology (MRSM) as simultaneous consideration of multiple responses involving first building an appropriate response surface model for each response and then trying to find a set of operating conditions that in some sense optimises all responses or at least keeps them in desired ranges. MRSM basically involves design of experiments, regression modelling, model selection and multi-response optimisation.

The modelling and model selection problems that an MRSM practitioner encounters include model uncertainty (Schomaker and Heumann, 2018), selection criterion uncertainty and response surface uncertainty (Pavolo and Chikobvu, 2019), regression parameter bias and variability (Mitra et al., 2019), small sample size selection criteria inefficiency (Seghouane and Bekara, 2004), and small sample size modelling credibility (Rawlings et al., 1998). This work proposes strategies for dealing with these problems using criterion-based frequentist model-averaging (CBFMA) and investigates effectiveness using a typical MRSM dataset. The strategies include:

1 avoiding the use of model selection criteria to evade both criterion uncertainty and small sample size bias (Pavolo and Chikobvu, 2019)

2 use of CBFMA to deal with model uncertainty and parameter variability (Zhang and Liu, 2018)

3 arithmetic averaging (Hu et al., 2015) of similar functional form estimators to deal with parameter bias

4 rigour to ensure credibility of MRSM results.

The results of the empirical study suggest that the proposed strategies are effective and CBFMA produces model estimators that have better prediction accuracy, measured by mean squared forecasted error (MSFE), than the 'best' ordinary least squares (OLS) candidate model. However, CBFMA produces model estimators whose fit to data, measured by mean squared error (MSE), is always second to the 'best' OLS candidate model. In the MRSM context, CBFMA is not a direct solution to both criterion uncertainty and response surface uncertainty.

2 Literature review

Khuri (2017) defines an experiment in which a number of responses can be measured for each setting of a group of input variable, $x_1, ..., x_k$, as a multi-response experiment. Myers et al. (2016) summarise MRSM as simultaneous consideration of multiple responses involving first building an appropriate response surface model for each response and then trying to find a set of operating conditions that in some sense optimises all responses or at least keeps them in desired ranges. MRSM experiments are designed to be cost efficient and are expected to provide optimum information. MRSM involves design of experiments, regression modelling, model selection and multi-response optimisation.

Model uncertainty (i.e., which model to use?) is an old problem in statistics (Steel, 2017). Literature presents model selection (MS) and model averaging (MA) as the two major approaches of dealing with model uncertainty (Schomaker and Heumann, 2018).

Alhorn et al., (2019) discusses the problem of constructing efficient experimental designs to address model uncertainty.

MS is an important concept in any data analysis and evaluation of regression models (Emmert-Streib and Dehmer, 2019). MS criteria provide a useful tool in identifying a model of appropriate structure and dimension among a set of candidate models. A selection criterion assesses fitness to data, balance between parameter bias and variability, and prediction accuracy of a fitted model (Cavanaugh and Neath, 2019). A variety of MS criteria have been proposed in the literature for selecting the 'best' model from a set of candidate models but different criteria choose differently (Schomaker and Heumann, 2018); this further introduces the complication of criterion uncertainty (i.e., which criterion to use?). MS does not deal satisfactorily with the problem of the regression parameters of models often having bias and the problem of their standard errors being too small since they do not reflect the uncertainty related to the model selection process (Mitra et al., 2019; Schomaker and Heumann, 2018). In the context of MRSM, MS has its own problems: firstly, response surface uncertainty (Pavolo and Chikobvu, 2019); secondly, the small sample size selection criterion inefficiency problem (Hurvich and Tsai, 1989; MaQuarrie and Tsai, 1998; Seghouane and Bekara, 2004). Rawlings et al. (1998) pointed out the credibility problem associated with regression modelling of small sample size datasets.

The authors propose a way to deal with the problems of criterion uncertainty and small sample size inefficiency, in MRSM, by avoiding the direct use of model selection criteria in model selection. This can be achieved by proceeding to problem solution with candidate models chosen because of response surface conformity and, if the results are different, averaging them. A novel method would be to first combine the candidate models with weights generated from different MS criteria to come up with as many averaged estimator models, which are further, combined using equal weights, i.e. arithmetic model averaging (AMA) since the models will be having the same functional form. The problem of small sample size credibility demands convincing rigour in the solution methodology. The uncertainty of response surfaces makes it imperative to do in-depth response surface analysis and base choice of candidate models on response surface conformity.

Xie (2015) presents model averaging (MA), unlike MS, as a methodology which incorporates model uncertainty by not choosing a best model among competing candidates but, instead, combining them. Yang (2003) suggests that MA tries to harness the best performance among candidate models. Juditsky and Nemirovsky (2000) state that MA aims to produce a model which has a better performance compared to any of the original candidate models. Leung and Barron (2006) add that MA improves the risk in estimation by providing insurance against selecting a poor model. Zhao et al. (2018) suggest that MA aims to achieve a trade-off between efficiency and biases. Schomaker and Heumann (2018) present MA as a successful alternative to the MS. Because of such tempting reasons, the use of MA in research and practice has been growing in other fields of science such as econometrics, ecology, biostatistics etc., while little progress is reported in MRSM where modelling sensitivity of small sample size datasets is an obvious concern.

Nguefack-Tsague (2014) gives another view of the relationship between MS and MA. It is considering MS as one extreme of MA in which one candidate model is accorded a weight of 1 and the rest zeros. In the same vein, the other extreme of MA is the allocation of equal weights to the candidate models (or arithmetic averaging, AMA). AMA allocates the same weight (1/M, where M is the number of candidate models) to estimators. Hu et al. (2015) reported a slightly better performance of AMA over Akaike information criteria (AIC) based weighted averaging. Lin and Kuo (2016) state that AMA is appropriate to use if all the candidate models have similar prediction powers. Okoli et al. (2018) in a study of estimating designs associated with flooding obtained the same performance for AMA and weighted MA when the AIC values of candidate model were almost similar.

MA has developed in two fronts which are Bayesian and frequentist (Zhang et al., 2016) paradigms. Bayesian model averaging (BMA) was the first to gain popularity but is complex. Frequentist model averaging (FMA) has been gaining popularity among practitioners because of its simplicity. Schomaker and Heumann (2018) summarize the developments in FMA into two:

1 criterion-based frequentist model averaging (CBFMA)

2 optimal model averaging (OMA).

The basis of CBFMA is using classical MS criteria to come up with weights for each candidate model. According to Schomaker and Heumann (2018) CBFMA is focused at purely descriptive work and work that helps to identify associations. Considering a simple linear regression model

$$y = \beta_0 + \sum_{j=1}^{k} \beta_i x_i + \varepsilon_i \tag{1}$$

a frequentist average estimator of β is given by

$$\beta = \sum_{j=1}^{k} \mu_i \beta_i \tag{2}$$

with weights satisfying $\mu_i \geq 0$ and $\sum_{j=1}^{k} \mu_i = 1$.

A popular weight choice would be based on the exponential Akaike information criterion (AIC) or Bayesian information criteria (BIC) as shown

$$\mu_i^{AIC} = \frac{e^{\left(\frac{-\Delta AIC_i}{2}\right)}}{\sum_{i=1}^{k} e^{\left(\frac{-\Delta AIC_i}{2}\right)}} \tag{3}$$

or

$$\mu_i^{BIC} = \frac{e^{\left(\frac{-\Delta BIC_i}{2}\right)}}{\sum_{i=1}^{k} e^{\left(\frac{-\Delta BIC_i}{2}\right)}} \tag{4}$$

where one calculates the weighted average $\beta = \sum_{i=1}^{k} \mu_i \beta_i$ from the *i* estimators of β_i.

A general class of random weights considered in CBFMA is given by:

$$\mu\left(\hat{\theta}, \hat{\sigma}^2\right) = \frac{a^{q_i}(n-q)^b\left(\hat{\sigma}^2\right)^c}{\sum_{i=1}^{M} a^{q_i}(n-q)^b\left(\hat{\sigma}^2\right)^c} \tag{5}$$

where $a(> 0)$, $b(\geq 0)$and $c(\leq 0)$, n is sample size, q is the number regressors and $\hat{\sigma}^2$ the MSE.

Most of the weight choices adopted in practice are various forms of the above, of which examples are given in Table 1.

Table 1 Examples of criterion-based FMA weights

#	*Weight*	*a*	*b*	*c*
1	Smoothed AIC	e^{-1}	0	$-n/2$
2	Smoothed BIC	$(n)^{-1/2}$	0	$-n/2$
3	Smoothed RMS	1	1	–1
4	Smoothed GVC	1	2	–1

A seminal paper by Hansen (2007) on OMA based on finding optimal weights that minimize Mallows' Cp criterion (MMA), opened the flood gates to optimal weights research. The Jackknife model averaging (JMA) criterion by Hansen and Racine (2012) focused on accommodating heteroskedasticity. Nguefack-Tsague (2014), however, argued that though optimal weights exist in theory, once estimated they are no longer optimal. Zhang et al. (2016) established the conditions under which Hansen's MMA weights dominated OLS estimators. Zhao et al. (2018) came up with a class of model averaging estimators and its dominant conditions over OLS in finite sample size conditions. Schomaker and Heumann (2018) categorized OMA model estimators as focused on prediction and forecasting and that they should not be included among MA models that are used for description, studying associations and dealing with model uncertainty.

Magnus et al. (2010) state that MA solves the problem of parameter and estimation variability as it calculates a weighted average of each regression coefficient from the set of competing candidate models combining them into a single estimator. The MA weights are determined in such a way that 'better' models receive a higher weight. Schomaker and Heumann (2018) point out that, with MA, both the variance related to the parameters of each model and the variance between the different model estimates is considered. MA, however, does not effectively tackle the problem of parameter bias if the candidate models do not have the same functional form because MA estimators behave like shrinkage estimators hence regression coefficients which belong to variables which are not supported among all candidate models are shrunk and are therefore possibly biased. Literature also warns that in situations where the data supports one particular model uncompetitively, MA may not improve the overall performance of an estimator.

In this paper, CBFMA is investigated in an MRSM environment as a solution to model uncertainty, criterion uncertainty, response surface uncertainty, bias and variability, and MRSM small sample size dataset modelling and model selection problems. The effectiveness of the proposed solutions and strategies to these problems are empirically tested. The authors could not find anywhere in MRSM literature where strategies to deal with these problems are discussed and empirically investigated.

The remainder of this paper is organised as follows. Section 3 presents the methodology used in this empirical investigation. Section 4 presents the results whilst Section 5 discusses the findings from the results. Section 6 concludes and proposes direction of future research.

3 Methodology

In this section, the performance of CBFMA in MRSM is explored using an adopted typical small sample size MRSM dataset.

3.1 The adopted dataset

The experimental design shown in Table 2 is a two-factor central composite design adopted from Pavolo and Chikobvu (2019). These experimental runs were done to determine the best cure times for different conveyor belt thicknesses for a Southern African rubber covered mining conveyor belts manufacturing company to ensure product quality and process productivity.

Table 2 Experimental design and averaged results from the MRSM experiment

Run	*T (time) Min.*	*RT (rubber thickness) mm*	*Ave. adhesion (N/mm)*	*Ave. hardness (^{0}shore A)*
1	16	7.2	10.60	60
2	30	7.2	13.34	63
3	16	22.8	6.20	53
4	30	22.8	12.10	61
5	23	15	11.80	58
6	23	15	12.10	58
7	13	15	6.5	44
8	33	15	13.30	63
9	23	4	13.30	63
10	23	26	3.50	56
11	23	15	12.20	58
12	23	15	12.30	57
13	23	15	12.10	58

The sample size n from the MRSM dataset is 13, the number of experimental runs. The full model number of regressors including the intercept $k = 6$. Now $(n/k) = (13/6) = 2.167$ and this is much less than 40, which shows that this is a typical small sample size case. Also, $(n - k) = (13 - 6) = 7 < 10$, which means results have an obvious modelling credibility problem. The reasons this dataset is chosen are that, in addition to being a typical small sample size MRSM dataset, two factor experiments produce response models that have response surfaces that can be constructed in three dimensions and analysed.

3.2 All possible regression modelling

All possible regression modelling was used to obtain the permutations of OLS response models for the adhesion and hardness responses from Table 2.

3.3 Goodness of fit check using mean squared error

The MSEs of the models were determined and used to estimate and compare the goodness of fit of the models to the MRSM dataset. The formula for MSE is shown below for a sample size *n*.

$$MSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}} \tag{6}$$

where Y_i is the measured response, $\hat{Y}_i$ is the predicted response.

3.4 Response surface analysis and candidate model selection

With prior expectation and knowledge of response surface structure, models with conforming response surfaces were selected from the two sets of the all regression models of the adhesion and hardness responses as candidate models for multi-response optimisation.

3.5 Analysis of the effect of CBFMA on response surfaces

This subsection, in Table 3, summarises the empirical investigation performed on the adhesion and hardness responses sets of 31 response models.

Table 3 Tabulation of summary investigations of CBFMA

#	*ANALYSIS*	*QUESTIONS TO BE ANSWERED*
i)	All the 31 all regressions models of each response are combined into one using CBFMA smoothed weights. Response models are combined using S-AIC, S-AICc, S-BIC, S-RMS, S-GVC, S-HQc, and S-HQ. Data arrays are produced and response surfaces analysed.	a. Does combining give models with conforming response surfaces? b. Do weights from different classical model selection criteria give the same response surfaces and MSE results? c. How do the MSE's of the combined models compare with the best OLS candidate models?
ii)	The four adhesion models with conforming response surfaces are combined using S-BIC weights. Data arrays are produced and response surfaces analysed.	a. Do candidate models with conforming response surfaces give combined models with conforming response surfaces? b. How does the MSE of the averaged model estimator compare with the OLS candidate models?

Table 3 Tabulation of summary investigations of CBFMA (continued)

#	*ANALYSIS*	*QUESTIONS TO BE ANSWERED*
iii)	The only hardness model with a conforming response surface is combined with the all possible regressions OLS models built from the set of regressors making up its functional form, (T, RT, T*RT, T^2) using S-BIC weights. This is repeated using the adhesion response	a. Can a conforming response model be distorted by nonconforming response models made up with regressors from the set of regressors making it? b. How do the MSE's of averaged model estimators compare with the OLS candidate models?
iv)	All the 27 all regression adhesion models without conforming response surfaces are combined using S-BIC weights. Data arrays are produced and response surfaces analysed.	a. Can candidate models without conforming response surfaces produce a combined model with a conforming response surface? b. How does the MSE of the averaged estimator compare with the best OLS model? c. What % of the total sum of weights do these 27 model weights add up to?
v)	All the 30 all regression hardness response models without conforming response surfaces are combined using S-BIC weights. Data matrices are produced and response surfaces analysed.	a. Can candidate models without conforming response surfaces produce a combined model with a conforming response surface? b. How does the MSE of the averaged estimator compare with the best OLS model? c. What % of the total sum of weights do these 30 add up to?

Table 3 summarises the empirical investigations carried out to answer questions related to the usability of CBFMA in MRSM focused on model fitness to dataset (MSE), response surface uncertainty and criterion uncertainty.

3.6 Determination of the MRSM results

A permutation of adhesion-hardness response model pairs with conforming response surfaces was determined. The data arrays of each pair of response models were overlaid. The region satisfying customer expectations was determined and results of each pair tabulated. The permutation of results was averaged to determine final results.

3.7 Prediction accuracy check using mean squared forecasted error

Comparison of the prediction accuracies of models was done using the MSFE computed with validation results adopted from Pavolo and Chikobvu (2019). The formula for MSFE for a sample size *n* is:

$$MSFE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - Y_{Fi})^2}{n}} \quad (7)$$

where Y_i is i^{th} the measured response, Y_{Fi} is the i^{th} forecasted response.

3.8 Assessment of the sufficiency of the solution strategies to problems

The sufficiency of strategies to the problems of model uncertainty, criterion uncertain, response surface uncertainty, parametric bias, and MRSM modelling and model selection problems was empirically assessed. This was to check the applicability of the strategies to the MRSM context.

4 Results

4.1 Results of all possible regression modelling

The all possible regression modelling approach generated 31 models for each response (see Appendix A).

4.2 Results of response surface analysis and candidate model selection

Table 4 shows the models with conforming response surfaces for both adhesion and hardness responses after analysing the response surfaces of all the response models.

Table 4 Showing the adhesion and the hardness response models with conforming response models

PAIR	*Adhesion response surface model*	*Hardness response surface model*
1	T, RT, T*RT	T, RT, T*RT, T^2
2	T, RT, T*RT, T^2	"
3	T,RT, T*RT, RT^2	"
4	T, RT, T*RT, T^2, RT^2	"

Only four adhesion response models and one hardness response models out of thirty one had conforming response surfaces.

4.3 Summary of results of MSE (goodness of fit) analysis

Table 5 shows the MSE's of the models with conforming response surfaces. The adhesion response models are arranged in MSE ascending order. The subscript 'A' stands for adhesion and 'H' for hardness. A subscript number like 31, 27 and 4 with an S-BIC weight indicates the number of OLS response models combined (e.g., S-BIC_{A27}, indicates 27 models combined). S-$BIC_{A(iii)}$ and S-$BIC_{H(iii)}$ are the adhesion and hardness combined models resultant from investigation iii) of Table 3, respectively.

Table 5 Showing the MSE's of four conforming and eight averaged adhesion models plus the conforming hardness model

ADHESION	*MSE*	*HARDNESS*	*MSE*
T, RT, T*RT, T^2, RT^2	0.97615	*T, RT, T*RT, T^2*	*5.38060*
S-HQ_A	1.02168	S-$BIC_{H(iii)}$	5.47910
S-HQc_A	1.03504		
S-BIC_{A31}	1.03504		
S-BIC_{A27}	1.06762	T, RT, T*RT, T^2, RT^2	2.144314
S-AIC_A	1.11023	S-AIC_H	2.193240
S-$AICc_A$	1.26531	S-HQ_H	2.212810
S-BIC_{A4}	1.27401	S-BIC_{H31}	2.212810
T, RT, T*RT, RT^2	1.27893	S-HQc_H	2.212810
S-RMS_A	1.39234	S-BIC_{H30}	2.831040
S-GVC_A	1.44110	S-$AICc_H$	3.509600
T, RT, T*RT, T^2	2.17773	S-RMS_H	6.360080
S-$BIC_{A(iii)}$	2.26892	S-GVC_H	6.618710
T, RT, T*RT	2.34407		

In Table 5, it is seen that the best OLS model, [T, RT, T*RT, T^2, RT^2], has the least MSE.

Table 6 Showing the MSE's of the four conforming adhesion models with their S-BIC averaged estimator

MODELS	*MSE*	*MSFE*
T, RT, T*RT, T^2, RT^2	0.97615	0.4386
S-BIC_{A4}	1.27401	0.4185
T, RT, T*RT, RT^2	1.27893	0.4712
T, RT, T*RT, T^2	2.17773	0.7827
T, RT, T*RT	2.34407	0.6352

S-BIC_{A4} of Table 6 has an MSE greater than the best OLS model but less than the other three OLS models, but has the least MSFE.

The averaged model estimators MAA and MAH in Table 7 have MSE's that are second best to the best OLS models.

In Table 8, the AMA estimators are shown with the CBFMA estimators subjected to AMA. The subscript number signifies the number of CBFMA models combined.

4.4 Results of the analysis of the effect CBFMA on response surfaces

This section presents the results of the empirical analysis of Table 3. The results are presented in accordance with the way the investigations are presented.

Table 7 Showing the MSE's of investigation number iii) in Table 3

MODEL	*ADHESION MSE*	*ADHESION MSFE*	*MODEL*	*HARDNESS MSE*	*HARDNESS MSFE*
T, RT, T*RT, T^2	2.177730	0.7827	T, RT,T*RT,T^2	5.380600	0.3047
MA_A	*2.270287*	0.5935	MA_H	*5.674753*	*1.3885*
T, RT, T*RT	2.344071		T, RT, T^2	6.329171	
T, RT, T^2	2.369766		T, T*RT, T^2	6.787514	
RT, T*RT, T^2	2.384709		T, T^2	7.545351	
RT, T*RT	2.385549		RT, T*RT, T^2	9.994493	
T, RT	2.536079		RT, T*RT	10.699603	
RT, T^2	2.810610		T, RT	11.634219	
T, T*RT, T^2	2.868295		T, T*RT	12.089156	
T, T*RT	3.034651		T	12.851626	
T*RT, T^2	3.516384		T, RT, T*RT	13.049738	
RT	5.742584		RT, T^2	13.539745	
T, T^2	6.021962		T^2	14.757444	
T	6.187742		RT	19.640166	
T^2	6.618071		T*RT	20.428447	
T*RT	9.056869		T*RT,T^2	128.399578	

Table 8 Showing the MSE's of best OLS, CBFMA and AMA estimators

ADHESION	*MSE*
T, RT, T*RT, T^2, RT^2 (OLS)	0.97615
S-HQ_A	1.02168
AMA_2 (S-HQ_A, S-HQc_A)	1.02797
AMA_3(S-HQ_A, S-HQc_A, S-BIC_{A31})	1.03024
S-HQc_A	1.03504
S-BIC_{A31}	1.03504
S-AIC_A	1.11023
AMA_7(All the seven CBFMA estimators)	1.13010
S-$AICc_A$	1.26531
S-RMS_A	1.39234
S-GVC_A	1.44110

Table 10 summarises the findings of the empirical analysis based on the four questions motivating the investigations.

Table 9 Tabulation of summary investigation results of criterion-based frequentist model averaging

EMPIRICAL ANALYSIS #		*FINDINGS*
i)	a	Ans.: Uncertain. The adhesion response gives an averaged model (S-BIC$_{A31}$) with a conforming response surface but the averaged hardness response model (S-BIC$_{H31}$) is not conforming.
	b	Ans.: No. Averaged models from different CBFMA weights have different MES's. All the adhesion models have conforming response surfaces but all the hardness models do not (see Appendix B).
	c	Ans.: The MSE's of the combined models are good but they are all less than the best OLS candidate models (ref. Table 6).
ii)	a	Ans.: In this case, yes, OLS models with conforming response surfaces combine to give an averaged model (S-BIC$_{A4}$) with a conforming response surface.
	b	Ans.: The MSE of the averaged model is larger than that of the best OLS candidate model but better than the other three candidate models (ref. Table 7).
iii)	a	Ans.: The combined response model remains with a conforming response surface but has an MSE that is larger than the OLS model with a conforming response surface (ref. Table 8). This shows that there is a distortion in accuracy.
	b	Ans.: Greater (ave. estimator MSE: best OLS candidate MSE: for hardness = 5.479: 5.381; for adhesion = 2.26892: 2.1773).
iv)	a	Ans.: Yes, for the adhesion response. 27 models with non-conforming models combine to give an averaged model S-BIC$_{A27}$ with a conforming response surface.
	b	Ans.: The MSE of the averaged estimator is greater than the best OLS model. (MSE S-BIC$_{A27}$: best OLS = 1.06762: 0.984094 [T, RT, T^2, RT^2])
	c	What % of the total sum of weights do these 27 model weights add up to? Ans.: 45%
v)	a	Ans.: No, for the hardness response.
	b	Ans.: the MSE of the averaged estimator is greater than for the best OLS model. (MSE S-BIC$_{H30}$: Best OLS = 2.831038: 2.144314 [T, RT, T*RT, T^2, RT^2]) ref. Table 6.
	c	What % of the total sum of weights do these 30 add up to? Ans.: 99%

4.5 MRSM results

This subsection presents the results of optimising the pairs of response models with conforming response surfaces. Figure 1 shows the results of optimising each model–pair as shown in Appendix C. Each of the results tables can be used since they are results of overlaying two response models with conforming response surfaces.

Table 10 Showing summary of Table 9

EMPIRICAL INVESTIGATION (#)QUESTION	*(i)*	*(ii)*	*(iii)*	*(iv)*	*(v)*	*ANALYSIS CONCLUSIONS*
Q1: Does CBFMA solve the model, criterion and response surface uncertainty problems?	N	Y	Y	N	N	Three N's and the two Y's signify uncertainty. In the MRSM case, there remains the response surface uncertainty and the criterion uncertainty
Q2: Does CBFMA adapt and capture the best performance among candidate models?	Y	Y	Y	Y	N	Yes. In (v) the CBFMA model estimator does not have conforming response surface though it has a good MSE
Q3: Does CBFMA produce a model which has a better performance than any of the original candidate models?	N	N	N	N	N	The MSE of the CBFMA estimator is second best in all investigated incidences to best OLS model.
Q4: Does CBFMA provide insurance against selecting a poor model, hence improving the risk of estimation?	N	Y	Y	Y	N	The two N's and three Y's indicate uncertainty to the answer to this question. In the MRSM case a poor model does not have a conforming response surface.

Note: N implies no and Y implies yes.

Two averaged results were compared:

1 the averaged results from the four adhesion OLS models with conforming response surfaces

2 the averaged results from all the pairs in Figure 1.

Cure time was plotted against rubber thickness. The averaged results of 1) have the best fit. Table 11 shows the result of combining the four tables of adhesion models with conforming response surfaces. This result gives the cure time-adhesion model with the best fit (R-sq. = 99.4% and R-sq.(adj.) =99.2%) and fit (S = 0.225032).

Table 11 Showing the averaged results

Rubber thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Cure time (min.)	21	22	23	24	24	25	25	26	26	27	27.25	28	28.75	29.75

4.6 Results of prediction accuracy check (MSFE results)

The CBFMA and AMA combined models showed better predictive accuracy (MSFE) than the OLS best model. The MSE column shows that the OLS best model has an MSE better than the CBFMA and AMA models signifying that although the OLS best model has the best fitness to data (MSE) but its predictive accuracy (MSFE) is not better than CBFMA and AMA combined models.

Figure 1 Showing the results from the fourteen pairs of adhesion-hardness models

[T, RT, T*RT, T^2, RT^2] vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

S-HQ vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

S-BIC$_{A31}$ vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

S-HQc vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

S-AIC vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

S-AICc vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

S-BIC$_{A4}$ vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

[T, RT, T*RT, RT^2] vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	29

S-RMS vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	29

S-GVC vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	29

S-BIC$_{A27}$ vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	28

[T, RT, T*RT] vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	28	28	29

[T, RT, T*RT, T^2] vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	29

S-BIC$_{A(iii)}$ vs. [T, RT, T*RT, T^2]

Rubber Thickness (mm)	7	8	9	10	11	12	13	14	15	16	17	18	19
Cure Time (min.)	21	22	23	24	24	25	25	26	26	27	27	28	29

Figure 2 Showing the polynomial relationship of cure time and rubber thickness from Minitab 17 (see online version for colours)

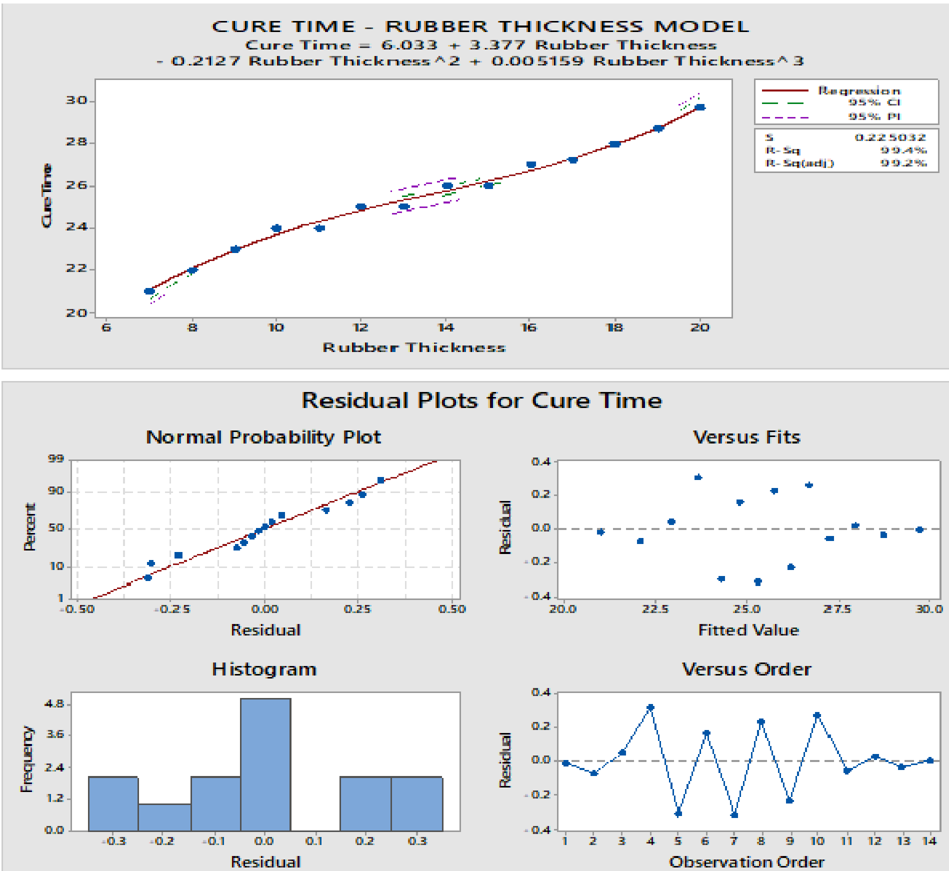

Table 12 Showing comparison of MSE and MSFE results

ADHESION	*MSE*	*MSFE*
S-HQ$_A$	1.0217	*0.3963*
AMA$_2$ (S-HQ$_A$, S-HQc$_A$)	1.0280	0.3977
AMA$_3$ (S-HQ$_A$, S-HQc$_A$, S-BIC$_{A31}$)	1.0302	0.3982
S-HQc$_A$	1.0350	0.3993
S-BIC$_{A31}$	1.0350	0.3993
S-AIC$_A$	1.1102	0.4184
AMA$_7$ (all the seven CBFMA estimators)	1.1301	0.4262
T, RT, $T*RT$, T^2, RT^2 (OLS)	*0.9762*	0.4386
S-AICc$_A$	1.2653	0.4658
S-RMS$_A$	1.3923	0.4784
S-GVC$_A$	1.4411	0.4884

4.7 Results of the assessment of the sufficiency of solution strategies to problems

Table 13 Showing how the sufficiency of problem solution strategies were assessed

#	*Problem description*	*Solution strategy*	*Method of assessment*	*ASSESSMENT RESULT*
1	Selection criteria problems. • Small sample size dataset inefficiency • Small sample size dataset modelling credibility	• Avoid MS • Use rigour to justify credibility	• Avoiding MS ensures avoiding the criteria small sample size problems • The solution methodology used must demonstrate rigour	• Results produced were validated with MSFE and compared with OLS. MSFE of CBFMA better than for OLS. • Results of permutation of candidate model response pairs where obtained without averaging then averaged.
2	Parameter bias and variability • Regression parameters are often biased • Standard errors too small because they do not reflect uncertainty	• Combine CBFMA estimators using AMA • Use MA	• Compare MSE's and MSFE's • Compare MSE's and MSFE's of OLS, MS vs. MA	• Ref. Table 13. MSFE's of AMA results are better than MS though less than CBFMA. • Ref. Table 13. CBFMA results of MSE's second to best OLS. MSFE's are better.
3	Criterion uncertainty	• Avoid MS, proceed to results with candidate models • Combine CBFMA estimators using AMA	• Compare results and average if necessary • Compare MSE's and MSFE's of OLS, CBFMA and AMA	• Averaged result was validated with MSFE • Ref. Table 13. AMA MSE's are competitive. MSFE's of AMA are better than OLS.

5 Discussion of findings

In this section, discussion of results from Section 4 is done and the impact of the findings to MRSM small sample size dataset problems is discussed.

5.1 Analysis of results in Table 10

1. The findings from the results observed in averaging all the 31 all regression models for both adhesion and hardness by the seven CBFMA weights suggest that:
   - For MRSM small sample size datasets, CBFMA does not deal with model uncertainty where response surface conformity is concerned. The two Y’s three N’s in the first row of Table 10 suggest uncertainty. The seven criterion-based model estimators for adhesion have conforming response surfaces, whilst the seven for hardness do not have.
   - The MSE’s of the seven CBFMA model estimators are different but show good fitness to data. However, the MSE’s of the entire seven CBMA model estimators are greater than the best OLS and MS model. The MSFE’s suggest better prediction accuracy than the OLS candidate models.
2. The findings from the combining of the four adhesion models with conforming response surfaces are:
   - The four adhesion response models with conforming response surfaces combine to give an averaged estimator model with a conforming response surface suggesting that if candidate models with conforming response surfaces are combined the result has certainty. This is one possible way we can eliminate response surface uncertainty in MRSM.
   - The MSE of the averaged estimator model is better than of three OLS candidate models but less than the best accurate OLS candidate model. However, the MSFE is smaller than for all the candidate models.
3. Mixing a response model with a conforming response surface with the permutation of models whose constituent regressors are a subset of its regressors gives the following findings:
   - It does not necessarily distort the response surface beyond conformity. This is demonstrated by model [T, RT, T*RT, T^2], the only hardness model with a conforming response surface. This finding is suggested by both adhesion and hardness responses.
   - Again, the OLS model [T, RT, T*RT, T^2], in both adhesion and hardness cases, has the least MSE; less than the MSE’s of the combined estimators. However, the MSFE for the combined estimator adhesion is less than the OLS model.
4. The effect of combining the 27 adhesion response models with nonconforming response surfaces gives the following results:
   - The averaged estimator has a conforming response surface.
   - The averaged estimator has an MSE suggesting good fitness to data though still less than the best OLS model. However, the averaged estimator shows good prediction accuracy (MSFE = 0.3979) better than best OLS model.
5. The effect of combining the 30 hardness models without conforming response surfaces demonstrates the response surface uncertainty problem. CBFMA does not always solve the uncertainty problem where response surfaces are concerned.

Q1 Does CBFMA solve the model selection uncertainty problem?

(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*	*COMMENT*
N	Y	Y	N	N	Three N's and the two Y's signify uncertainty. In the MRSM case, there remains the response surface uncertainty and the criterion uncertainty

The finding here does not suggest agreement with Wan et al. (2010). It is suggested that:

a If combining all the all possible regressions models of a response gives a model with a conforming response surface, then the combining of all the models without conforming response surfaces may give a conforming response surface.

b If combining all the all possible regressions models of a response gives a non-conforming response surface, combining all the non-conforming response surfaces gives a non-conforming response surface.

c If there is certainty in the way that candidate response models are selected, there is certainty in the averaged estimator.

Q2 Does CBMA adapt and capture the best performance among candidate models?

(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*	*COMMENT*
Y	Y	Y	Y	N	Yes. In (v) the CBFMA model estimator does not have conforming response surface though it has a good MSE

Empirical investigations (i), (ii), (iii) and (iv) give results in agreement with Yang (2003). The failure of 30 hardness non-conforming response models to combine to give a conforming response model in (v) is the only negative though the MSE shows good fitness to data.

Q3 Does CBFMA produce a model which has a better performance than the original candidate models?

(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*	*COMMENT*
N	N	N	N	N	The MSE of the CBFMA estimator is second best in all investigated incidences to best OLS model. However, MSFE of the CBFMA estimator performs better than the OLS model.

MSE findings suggest disagreement with Juditsky and Nemirovsky (2000), however, analysis of MSFE results seem to agree with Juditsky and Nemirovsky. Best fitness to data does not necessarily imply best prediction accuracy.

Q4 Does CBFMA provide insurance against selecting a poor model?

(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*	*COMMENT*
N	Y	Y	Y	N	The two N's and three Y's indicate uncertainty to this question. In the MRSM case a model is poor because it does not have a conforming response surface.

The findings from investigations (ii), (iii) and (iv) suggest agreement with this finding by Leung and Barron (2006) but those of (i) and (v) suggest disagreement. CBFMA simply avoids selection by opting to combine the competing models. If the candidate models are all poor models, the combined estimator may be poor as well.

Q5 Does CBFMA improve model forecast accuracy?

The MSFE's of averaged models suggest that CBFMA produces models with good accuracy that may even be better than the best OLS models.

Q6 When can CBFMA be best applied in MRSM small sample dataset problems?

Where prediction accuracy is required, MSFE results suggest that CBFMA may be the best option. CBFMA produces a number of good accurate models which may make averaging results the best option. Where CBFMA is used, it is better to use a number of criterion-based frequentist weights to combine the candidate models whose results are then further averaged for the final result. Two stages of averaging, that is:

1 MA

2 AMA or averaging results, are therefore suggested.

Q7 Can we adequately deal with the small sample size problems of MRSM datasets?

Table 14 shows that it is possible to adequately deal with the small sample size problems of MRSM datasets by using the right strategies.

6 Conclusions and future research

6.1 Conclusions

The empirical findings suggest that CBFMA produces model estimators with better prediction accuracy than candidate OLS models. Although the fitness to data of the averaged model estimator is improved, it is always less than the best OLS candidate model. However, CBFMA does not directly solve the criterion and response surface uncertainty problems in the MRSM context.

6.2 Unique contributions of the paper

- The paper lists together five data modelling and model selection problems that an MRSM practitioner encounters in MRSM work and suggests deliberate strategies to employ to achieve credible results from the small sample size MRSM dataset using CBFMA.

- That CBFMA produces model estimators that have better prediction accuracy (MSFE), but always have less good fit to data (larger MSE), than the “best” OLS candidate model.
- That not all CBFMA weights produce averaged models with prediction accuracy (MSFE) better than the best OLS candidate model, hence selection criterion uncertainty still exists.
- That AMA can be useful where models have the same functional structure producing models with good fitness to data (MSE) and prediction accuracy (MSFE).
- The paper also highlights to MRSM practitioners to be on the watch for criterion uncertainty (which selection criterion to use for CBFMA?), response surface uncertainty (good-fitness-to data does not necessarily imply conforming response surface) just as for model uncertainty (which model to use?).

6.3 Theoretical and managerial implications

The paper implies that MRSM practitioners need to be more careful in their work as, statistically, they deal with small sample size datasets that lack modelling credibility and therefore demand rigour. CBFMA should be used together with the suggested strategies to ensure credible and accurate results.

6.4 Limitations of the research

The research work is an empirical analysis designed to add to literature on the applied statistics side of MRSM. Some areas would definitely require simulation studies to come out with weighted positions that is why findings are presented as suggestions.

6.5 Future research

A number of issues for future research arise from this work:

1 The impact and usability of frequentist OMA in small sample size MRSM datasets.

2 The idea of cloning conforming response surfaces by combining models with nonconforming response surfaces as a strategy of increasing the number of candidate models with the same functional form that can be combined using AMA to minimise parametric bias.

3 And the usability and applicability of re-sampling methods (jackknifing and bootstrapping) in small sample size MRSM datasets.

Appendices/Supplementary materials are available on request by emailing the corresponding author.

References

Alhorn, K., Schorning, K. and Dette, H. (2019) 'Optimal designs for frequentist model averaging', *Biometrika*, https://arxiv.org/pdf/1807.05234.pdf.

Cavanaugh J.E. and Neath, A.A. (2019) *The Akaike Information Criterion: Background, Derivation, Properties, Application, Interpretation and Refinements*, WIRES Computational Statistics, No. e1460.

Emmert-Streib, F. and Dehmer, M. (2019) 'Evaluation of regression models: model assessment, model selection and generalization error', *Machine Learning and Knowledge Extraction*, Vol. 1, pp.521–550, DOI: 10.3390/make 1010032.

Hansen, B.E. (2007) 'Least squares model averaging', *Econometrica*, Vol. 75, No. 4, pp.1175–1189.

Hansen, B.E. and Racine, J. (2012) 'Jack-knife model averaging', *Journal of Econometrics*, Vol. 167, No. 1, pp.38–46.

Hu, X., Madden V.L., Edwards, S. and Xu, X. (2015) 'Combining models is more likely to give better predictions than single models', *Phytopathology*, Vol. 105, No. 9, pp.1174–1182.

Hurvich, C., and Tsai, C. (1989) 'Regression and time series model selection in small samples', *Biometrika*, Vol. 76, No. 2, pp.297–307.

Juditsky, A. and Nemirovsky, A. (2000) 'Functional aggregation for nonparametric estimation', *Annals of Statistics*, Vol. 28, No. 3, pp.681–712.

Khuri, A.I. (2017) 'A general overview of response surface methodology', *Biom. Biostat. Int. J.*, Vol. 5, No. 3, p.133, DOI: 10.15406/bbij.2017.05.00133.

Leung, G. and Barron, A.R. (2006) 'Information theory and mixing least squares regressions', *IEEE Transactions on Information Theory*, Vol. 52, No. 8, pp.3396–3410.

Lin, C. and Kuo, B. (2016) 'Model averaging in predictive regressions', *Econometrics Journal*, Vol. 19, No. 2, pp.203–231.

Magnus, J.R., Powell, O. and Prüfer, P. (2010) 'A comparison of two averaging techniques with an application to growth empirics', *Journal of Econometrics*, Vol. 154, No. 2, pp.139–153.

MaQuarrie, A.D.R., and Tsai C.L. (1998) *Regression and Time Series Model Selection*, World Scientific Publishing Co. Pte Ltd., Singapore.

Mitra, P., Lian, H., Mitra, R., Liang, H. and Xie, M. (2019) 'A general framework for frequentist model averaging', *Science China Mathematics*, Vol. 62, No. 2, pp.205–226, https://doi.org/10.1007/s11425-018-9403-x.

Myers, R.H., Montgomery, D.C. and Anderson-Cook, C.M. (2016) *Response Surface Methodology: Process and Product Optimisation Using Designed Experiments*, 4th ed., ISBN: 978-1-118-91601-8.

Nguefack-Tsague, G. (2014) 'On optimal weighting scheme in model averaging', *American Journal of Applied Mathematics and Statistics*, Vol. 2, No. 3, pp.150–156.

Okoli, K., Breini, K., Brandimarte, L., Botto, A., Volpi, E. and De Baldassarre, G., (2018) 'Model averaging versus model selection: estimating design floods with uncertain river flow data', *Hydrological Sciences Journal*, Vol. 63, Nos. 13–14, pp.1913–1926, DOI: 10.1080/02626667.2018.1546389.

Pavolo, D. and Chikobvu, D. (2019) 'A practical solution to the small sample size bias and uncertainty problems of model selection criteria in two-input process multiple response surface methodology datasets', *Open Journal of Statistics*, Vol. 9, pp.109–142, https://doi.org/10.4236/ojs.2019.91010.

Rawlings, J.O., Pantula, G.S. and Dickey A.D. (1998) *Applied Regression Analysis: A Research Tool*, 2nd ed., Springer-Verlag New York Inc., New York.

Schomaker, M. and Heumann, C. (2018) *When and When Not to Use Optimal Model Averaging*, Statistical Papers, in press.

Seghouane, A.K. and Bekara, M. (2004) 'A small sample model selection criterion based on the Kullback symmetric divergence', *IEEE Transactions Signal Processing*, Vol. 52, No. 12, pp.3314–3323.

Steel, M.F.J. (2017) *Model Averaging and its Use in Economics*, MPRA Paper 81568, University Library of Munich, Germany.

Wan, T.K.A., Zhang, X. and Zou, G. (2010) 'Least squares model averaging by Mallows criterion', *Journal of Econometrics*, Vol. 156, No. 2, pp.277–283.

Xie, T. (2015) 'Prediction model averaging estimator', *Economics Letters*, Vol. 131, No. C, pp.5–8.

Yang, Y. (2003) 'Regression with multiple candidate models: selecting or mixing?', *Statistica Sinica*, Vol. 13, No. 3, pp.783–809.

Zhang, X. and Liu, C-A. (2018) *Inference After Model Averaging in Linear Regression Models*, IEAS Working Paper No. 17-A005.

Zhang, X., Ullah, A. and Zhao, S. (2016) 'On the dominance of Mallows' averaging estimator over ordinary least squares estimator', *Economic Letters*, Vol. 142, No. C, pp.60–73.

Zhao, S., Ullah, A. and Zhang, X. (2018) 'A class of model averaging estimators', *Economic Letters*, Vol. 162, No. C, pp.101–106.