American Journal of Theoretical and Applied Statistics
2014; 3(1): 1-5
Published online December 10, 2013 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20140301.11

Bayesian model averaging: An application to the determinants of airport departure delay in Uganda

Wesonga Ronald[1, *], Nabugoomu Fabian[2]

[1]School of Statistics and Planning, Makerere University, Kampala, Uganda
[2]Kyambogo University, Kampala, Uganda

Email address:
wesonga@wesonga.com (W. Ronald), fnabugoomu@yahoo.com (N. Fabian)

To cite this article:
Wesonga Ronald, Nabugoomu Fabian. Bayesian Model Averaging: An Application to the Determinants of Airport Departure Delay in Uganda. *American Journal of Theoretical and Applied Statistics.* Vol. 3, No. 1, 2014, pp. 1-5. doi: 10.11648/j.ajtas.20140301.11

Abstract: Bayesian model averaging was employed to study the dynamics of aircraft departure delay based on airport operational data of aviation and meteorological parameters collected on daily basis for the period 2004 through 2008 in matrix X. Models were evaluated using the R programming language mainly to establish the combinations of variables that could formulate the best model through assessing their importance. Findings showed that out of the sixteen covariates, 62.5% were suitable for model inclusion to determine aircraft departure delay of which 40% exhibited negative coefficients. The following parameters were found to negatively affect departure delay; number of aircrafts that departed on time (-0.562), number of persons on board of the arriving aircrafts (-0.002), daily average visibility (-0.001) and year (-1.605). Comparison between Posterior Model Probabilities (PMP Exact) and that based on Markov Chain Monte Carlo (PMP MCMC) revealed a high correlation (0.998; p<0.01).The study recommended the MCMC as providing a more efficient approach to modelling the determinants of aircraft departure delay at an airport.

Keywords: Airport Departure Delay, Prior, Posterior Model Probability, Model Selection

1. Introduction

Aviation provides an interesting example of interdependences of functions and activities. Aircraft delay is the most destabilizingcovariate in estimations ofairportoperational efficiency, in comparison to other determinants known to influence airport's efficiency at any given time of the day. Uncertainties always prevail when estimations are derived to determine suitable models for aircraft departure delay. To deal with model uncertainties that arise during model selection, Bayesian Model Averaging, BMAwas employed in the analysis[1].In this study, aviation and meteorological parameters for the period 2004 through 2008 in matrix X were evaluated mainly to establish the combinations of variables $X_\gamma \in \{X\}$ that could formulate the best model[2, 3]. Secondly, the study sought to establish the importance of the variables in the selected model for determinants of departure delay[4].

2. Methodology and Data

Parameter-based framework was recommended asone best suited to determine the probability of aircraft delay at an airport[5, 6]. However, the method did not fall short of the usual practice whereby standard statistical approaches tend toignore model uncertaintyanda model is selected from some class of models then proceed as if the selected model had generated the data[1]. In order to develop a coherent mechanism for accounting for model uncertainty, BMA[7] was employed in this study.

To develop a model for determinants of aircraft departure delay, we assumed a canonical regression[8-11]problem as presentedin Equation 1.

$$y = \alpha_\gamma + X_\gamma\beta_\gamma + \varepsilon_\gamma \tag{1}$$

$\varepsilon \sim N(0, \sigma^2 I) = error\ term$

$y\ \ = aircrafts\ delayed\ at\ departure\ per\ day$

$X_\gamma = covariates\ of\ y\ for\ model\ \gamma$

$\beta_\gamma = regression\ coefficients\ for\ model\ \gamma$

$\alpha_\gamma = regression\ constant\ for\ model\ \gamma$

The posterior model probability, PMP was assumed to be proportional to the marginal likelihood of the model times a prior model uncertainty [12, 13].Thus,

$$PMP \propto p(M_\gamma / y, X) \quad (2)$$

$$p(M_\gamma / y, X) = \frac{p(y/M_\gamma, X)p(M_\gamma)}{p(^{y}/_{X})} \quad (3)$$

The posterior model probability, PMP was used to answer how probable we were that the model M_γ would provide reliable estimates before the data could be examined.Re-normalization led to the PMPs and the model weighted posterior distribution for the statistic Θ. Thus, the β-coefficients were given by:

$$p(^{\theta}/_{y}, X) = \sum_{\gamma=1}^{2^k} p\left(^{\theta}/_{M_\gamma}, y, X\right) p(^{M_\gamma}/_{X}, y) \quad (4)$$

$p(M_\gamma) =$ our established beliefs about the covariates
$p(M_\gamma) \propto 1 =$ implied lack of knowledge apriori
$2^k =$ the number of variable combinations (the numb
$k =$ number of variables in the model

Subsequently, the posterior model probability obtained through thecomputation was considered to be the exact PMP. Given that MCMC lends another approach of computing PMPs[14] so a comparison of the results between the exact PMP and MCMC PMP[15, 16]was deemed necessary.

In thestudy, sixteen covariates in matrix X were assessed. The first question we attempted to answer was to establish which variables $X_\gamma \in \{X\}$would be included in the model. Secondly, we evaluated their importance in estimation of aircraft departure delay.One way was doing inference on a single model that could include all variables, but this proved to be inefficient and infeasible. We therefore employed the Bayesian model averaging; BMA that took the problem through estimating models for all possible combinations of {X} and then constructed a weighted average over all of them. Since X containedsixteen (16) potential variables as determinants, this meant estimating 2^{16}(65,536) variable combinations and thus the same number of models. The model weights for this averaging stemmed from posterior model probabilities that arise from the Bayes' theorem.

Specific expressions for maximum likelihoods $p\left(^{M_\gamma}/_{y}, X\right)$ and posterior distributions $p\left(^{\theta}/_{M_\gamma}, y, X\right)$ were found to depend on the chosen estimation framework. The literature standard was to use a 'Bayesian regression' linear model with a specific prior structure called the 'Zellner's g prior' that has remained robust and popular over time[17]. Thus, for each individual modelM_γ, we assumed a normal error structure as shown in Equation 1. This resulted into posterior distributions which are requiredto specify the priors on the model parameters. We then placed improper priors on the constant and error variance with an assumption that they were evenly distributed over their domain. Thus, $p(\alpha_\gamma) \propto 1$,so as to representa complete prior uncertainty where prior was located. Similarly, we set $p(\alpha_\gamma) \propto \alpha^{-1}$. The crucial prior was the one on regression coefficients β_γ assumed before looking into the data (y,X); we formulated prior beliefs on coefficients into a normal distribution with a specified mean and variance. We therefore, assumed a prior conservative mean of zero for the coefficients to reflect that much was known about regression coefficients. Their variance structure was defined according to Zellner's g: $\sigma^2(\frac{1}{g}X_\gamma^1 X_\gamma)^{-1}$[18]. Thus,

$$^{\beta_\gamma}/_{g} \sim N\left(0, \sigma^2\left(\frac{1}{g}X_\gamma^1 X_\gamma\right)^{-1}\right) \quad (5)$$

In this case, a small g would suggestthat there werefewer prior coefficient variances and therefore imply that wewere quite certain that the coefficients are indeed zero and vice versa. The smaller the value of g, the more important is the prior, and the more the expected value of coefficients wouldshrink towards the prior mean zero.However, when $g \to \infty$, the coefficient estimator approaches the ordinary least square estimator, OLS.

Similarly, the posterior variance of $\beta_\gamma$was affected by the choice of g[19]:

$$cov(\beta_\gamma, X, g, M_\gamma) = \frac{(y-\bar{y})^1(y-\bar{y})}{N-3}\frac{g}{1+g}\left(1-\frac{g}{1+g}R_\gamma^2\right)\left(X_\gamma^1 X_\gamma\right)^{-1} \quad (6)$$

This implied that the posterior covariance would be similar to that of the OLS estimator; times a factor that included g andR_γ^2, the OLS R-squared for model γ. The main form of the hyper-parameter, g, we used was the popular default approach of the unit information prior, UIP, which set g=N for all models and thus attributed about the same information to the prior as was contained in ourobservation[20].

The data for the study comprised of 1827 cases and 17 variables; one was the dependent while the rest were covariates. The variables were of scale type, derived from aviation[21] and meteorological parameters that affect departure of aircrafts at the airport while cases were daily records collected over a period of five years.

2.1. Algorithm for Model Selection

The following algorithm was developed and model analysis performed in R statistical computing language so as to develop the best model that could determine airport departure delay:

1. Estimated model parameter inclusion probability;
2. Examined performance of PMP (Exact) and PMP (MCMC) on selected best models and parameter inclusion;
3. Determined the parameter inclusion in the model while distinguishing between signs of their coefficients;
4. Examined how far the posterior model size distribution matchedup the prior on different indices of models.

3. Findings and Discussions

The dependent variable for the study was 'number of aircrafts that delay to depart'which was regressed on sixteen covariates so as to developthe best model. Table 1 shows parameter grading according to the posterior inclusion probabilities (PIP) for each of the sixteen model parameters. They represent the sum of PMPs for all models wherein a covariate is included. The higher PIPs signify the importance of the covariate in the model.Therefore, from the analysis, there were ten covariates with maximum PIPs of one, four of which had negative post means. The Bayesian model averaging computations used on average 11 regressors, 3000 draws, 1000 burnins in about 0.5 seconds on an Intel Core i5 computer processor.During the same time, the statistics including coefficients as presented were averaged over a total of 413 models visitedby applying the hyper parameter of unit information prior (UIP).

Table 1.*Posterior model estimation showing parameter inclusion probability*

Parameter	Inclusion Probability	Posterior Mean ($\bar{X}$)	Standard deviation(δ)
Scheduled flights	1.000	0.189	0.014
Freighters	1.000	0.161	0.036
Non-commercial flights	1.000	0.071	0.012
Aircrafts departing on time	1.000	-0.562	0.017
Aircrafts arriving on time	1.000	0.732	0.017
Aircrafts delaying to arrive	1.000	0.649	0.014
Persons on board-in	1.000	-0.002	0.001
Persons on board-out	1.000	0.002	0.001
Visibility	1.000	-0.001	0.001
Year (2004-2008)	1.000	-1.605	0.108
Queen's nautical height	0.588	0.004	0.003
Chartered aircrafts	0.306	0.011	0.018
Dew point temperature	0.057	0.001	0.011
Wind direction	0.034	0.000	0.001
Air temperature	0.019	0.001	0.005
Wind speed	0.0170	0.001	0.005

Note: (i) Columns show Posterior Inclusion Probability (value of 1 is highly desired) (ii) posterior mean whose sign show the effect of the parameter on the departure delay (iii) posterior standard deviation

The posterior mean coefficients show some interesting findings as far as their effect on aircraft departure delay is concerned. Four of the ten covariates found suitable for the model indicated negative coefficients. The four covariates were number of aircrafts that departed on time (-0.562), number of persons on board of the arriving aircrafts (-0.002), daily average visibility (-0.001) and year(-1.605). All the four covariates, according to the analysis confirmed that the higher their values, the lower were the number of aircrafts that delay to depart.

Markov Chain Monte Carlo, MCMC samplers were applied on the 16 covariates to gather results on the most important part of the posterior model distribution and thus approximate it as closely as possible. The MCMC method applied mostly relied on the Metropolis-Hastings algorithm which is known to walk through the model space[22]. The number of times each model was kept then converged to the distribution of posterior model probability $PMP = p\left({}^{M_i}/{}_{y}, X\right)$.Comparison of the top six models kept is shown in Table 2andthe MCMC PMPs showed statistically significant correlation withthe Exact PMPs (ρ=0.998, p<0.01) in the results of the computations.

Table 2.*Posterior model probability estimates for the exact and Markov Chain Monte Carlo for the six best models*

Parameter	Mod1	Mod2	Mod3	Mod4	Mod5	Mod6
Scheduled flights	1	1	1	1	1	1
Chartered flights	0	0	1	1	0	0
Freighters	1	1	1	1	1	1
Non-commercial flights	1	1	1	1	1	1
Aircraft on-time departures	1	1	1	1	1	1
Aircrafts arriving on time	1	1	1	1	1	1
Aircrafts delaying arrival	1	1	1	1	1	1
Persons on board-in	1	1	1	1	1	1
Persons on board-out	1	1	1	1	1	1
Wind direction	0	0	0	0	1	0
Windspeed	0	0	0	0	0	1
Visibility	1	1	1	1	1	1
Airtemperature	0	0	0	0	0	0
Dewpoint temperature	0	0	0	0	0	0
Queen's nautical height	1	0	1	0	1	1
Year	1	1	1	1	1	1
PMP (Exact)	0.39	0.27	0.14	0.10	0.01	0.01
PMP (MCMC)	0.36	0.26	0.14	0.11	0.02	0.01

Note: (i)1 indicates that parameter was included in the model (ii)0 indicates parameter exclusion from the model (iii)the last two rows show a comparison between Exact and MCMC approaches

From the best six models generated, we were able to see how different models treated different parameters in the estimation of determinants of aircraft departure delay. The binary digits were used to indicate performance of different parameters over the six best models with zero (0) indicating the variable did not qualifywhile one (1) indicates that the variable qualified and was considered under the prescribed model.

Figure 1 further shows the graphical illustration of the candidate variables selected from the best 20 modelsthat qualified for model inclusion. We note an agreement with the results in Table 1 when the posterior inclusion probabilities were used. We also note that the 20 best models represent a cumulative probability of 0.98and the selected variables have full colouring where dark shadesindicatedthat the variable had a positive influence while lighter shades showed that the variable had a negative influence to departure delay.It is further noted that for Entebbe International Airport in Uganda, 80% of the variables selected for the model are actually aviation parameters.Hence, according to the results, the only and most significant meteorological parameter selected and recommended for the model was visibility. Important to note further is the variability of these parameters over the days for the periodof the study.

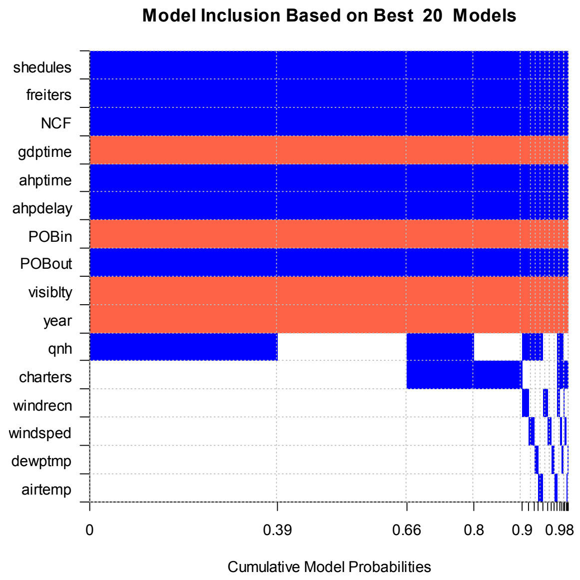

Figure 1.*Candidate model parameter against their cumulative probabilities*

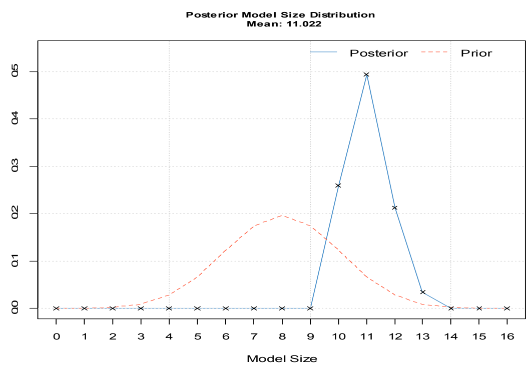

Figure 2.*Variations of posterior and prior model size distributions*

We examined how far the posterior model size distribution matched up the prior. Findings showed that the mean number of regressors was 11 whereby 100,000 draws were performed resulting in 50,000 burnins which lasted for about 18 seconds. The number of models visited was 14657 constituting 22% of the model space of 65536 to generate 100% of the top models for the study with the correlation of 0.9985 for the PMPs.

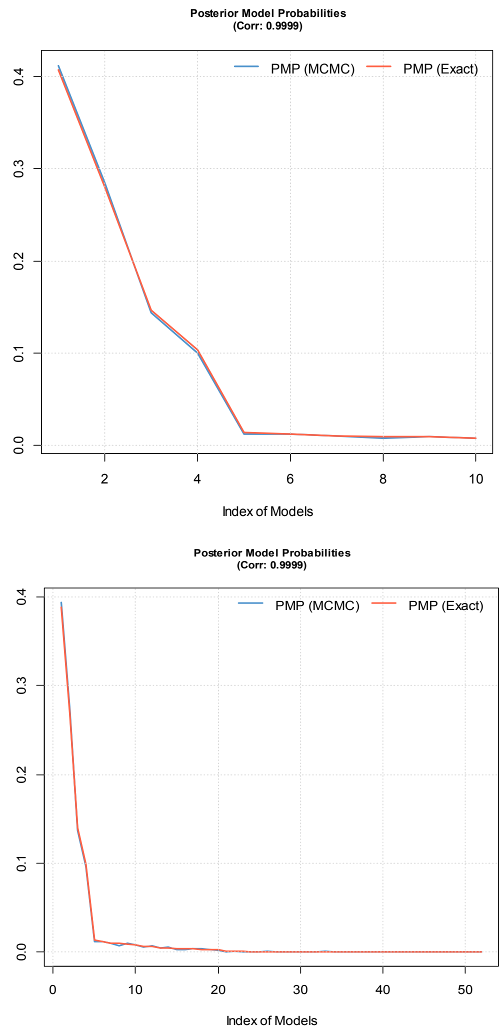

Figure 3. *Posterior model probabilities for PMP(Exact) and PMP (MCMC) when the index of models is 10 and 50 respectively*

As recommended [22],to evaluate and accelerate MCMC sampler convergence, application of diagnostic procedures to a small number of parallel chains, monitoring of autocorrelations and cross-correlations and where necessary modification ofparameterizations or sampling algorithms[19]were performed appropriately. In this studytherefore, before looking at the coefficients, we checked convergence of PMP (Exact) versus PMP (MCMC). Figure 3 shows that there was a sufficient convergence with high values of correlations for varying index of models 10 and 50 respectively.

4. Conclusion

Of the sixteen parameters, ten were recommended for inclusion into the model, six of which had positive coefficients; number of scheduled flights, number of freighters, number of non-commercial flights, proportion of flights arriving on time, proportion of aircrafts delaying to arrive, persons on board departing aircrafts. MCMC was found to perform equally and very often more efficiently with the correlation coefficient between PMP(Exact) and PMP(MCMC) (0.9985;$p<0.01$)which confirmedstatistically significant relationship between the two computational approaches. Forty percent(40%) of the parameters that qualified for model inclusion; aircrafts that departed on time, persons on-board of incoming aircrafts, visibility and year under study (2004-2008) exhibited negative signs. This was found to be in conformity with other findings[5, 21, 23].Although the posterior mean (11.022) varied from the prior mean (8.000), this still indicated a good estimate by the prior. The findings show that Markov Chain Monte Carlo provides a moreefficient approach to derive the best model for estimation of departure delay at Entebbe International Airport in Uganda. Like[24], this study established that efficient tools such as MCMC would facilitate optimal solutions through airport analysis, planning and design for demand and capacity to support sustainable development. Thus, further studies and development endeavors of the airport need to critically consider parameters derived and recommended in this study for efficient management of air traffic flow.

Acknowledgements

We extend our appreciation to the staff and management of the School of Statistics and Planning for the moral, material and financial support contributed towards this study. We also thank the management of the Civil Aviation Authority, Uganda and the Department of Meteorology for availing us with the data.Finally, weare highly indebted to the anonymous reviewers for their positive criticisms.

References

[1] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," Statistical science, pp. 382-401, 1999.

[2] M. Salibian-Barrera and S. Van Aelst, "Robust model selection using fast and robust bootstrap," Computational Statistics & Data Analysis, vol. 52, pp. 5121-5135, 2008.

[3] H.-B. Wang, "Bayesian estimation and variable selection for single index models," Computational Statistics & Data Analysis, vol. 53, pp. 2617-2627, 2009.

[4] A. Hadjaz, G. Marceau, P. Savaent, and M. Schoenauer, "Increasing Air Traffic: What is the Problem?," arXiv preprint arXiv:1212.3996, 2012.

[5] R. Wesonga, F. Nabugoomu, and P. Jehopio, "Parameterized framework for the analysis of probabilities of aircraft delay at an airport," Journal of Air Transport Management, vol. 23, pp. 1-4, 2012.

[6] E. L. Webb and J. J. Forster, "Bayesian model determination for multivariate ordinal and binary data," Computational Statistics & Data Analysis, vol. 52, pp. 2632-2649, 2008.

[7] P. Congdon, "Bayesian modelling strategies for spatially varying regression coefficients: A multivariate perspective for multiple outcomes," Computational Statistics & Data Analysis, vol. 51, pp. 2586-2601, 2007.

[8] R. J. Freund, W. J. Wilson, and P. Sa, Regression analysis : statistical modeling of a response variable, 2nd ed. Burlington, MA: Elsevier Academic Press, 2006.

[9] A. Alkenani and K. Yu, "A comparative study for robust canonical correlation methods," Journal of Statistical Computation and Simulation, vol. 83, pp. 692-720, 2013/08/06 2011.

[10] A. Alkenani and K. Yu, "A comparative study for robust canonical correlation methods," Journal of Statistical Computation and Simulation, vol. 83, pp. 692-720, 2013/08/06 2013.

[11] C. Matthew, C. R. O. Lawoko, C. J. Korte, and D. Smith, "Application of canonical discriminant analysis, principal component analysis, and canonical correlation analysis as tools for evaluating differences in pasture botanical composition," New Zealand Journal of Agricultural Research, vol. 37, pp. 509-520, 2013/08/06 1994.

[12] A. S. M. Arroyo, A. Garca-Ferrer, A. de Juan Fernandez, and R. Sanchez-Mangas, "Lower posterior death probabilities from a quick medical response in road traffic accidents," Journal of Applied Statistics, vol. 40, pp. 40-58, 2013/08/06 2013.

[13] R. C. Weng and C.-H. Hsu, "A Study of Expansions of Posterior Distributions," Communications in Statistics - Theory and Methods, vol. 42, pp. 346-364, 2013/08/06 2013.

[14] G. O. Roberts and J. S. Rosenthal, "Examples of Adaptive MCMC," Journal of Computational and Graphical Statistics, vol. 18, pp. 349-367, 2013/08/08 2009.

[15] T. H. Oum, J. Yan, and C. Yu, "Ownership forms matter for airport efficiency: A stochastic frontier investigation of worldwide airports," Journal of Urban Economics, vol. 64, pp. 422-435, 2008.

[16] N. Xu, G. Donohue, K. B. Laskey, and C.-H. Chen, "Estimation of delay propagation in the national aviation system using Bayesian networks," in 6th USA/Europe Air Traffic Management Research and Development Seminar, 2005.

[17] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for Bayesian variable selection," Journal of the American Statistical Association, vol. 103, 2008.

[18] D. G. Brown, S. Page, R. Riolo, M. Zellner, and W. Rand, "Path dependence and the validation of agents' based spatial models of land use," International Journal of Geographical Information Science, vol. 19, pp. 153-174, 2005.

[19] M. M. Nassar, S. M. Khamis, and S. S. Radwan, "On Bayesian sample size determination,"