Paper 175-2023

What's Next After Model Comparison?
Model Selection and Model Averaging in SAS®

Chong Ho (Alex) Yu, PhD, DPhil, Hawai'i Pacific University

Charlene J Yang, MA, Azusa Pacific University

ABSTRACT

To enhance the generalizability of the findings and to overcome the replication crisis, it is common practice to run multiple models under the framework of ensemble methods, instead of drawing a conclusion based on a single analysis. At the stage of model comparison, the analyst can choose between model selection and model averaging. In model selection, the best model is retained based on predictive accuracy, error rates, or parsimony, such as F1 score, generalized R-square, RASE, AIC, and BIC. By doing so, the second-best and other less adequate models are totally discarded. On the other hand, in model averaging, information from all models is utilized in several ways, such as averaging the prediction estimates from all models, selecting the highest estimates of all models, or returning the proportion of the models that can substantially contribute to the outcome. While these options are available in SAS® Viya, SAS® Enterprise Miner™, and JMP® Pro, no consensus exists on how model selection and model averaging should be properly used under different situations. There have been prior research studies that found, in most cases, model averaging and model selection lead to comparable results. In contrast, another study found that model averaging can yield more accurate results than model selection when researchers study complex models, such as nonlinear mixed-effect models. Further, some researchers argue that model averaging generally outperforms a single best model, especially when the analyst does not have information on the relative performance of the candidate models. Nonetheless, model averaging is based on the assumption that all models have certain merits that can contribute to the assembly of the final model, but in reality, it might not always be the case. In this paper, the advantages and disadvantages of both approaches will be discussed and illustrated.

INTRODUCTION

WHY IS MODEL COMPARISON IMPORTANT?

In the past, usually data analysis was a one-shot process. Specifically, a single data set was fed into a single modeling process. As a consequence, the model is overfitted to a particular sample (Kuhn & Johnson, 2013), resulting in a replication crisis (Colling & Szucs, 2018; Open Science Collaboration, 2015). In addition, popular traditional modeling methods, such as regression analysis, have certain limitations. For example, the data structure must conform to restrictive assumptions (Yu, 2022b). To address the preceding issues, data scientists have to compare multiple models, involving various techniques like neural networks, boosting, bagging, support vector machines, and more. In doing so, analysts avoid the risk of placing "all eggs into one basket." Instead, conducting multiple analyses with various data subsets tends to yield a more accurate model, echoing the concept of the "wisdom of the crowd" (Surowiecki, 2004). Further, modern data science methods are much more versatile than traditional modeling methods. For instance, most are non-parametric in essence and require very few or even no assumptions (Yu, 2022a).

AFTER MODEL COMPARISON: MODEL SELECTION AND AVERAGING

Model comparison serves as the initial phase of the solution. Once multiple models have been generated, the analyst faces the task of determining how to leverage the diverse outcomes. Diversity is key. If all modeling results are identical, then this work is redundant because it does not provide additional information. Facing different results from different models, the analyst can choose between two courses of

action: Model Selection (MS) which retains the best model, or Model Averaging (MA) which synthesizes information from multiple models to generate the final model.

MS and MA are not unique in data science. Before the rise of data science and machine learning (DSML), both MS and MA have been implemented in some traditional and Bayesian statistical procedures (Piironen & Vehtari, 2017; Raftery et al., 1997). However, in contrast to their DSML counterparts, MS and MA are typically confined to a single modeling technique, such as stepwise regression, forward selection, backward elimination, and all-subset regression (SAS® Institute, 2020). In this sense, there is not much diversity in traditional and Bayesian MS/MA. Indeed, many authors advise against using stepwise regression approaches and other traditional MS methods (Burnham & Anderson, 2002; Harrell, 2001; Lavery et al., 2016; Smith, 2018).

Both MS and MA come with their own set of strengths and weaknesses, and the analyst's choice depends on the specific problem, data, and objectives at hand. While SAS® software product documentation offers users in-depth information about various procedures, there is no clear guideline about which one is more appropriate. The primary focus of this paper is not to illustrate the procedures of MS and MA, as they are already documented; rather, the primary objective is to discuss the efficacy of both MS and MA and suggest proper applications for both. The following is an overview of the pros and cons of model selection and model averaging (Dietterich, 2000; Hastie et al., 2009; Gao et al., 2016).

ADVANTAGES AND DISADVANTAGES

MODEL SELECTION

Advantages:

1. Simplicity and Efficiency: Model selection simplifies the decision-making process by choosing a single "best" model from the candidate models. This can make the chosen model easier to interpret and implement. Moreover, it offers enhanced efficiency as it allows for the straightforward exclusion of all other models, essentially representing an all-or-none decision.
2. Interpretability: The selected model is often easier to interpret than an average of multiple models. This can be crucial for understanding the relationship between predictors and the target variable. The analyst is not required to provide justifications for incorporating predictors or predictor information from "inferior" models.
3. Computational Efficiency: Model selection typically requires less computational resources compared with model averaging, because after the top model is selected, no further action needs to be taken or considered with all the rest.

Disadvantages:

1. Risk of Overfitting: Model selection can lead to overfitting, especially if the selection criterion is used to choose the most complex model. Overfit models may not generalize well to new data.
2. Vagueness of “The Best”: There are different criteria for model selection, namely, R-square, RMSE, AIC, BIC, and so forth. While a model is considered the best by a certain criterion, it may be the second best when another reference is used. The selection process is subject to the preference of the analyst.
3. Model Uncertainty: One of the most pervasive disadvantages is that it does not account for model uncertainty. The chosen model may perform well on the training data, but might not capture the true underlying relationship in the data. Nevertheless, while traditional modeling approaches are prone to this error, it is less applicable to data science and machine learning, because ensemble methods can fine-tune the model by partitioning the data into numerous subsets and running repeated analyses on them.

4. Ignoring Valuable Information: Model selection discards potentially valuable information contained in the other candidate models. This can be a limitation when multiple models have complementary strengths. This drawback can be likened to a scenario where a university chooses a team to represent the school in a national Jeopardy contest following internal competitions between schools and colleges. In this case, the university selects a team from one college but disregards the other talents of the "losing" teams.

MODEL AVERAGING

Advantages:

1. Reduces Overfitting: Model averaging mitigates the risk of overfitting by incorporating multiple models, enhancing its ability to generalize effectively when faced with unseen data.
2. Accounts for Model Uncertainty: One of the primary justifications for using model averaging, is that it explicitly considers model uncertainty. It acknowledges that multiple models may have similar predictive performance but differ in their parameter estimates.
3. Improved Robustness: Averaging models can lead to more robust predictions, as they smooth out individual model idiosyncrasies and errors.

Disadvantages:

1. Complexity: Model averaging can be more complex to implement and manage than selecting a single model. It requires keeping track of multiple models and their associated weights.
2. Loss of Interpretability: Averaged models are often less interpretable than individual models. It can be challenging to explain how the ensemble of models arrived at a particular prediction.
3. Increased Computational Cost: Model averaging typically requires more computational resources as it involves training and evaluating multiple models. This may not be feasible in situations with strict computational constraints. This weakness is of lesser significance today given that high-performance computing can easily be performed with GPU-enabled computers on the cloud.

CRITERIA FOR EVALUATING MODEL GOODNESS

Across traditional statistics, the Bayesian approach, and DSML, there are various criteria for evaluating the goodness of models, each with its own advantages and disadvantages. The following are some of the popular criteria (James et al., 2013):

F1 Score: It measures accuracy by balancing precision and recall, providing a single metric that considers both false positives and false negatives. It is useful for imbalanced datasets where one class dominates. However, the F1 score may not be suitable for all scenarios, especially when the trade-off between precision and recall needs to be adjusted.

Generalized R-Squared: It provides a measure of the proportion of variance explained by the model. This approach has long been used in traditional regression analyses. It is well known, and thus, easy to interpret. Yet, it might not suit complex models featuring numerous predictors, as it tends to amplify variance when more predictors are included, even when their contribution to explanatory power is minimal.

RASE (Root Average Squared Error): It measures the difference between predictions and actual values which represents the average error magnitude in the original units of the dependent variable. However, it is sensitive to outliers and can be influenced by extreme values. Further, it is scale-dependent and is not always easy to interpret due to the squared error term.

AIC (Akaike Information Criterion), AICc (AIC corrected), and BIC (Bayesian Information Criterion): They estimate model quality based on balancing goodness of fit and complexity, but favor parsimony. This approach can prevent overfitting, and thus, it is useful for model selection, especially when comparing different models.

Interpretation can be challenging for those not familiar with the theory behind these criteria. The choice of the "best" evaluation criterion depends on the specific context. There is no one-size-fits-all answer. The most popular criterion may vary across different domains and applications. For example, in classification tasks, the F1 score is often used, whereas in models involving a continuous-scaled target variable, R-squared or mean squared error (MSE) might be more suitable.

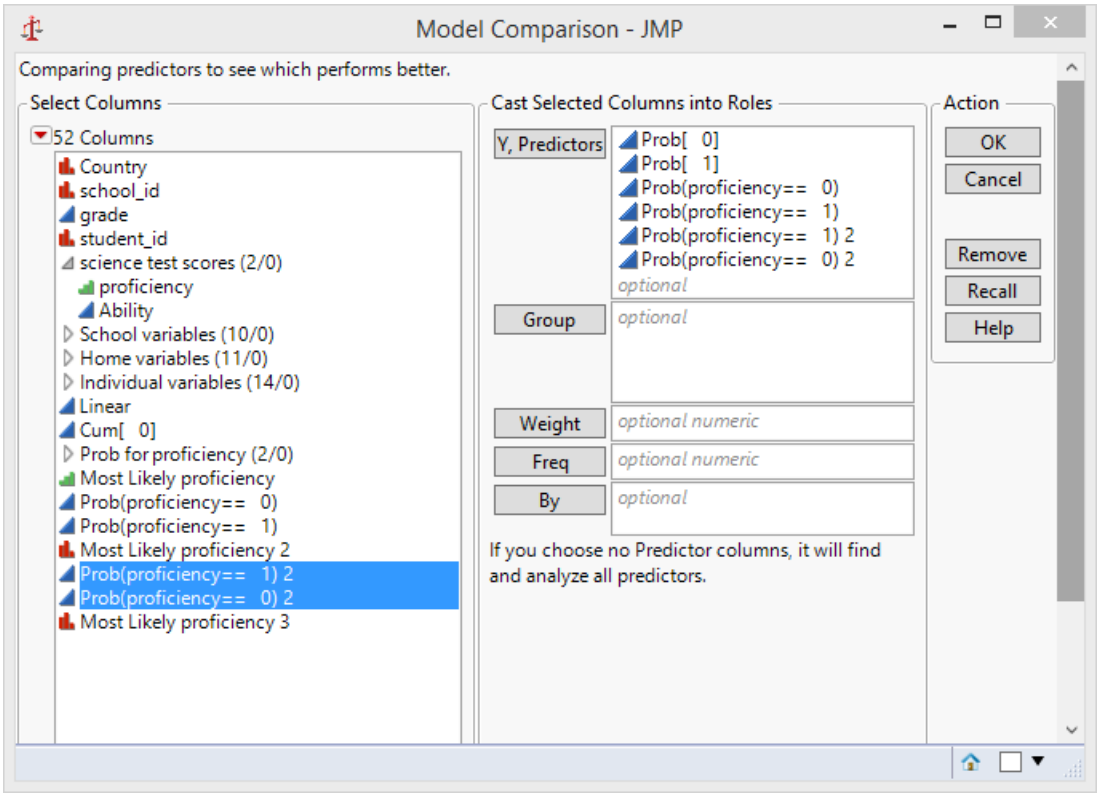
MODEL COMPARISON, MODEL SELECTION, AND MODEL AVERAGING IN SAS®

MS and MA are available in many SAS® products, including JMP® Pro, SAS® Studio, SAS® Enterprise Guide®, SAS® Enterprise Miner™, and SAS® Viya. Due to space constraints, only JMP® Pro and SAS® Enterprise Miner™ are discussed in this paper.

JMP® PRO

Within JMP® Pro, model selection is performed in *Model Screening* whereas *Model Comparison* offers the choice between model selection or model averaging. After the predicted outcomes for all observations of all models are generated, they can be inputted into *Model Comparison* for further investigation (see Figure 1).

Figure 1. Model Comparison in JMP® Pro

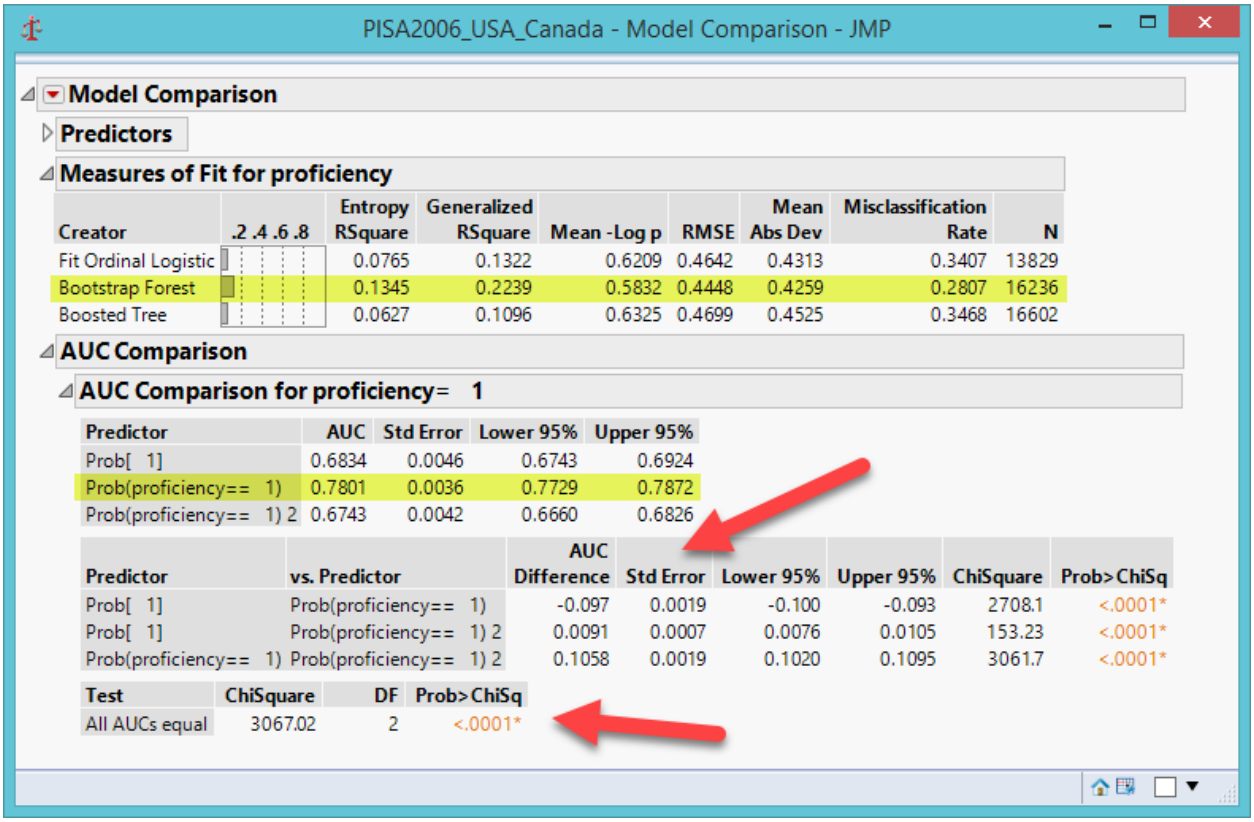


If the data analyst would like to select the single best model, then different criteria can be examined to determine which one is the best. In this example, the bootstrap forest model has the highest Entropy R-

square (based on purity), the lowest Root Mean Square Error, the lowest misclassification rate, the highest AUC, and the lowest standard error (see Figure 2).

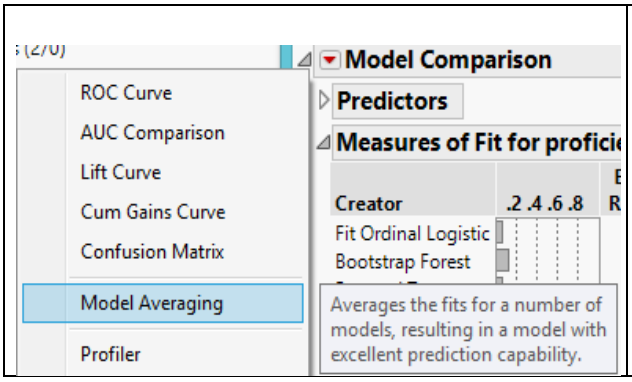
The bottom illustration in Figure 2 shows the test result of the null hypothesis. It is hypothesized that all AUCs are not significantly different from each other, but indeed they are. The table "AUC Comparison for proficiency = 1" in Figure 2 shows the multiple comparison results (logistic vs. bagging; boosting vs. logistic; bagging vs. boosting). All pairs are significantly different from each other.

Figure 2. Model Comparison Results in JMP® Pro



Alternatively, the analyst can also use model averaging, which creates a new field of the arithmetic mean of the predicted values across models (see Figure 3).

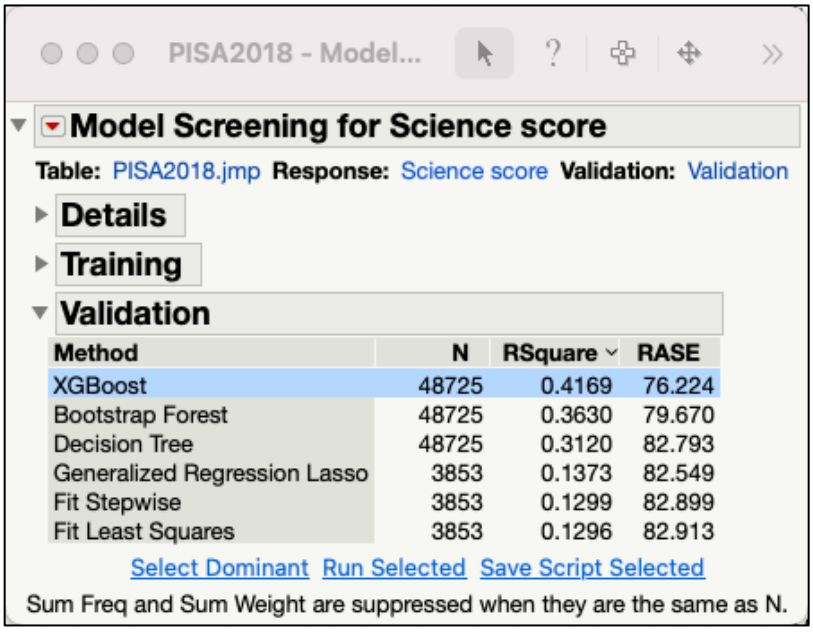
Figure 3. Model Averaging in JMP® Pro



proficiency	0 Avg Predictor	proficiency	1 Avg Predictor
	0.6089894538		0.3910105462
	0.5698144172		0.4301855828
	0.3208607923		0.6791392077
	0.5072586939		0.4927413061
	0.4536519821		0.5463480179
	0.3260864315		0.6739135685
	0.5267876581		0.4732123419
	0.6061649692		0.3938350308

JMP® Pro provides a quick tool for model selection, which is known as *Model Screening*. In this procedure, multiple methods are employed to analyze the same data set in tandem. Afterward, a summary table is displayed for selecting the dominant model. As shown in Figure 4, the information on model adequacy presented by *Model Screening* is less than that of *Model Comparison*. In this example, the dominant model is XGBoost in terms of both R-square and RASE while the poorest one is the least square regression model. However, no model averaging is allowed in *Model Screening*.

Figure 4. Model Screening in JMP® Pro

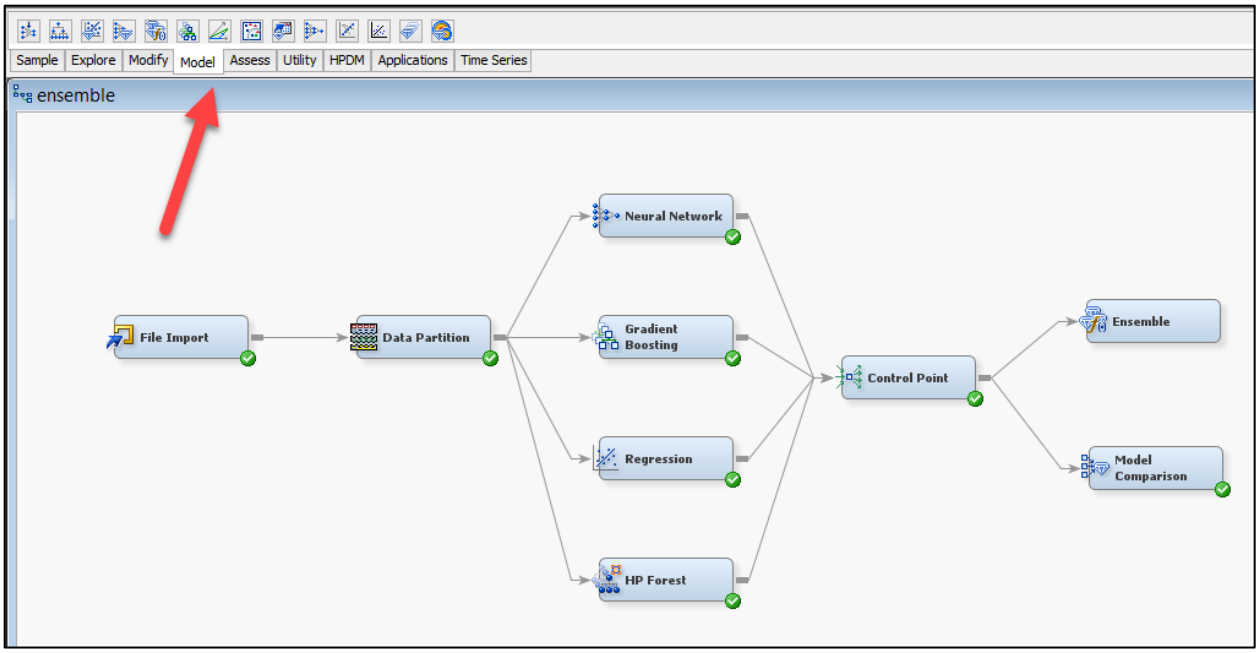


SAS® ENTERPRISE MINER™

SAS® Enterprise Miner™ offers the capability of utilizing both MS and MA techniques. In Figure 5, you can observe the utilization of four distinct modeling techniques: neural networks, gradient boosting, regression, and high-performance forest. After these methodologies are employed, all results are stored in the *Central Point* for model comparison or *Ensemble*. The difference between *Ensemble* and *Model Comparison* is that the former merges all modeling results whereas the latter simply retains the best. Specifically, the *Ensemble* node harmonizes component models derived from these four modeling methods to create the ultimate model solution. Specifically, the *Ensemble* method constructs new models by combining posterior probabilities (for class targets) or predicted values (for interval targets) derived from multiple precursor models. This newly created model is subsequently employed for scoring new data. Various techniques for amalgamating information from diverse models are available:

- **Average**: This method calculates the mean of the posterior probabilities (for categorical targets) or the predicted values (for interval targets) from different models, offering it as the prediction from the *Ensemble* node.
- **Maximum**: The Maximum method selects the highest posterior probabilities (for categorical targets) or the maximum predicted values (for interval targets) among the various models, presenting it as the prediction from the *Ensemble* node.
- **Voting**: Primarily applicable to categorical targets, the Voting method facilitates the computation of posterior probabilities. Two voting methods are available for this purpose: Average and Proportion.

Figure 5. Ensemble and Model Comparison in SAS® Enterprise Miner™



LITERATURE REVIEW

Table 1. Summary of Literature Review Comparing Between Model Selection and Model Averaging

Author(s) and Paper	Traditional, Bayesian, or DSML Methods	Criteria for Comparison	Favors Model Selection	Favors Model Averaging	Either Way or Inconclusive	Depends on the Conditions	Favors Combining MS and MA	Notes
Aoki et al. (2013)	DSML	AIC & BIC	X					Four methods are introduced that combine multiple candidate dose-response models: model selection, bootstrap model selection, model averaging, and bootstrap model averaging. Bootstrap model selection performed best overall. It had good accuracy for dose finding, decision-making, and estimating the probability of achieving the target response. MA reduced bias compared to just using a single-selected model, but was still outperformed by bootstrap approaches.
Berge (2017)	Bayesian & DSML	Quadratic Probability Score & ROC/AUC		X				Four methods are compared: equally weighted forecast, Bayesian Model Averaging (BMA), Linear Boosted Model, and Nonlinear Boosted Model. BMA and boosted models performed much better than equally weighted forecasts, as they can discriminate between useful predictors.

Buatois et al. (2018)	TR	AIC		X				In an informative design, MA and MS provided similar predictive performances and led to accurate prediction of the target dose. However, with less informative designs, by estimating weights for a predefined set of NLMEMs, MA showed better overall predictive performances than MS, increasing the likelihood of accurately characterizing the dose-response relationship.
Gao et al. (2016)	TR	AIC, BIC & MSFE		X				This investigation applied six commonly used MS criteria, including AIC, BIC, HQC, Mallows' C_p, LooCV, and LsoCV, and six FMA methods, namely S-AIC, S-BIC, S-HQC, JMA, LsoMA, and AFTER. The MSFE (mean-squared forecast error) was calculated for each selected model and MA method. The mean of MSFE was used to rank the performance of each method, and from the ranking, the LsoMA method was found to be the best.
Grainger et al. (2018)	TR, Bayesian & DSML	AIC & AICc					X	To address the problem of food waste, both MS and MA were employed to retain the most promising models out of 16,384 potential candidates.
Gu et al. (2018)	TR	AIC, AICc, BIC, & APRESS					X	APRESS outperforms AIC and BIC when it comes to comparing models, making it the preferred choice for model comparison in nonlinear modeling. While MA offers increased predictive robustness compared to choosing a single model, it is advisable to combine both MS and MA techniques for nonlinear modeling to achieve the best results.
Haggag (2014)	TR	AIC, BIC, TIC, HQC & Mallows' C_p		X				Results showed smaller values of bias, variance, and PMSE for regression coefficient estimates of MA than that of MS.
Okoli et al. (2018)	Bayesian	AIC & RMSE			X			These authors compared MS and two types of MA, arithmetic (unweighted) MA and weighted MA. When the sample size was small, both MS and MA outperformed a single model. When the sample size was large, MS and MA (weighted or unweighted) had similar variances.

Peng & Yang (2020)	TR & Bayesian	BIC			X			When there is no approximation error advantage in linear combining, it is unclear whether MA improves over MS. In a nested linear model setting, the paper shows the optimal risk of MA can significantly reduce the optimal risk of MS when the true coefficients decay slowly. When coefficients decay quickly, the risks of optimal MA and MS are asymptotically equivalent. MA does not provide benefits when the best model size is small compared to the sample size.
Richards et al. (2010)	TR	AIC			X			The MS approach provides substantial benefits in tackling complex modeling tasks. While MA can improve stability, evidence for its impact on accuracy remains inconclusive.
Schomaker & Heumann (2013)	TR	AIC		X				MA offers more stable estimates than MS in the presence of missing data, primarily due to its shrinkage properties which reduce variance at the expense of some bias.
Schorning et al. (2016)	TR	AIC, BIC & TIC		X				This study used both candidate models (6) and simulations (40). Overall from simulations, MA outperformed MS methods. The benefits were not large but they were consistent.
Symonds & Moussalli (2011)	TR	AIC				X		If one can clearly identify the best model, it is appropriate to make inferences based on that model alone. If not, turn to MA; however, there are still certain problems and pitfalls in using AIC. The so-called best model might consist of meaningless variables.
Turkheimer et al. (2003)	TR	AIC		X				MA reduces variance and bias compared to selecting a single model, especially when the true model is not contained in the candidate set. It also reduces generalization errors when applying the model to new data. Limitations include increased computation and requiring a common parameter across models.
Ullah & Wang (2013)	TR	AIC, BIC, Mallows' C_p, Cross Validation, GVY & FIC		X				Estimation and inference within the context of MS often overlook the inherent uncertainty associated with the selection process. This oversight frequently results in inefficient outcomes and confidence intervals that can be misleading. Conversely, MA mitigates model uncertainty.

Verrier et al. (2014)	TR	AIC, BIC & Bayesian			X			The authors are not explicit on which method is preferred; however, it seems to lean towards MA. This article focuses more on the modeling of the entire curve as an alternative.
Wheeler & Bailer (2009)	TR	AIC & BIC				X		It depends on the model space chosen. MA procedure averaged estimates from each of the 9 models' functional forms and not based on the average of individual models themselves.
Yang et al. (2022)	TR	RMSE			X			Monte Carlo simulation results demonstrate strong performance for both the MS and MA estimators in finite samples. Notably, the MS estimator exhibits asymptotic optimality, implying its efficiency is on par with the unattainable estimator utilizing the optimal spatial weights matrix in the limit.
Zhang et al. (2006)	TR	Maximum Likelihood Method & AIC		X				MS is limited to a subset of time-reversible models of nucleotide substitutions; therefore, MA can reduce biases arising from MS.
Zhang et al. (2012)	TR	AIC, BIC & FIC			X			The authors used both Tobit models and Monte Carlo simulations. There are merits in MA estimators over MS estimators, but it is not conclusive.
TOTAL			1	9	6	2	2	

Table 1 displays a summary of the literature review. This review is by no means comprehensive; nonetheless, it is still a fair representation of the diverse perspectives of MS and MA. Surprisingly, only a few of these discussions are contextualized within the field of DSML. Rather, most are concerned with traditional or Bayesian statistics. Hence, this literature gap awaits further investigation.

Nine of the 20 papers reviewed endorsed model averaging, six were inconclusive or said either was fine, two favored contextual usage, and two recommended combining the two. Only one favored model selection. However, it is important to note that this vote count is tallied by mixing MS and MA in traditional statistics, Bayesian statistics, and DSML. When considering only the articles related to DSML, it became evident that all of them preferred MA to MS.

Second, it was found that most studies comparing MS to MA overwhelmingly relied on AIC, BIC, or both. This implies that the avoidance of overfitting and preserving parsimony has become the dominant goal across the realms of traditional statistics, Bayesianism, and DSML. The phenomenon that analysts and researchers fully embrace AIC and BIC is understandable. As mentioned before, both AIC and BIC lean toward simplicity and penalize complexity. It can be explained by the fact that parsimony aligns with the principle of Occam's razor (Gamberger & Lavrač, 1997), suggesting that simpler explanations or models are generally preferred unless there is strong evidence to the contrary. This principle has a long history in the philosophy of science and is almost universally accepted.

Unlike complex models that are more sample-dependent, simpler models are more likely to generalize to new data and new contexts. Following this line of reasoning, simpler models are often more interpretable, making it easier to understand and communicate the relationships between variables. A model's interpretability is particularly important in fields like medicine, finance, and social sciences where decisions are made based on its interpretation. In contrast, complex models with many parameters, if not properly

regularized, can lead to overparameterization. When models are overparameterized, they can be computationally infeasible to train and may fail to converge to meaningful results. Following this line of reasoning, the popularity of AIC and BIC is understandable.

DISCUSSION

Although most researchers count on AIC and BIC for model comparisons, there is no single "best" metric. The "best" criterion depends on the objectives. If the analyst aims to maximize predictive accuracy, one might focus on criteria like mean squared error (MSE) or root mean squared error (RMSE). If the researcher wants to balance model fit and complexity, AIC or BIC might be appropriate. In both theory and practice, it is often advisable to consider multiple evaluation criteria, especially when comparing different models. This can provide a more comprehensive view of a model's performance, taking into account different aspects like accuracy, precision, recall, model complexity, and more.

Despite the fact that the majority of researchers prefer MA to MS, it is our conviction that choosing between MS and MA depends on the goal and availability of time and/or resources.

Generally speaking, model selection should be considered:

- When the goal is to identify a single model that can be used for both prediction and inference.
- When the number of candidate models is small.
- When the computational resources are limited.

On the other hand, model averaging should be taken into account:

- When the goal is to improve the predictive performance of the model.
- When the number of candidate models is large.
- When the computational resources are available.

Further, MA is based on the assumption that all models have certain merits that can contribute to the assembly of the final model, but in reality, it might not always be the case. Put bluntly, when traditional models, such as OLS regression models and stepwise regression models, are compared against their DSML counterparts, in most cases they are ranked at the bottom in terms of various criteria. In this case, it might not be beneficial to implement MA by incorporating less accurate predictors from these models. Further, domain knowledge plays a crucial role in MS and MA. If the so-called best model or final model carries variables that do not make sense (e.g., eating more ice cream can improve academic performance), then the analyst should toss out the variables no matter how good the numbers are and select the second-best model instead.

There is no definitive solution; there is always a trade-off. The researcher should consistently pose this question: *Does the additional work required for model averaging, in terms of time and effort, yield a significantly superior result compared to the single best model?* Regardless of what the response might be, model selection and model averaging are not mutually exclusive. The analyst can use model selection to choose a subset of candidate models, and then use model averaging to combine the predictions of those models. It is advisable to do so because the models near the bottom of the list might not have valuable information that is worth considering.

REFERENCES

Aoki, Y., Roshanmmar, D., Hamren, B., & Hooker, A. C. (2017). Model selection and averaging of nonlinear mixed-effect models for robust phase III dose selection. *Journal of Pharmacokinet Pharmacodyn, 44*, 581–597. DOI 10.1007/s10928-017-9550-0

Berge, T. J. (2015). Predicting recessions with leading indicators: Model averaging and selection over the business cycle. *Journal of Forecasting, 34*, 455–471. DOI: 10.1002/for.2345

Buatois, S., Ueckert, S., Frey, N., Retout, S., & Mentré, F. (2018). Comparison of model averaging and model selection in dose finding trials analyzed by nonlinear mixed effect models. *The AAPS Journal*, *20*, 1-9. DOI: 10.1208/s12248-018-0205-x

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel Inference* (2nd ed.). Springer-Verlag.

Colling, L., & Szucs, D. (2018). Statistical inference and the replication crisis. R*eview of Philosophy and Psychology, 12*(1), 121–147. https://doi.org/10.1007/s13164-018-0421-4

Dietterich, T. (2000). Ensemble methods in machine learning. Multiple Classifier Systems. https://doi.org/10.1007/3-540-45014-9_1

Gamberger, D., & Lavrač, N. (1997). Conditions for Occam's razor applicability and noise elimination. In Van Someren, M. & Widmer, G. (Eds.), *Machine Learning: ECML-97: 9th European Conference on Machine Learning Prague, Czech Republic, April 23–25, 1997 Proceedings* (9, pp. 108-123). Springer. (1997). https://doi.org/10.1007/3-540-62858-4_76

Gao, Y., Luo, M., & Zou, G. (2016). Forecasting with model selection or model averaging: a case study for monthly container port throughput. *Transportmetrica A: Transport Science*, *12*(4), 366-384. https://doi.org/10.1080/23249935.2015.1137652

Grainger, M. J., Aramyan, L., Piras, S., Quested, T. E., Righi, S., Setti, M., Vittuarai, M., Stewart, G. B. (2018). Model selection and averaging in the assessment of the drivers of household food waste to reduce the probability of false positives. LoS ONE 13(2): e0192075. https://doi.org/10.1371/journal.pone.0192075

Gu, Y., Wei, H. L., & Balikhin, M. M. (2018). Nonlinear predictive model selection and model averaging using information criteria. *Systems Science & Control Engineering*, *6*(1), 319-328. https://doi.org/10.1080/21642583.2018.1496042

Haggag, M. M. M. (2014). New criteria of model selection and model averaging in linear regression models. *American Journal of Theoretical and Applied Statistics*, *3*(5), 148-166. doi: 10.11648/j.ajtas.20140305.15

Harrell, F. E., Jr. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer

Lavery, R. et al. (2016). *An animated guide: Penalized variable selection techniques in https://documentation.sas.com/doc/en/statug/15.2/statug_glmselect_details14.htm and quantile regression*. Paper presented at Midwest SAS® User Group Conference.

Okoli, K., Breinl, K., Brandimarte, L., Botto, A., Volpi, E., & Di Baldassarre, G. (2018). Model averaging versus model selection: estimating design floods with uncertain river flow data. *Hydrological Sciences Journal*, *63*(13-14), 1913-1926. https://doi.org/10.1080/02626667.2018.1546389

Open Science Collaboration. (2015). Psychology. Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Peng, J., & Yang, Y. (2022). On improvability of model selection by model averaging. *Journal of Econometrics*, *229*(2), 246-262. https://doi.org/10.1016/j.jeconom.2020.12.003

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing, 27*(3), 711–735.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92,* 179–191.

Richards, S. A., Whittingham, M. J., & Stephens, P. A. (2011). Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behavioural Ecological Sociobiology, 65*, 77–89 DOI 10.1007/s00265-010-1035-8

SAS® Institute. (2020). Model selection issues. https://documentation.sas.com/doc/en/statug/15.2/statug_glmselect_details14.htm

Schomaker, M., & Heumann, C. (2014). Model selection and model averaging after multiple imputation. *Computational Statistics & Data Analysis*, *71*, 758-770. https://doi.org/10.1016/j.csda.2013.02.017

Schorning, K., Bornkamp, B., Bretz, F., & Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, *35*(22), 4021-4040. https://doi.org/10.1002/sim.6991

Smith, G. (2018). Step away from stepwise. *Journal of Big Data, 5*, 1-12. https://doi.org/10.1186/s40537-018-0143-6

Surowiecki, J. (2004). *The Wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* Anchor.

Symonds, M. R., & Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, *65*, 13-21. https://link.springer.com/article/10.1007/s00265-010-1037-6

Turkheimer, F. E., Hinz, R., & Cunningham, V. J. (2003). On the undecidability among kinetic models: From model selection to model averaging. *Journal of Cerebral Blood Flow and Metabolism, 23*, 490–498.

Ullah, A., & Wang, H. (2013). Parametric and nonparametric Frequentist model selection and model averaging. *Econometrics, 1*, 157-179. doi:10.3390/econometrics1020157

Verrier, D., Sivapregassam, S., & Solente, A. C. (2014). Dose-finding studies, MCP-Mod, model selection, and model averaging: two applications in the real world. *Clinical Trials*, *11*(4), 476-484. https://doi.org/10.1177/1740774514532723

Wheeler, M. W., & Bailer, A. J. (2009). Comparing model averaging with other model selection strategies for benchmark dose estimation. *Environmental and Ecological Statistics*, *16*, 37-51. https://link.springer.com/article/10.1007/s10651-007-0071-7

Yang, Y., Dogan, O., & Taspinar, S. (2022). Model selection and model averaging for matrix exponential spatial models. *Econometric Reviews, 41*(8), 827-858, DOI: 10.1080/07474938.2022.2047507

Yu, C. H. (2022a). *Data mining and exploration: From traditional statistics to modern data science.* CRC Press.

Yu, C. H. (2022b). Parametric tests. https://creative-wisdom.com/teaching/WBI/parametric_test.shtml

Zhang, X., Wan, A. T., & Zhou, S. Z. (2012). Focused information criteria, model selection, and model averaging in a Tobit model with a nonzero threshold. *Journal of Business & Economic Statistics*, *30*(1), 132-142. https://doi.org/10.1198/jbes.2011.10075

Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Wong, G. K. S., & Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, *4*(4), 259-263. https://doi.org/10.1016/S1672-0229(07)60007-2

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Chong Ho (Alex) Yu, PhD, DPhil
Hawai'i Pacific University
cayu@hpu.edu
chonghoyu@gmail.com
https://creative-wisdom.com/index.html
https://scholar.google.com/citations?user=mdGny3EAAAAJ&hl=en

Charlene J Yang, MA
Azusa Pacific University
cyang20@apu.edu
charlenejyang@gmail.com