Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015
11-13 August, 2015 Istanbul, Turkey. Universiti Utara Malaysia (http://www.uum.edu.my)

Paper No.
218

ALGORITHMIC APPROACHES IN MODEL SELECTION OF THE AIR PASSENGERS FLOWS DATA

Suzilah Ismail[1], Norhayati Yusof[2], and T Zalizam T Muda[3]

[1]*Universiti Utara Malaysia, Malaysia, halizus@uum.edu.my*
[2]*Universiti Utara Malaysia, Malaysia, norhayati@uum.edu.my*
[3]*Universiti Utara Malaysia, Malaysia, zalizam@uum.edu.my*

ABSTRACT. Algorithm is an important element in any problem solving situation. In statistical modelling strategy, the algorithm provides a step by step process in model building, model testing, choosing the 'best' model and even forecasting using the chosen model. Tacit knowledge has contributed to the existence of a huge variability in manual modelling process especially between expert and non-expert modellers. Many algorithms (automated model selection) have been developed to bridge the gap either through single or multiple equation modelling. This study aims to evaluate the forecasting performances of several selected algorithms on air passengers flow data based on Root Mean Square Error (RMSE) and Geometric Root Mean Square Error (GRMSE). The findings show that multiple models selection performed well in one and two step-ahead forecast but was outperformed by single model in three step-ahead forecasts.

Keywords: manual model selection, automated model selection, single equation, multiple equations

INTRODUCTION

Basically an algorithm is an important element in any problem solving situation. In statistical modelling strategy, the algorithm provides a step by step process in model building, model testing, choosing the 'best' model and even forecasting using the chosen model. Experiment organized by Magnus and Morgan (1999) demonstrated that different modellers with the same methodological approach specified several models for a given data set. It proves the existence of a huge variability in manual modelling process especially between expert and non-expert modellers due to tacit knowledge. Since then, many algorithms (automated model selection) have been developed to bridge the gap such as *PcGets* (Hendry & Krolzig, 2001) and *Autometrics* (Doornik, 2009). These algorithms however focused on single equation modelling. The extended algorithms developed for multiple equations modelling specifically the seemingly unrelated regression equations are *SURE-PcGets* (Ismail, 2005) and *SURE-Autometrics* (Yusof & Ismail, 2014), respectively. The main element in these algorithms is the search procedure in finding the 'best' parsimonious model from a very general model. Thus, it is a model selection approach. This study aims to evaluate the forecasting performances of several selected algorithms on a real data set (i.e. air passengers flow data).

The evaluation using empirical data is very crucial in identifying whether these algorithmic approaches exhibits 'data mining' characteristics which has very common problem amongst model builders since data mining permitted the selection of best models withinsample fitted model and able to satisfy all measures of goodness of fits. However, the data mining models might fail when it comes to forecasting.

MODEL SELECTION ALGORITHMS

In this study, the model selection algorithms are classified into three approaches. The first is the individual selection approach for single model while employing an Ordinary Least Squares (OLS) method of estimation which are *Stepwise*, *Autometrics* and *MINE*. The *Stepwise* starts from an empty model, adding a variable and remove if it is insignificant. The process continues until no more variables can be added into the model and often failed in finding the best model (Lovell, 1983; Whittingham, Stephens, Bradbury, & Freckleton, 2006). Hence, *PcGets* (Hendry & Krolzig, 2001, 2002; Hoover and Perez, 1999) is introduced. Unlike *Stepwise*, this algorithm starts from the other end which is from a general model that comprised of all variables, and it is reduced to a simpler model using a 'testing-down process by eliminating variables with coefficients that are not statistically significant. Both techniques are known as expanding or specific-to-general and contracting or general-to-specific (GETS) method (Hendry & Doornik, 2014). An algorithm that contains hybrid of these methods is known as the *Autometrics* (Doornik & Hendry, 2007; Doornik, 2009). The algorithm implements a tree search that systematically navigates the whole model space. However, to find the all possible models is a computationally inefficient. Thus, several strategies such as pruning, bunching, and chopping are implemented to cut-off irrelevant paths and speed up the process. These will achieved the goal of *Autometrics* in improving the computational strategies by avoiding repeated estimation of the same model, diagnostic tests delayed, and recollect terminals between iterations. *Stepwise* and *Autometrics* are automated model selections as for *MINE* is manual selection procedures by employing our own tacit knowledge based on theory and judgment in statistical modelling.

The second approach is a simultaneous selection of multiple models while employing a Feasible Generalised Least Squares (FGLS) as a method of estimation. The algorithms included are the *SURE-Autometrics*, *SURE-PcGets*, and *SURE-MINE*. There are many types of multiple equations but this study focuses on a seemingly unrelated regression equations (SURE) model. This model introduced by Zellner (1962) to increase the efficiencies in several single equations that are related through the disturbances amongst equations thus the named is seemingly unrelated. The *SURE-Autometrics*, *SURE-PcGets* and *SURE-MINE* are the extended version of Autometrics, *PcGets* and *MINE*, respectively from the application of single equation to multiple equations modelling.

The last category involves individual selection with OLS estimation, except the final selected multiple models employs FGLS method of estimation. Thus, the procedures involved are *Autometrics-SURE*, *Stepwise-SURE* and *MINE-SURE*. Hence, there are nine different model selections algorithmic approaches involve in this study.

ANALYSIS AND FINDINGS

The data set used in this study is from Fildes et. al (2011) which comprise the total annual passenger (dependent variable, Y_{it}) from and to UK based on six countries (Germany, Sweden, Italy, Japan, USA and Canada). Figure 1 displays the trends of air passenger according to countries from 1961 to 2002. Overall the trends are increasing where the lowest is Japan and the highest is USA but decrease in 2001 and 2002 due to terrorist incidents in September 11, 2001.

The independent variables included are income (x_{i1t}), trade (x_{i2t}), price (x_{i3t}) and 'world' trade (x_{i4t}), population (x_{i5t}), gross domestic products (GDP, x_{i6t}) and consumer price index (CPI, x_{i7t}) which had proved important in earlier studies of the demand for air travel (Fildes et. al, 2011; Jorge-Caleron, 1997; Kaemmerle, 1991; Quandt and Baumol, 1966; O'Conner, 1989). Autoregressive Distributed Lag (ADL) model was used. Therefore additional of three lags of Y_{it} and one lag of each independent variables are included in the general model as independent variables, thus a total number of independent variables used in this study are 17. The data is transformed by taking log and first differencing to achieve stationarity. The first thirty eight data is used for model estimation and the last five is for model evaluation (i.e. recursive evaluation) which based on Root Mean Square Error (RMSE) and Geometric Root Mean Square Error (GRMSE).

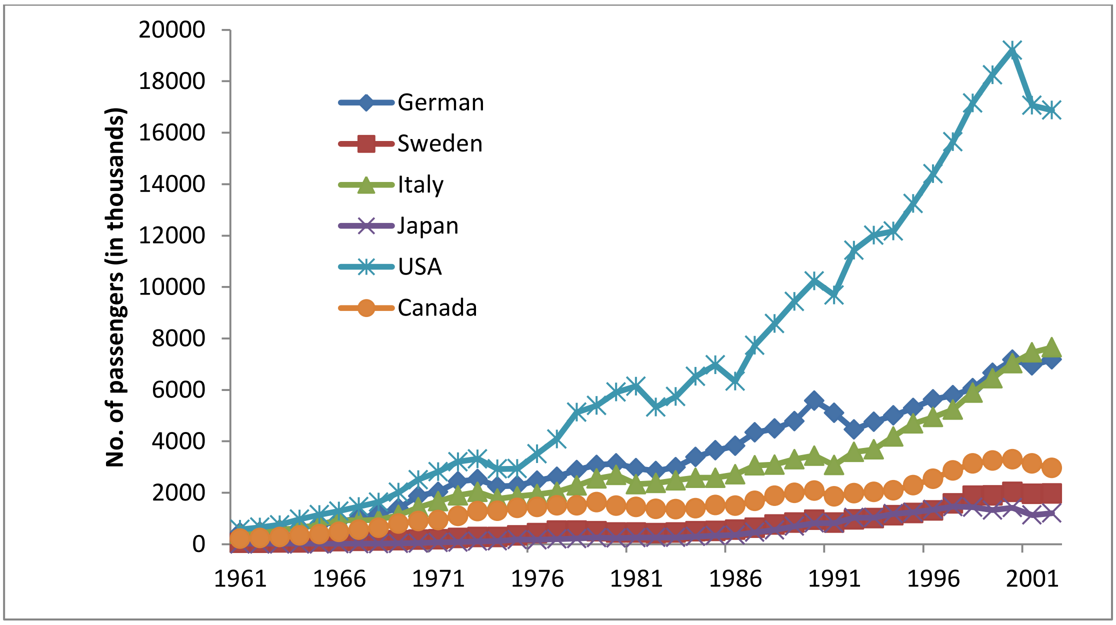

Figure 1. International passenger from and to UK.

Table 1 presents the adjusted R square ($\bar{R}^2$) and standard error (SE) based on different model selection algorithm. Canada has the highest ($\bar{R}^2$) while Japan and Sweden have the lowest. It is very obvious Japan has the largest standard error (SE). Perhaps this is due to different passenger behaviour where Japan is the only Asia country involves in this study and also long haul route as compared to other western countries.

Table 2 and 3 display the evaluation results for algorithmic approaches for one, two and three step ahead forecast. There are difference findings based on RMSE and GRMSE where in one and two steps, the RMSE indicate individual selection approach for single model as the 'best' approach (rank 1) but GRMSE oppositely specify multiple models approach (*SURE-PcGets*). This is due to large errors (or outliers) cause by the down fall of air passenger in 2001 and 2002 data which related to September 11, 2001. Since RMSE are easily affected by outliers (Lazim, 2011) therefore GRMSE is more appropriate to be used. Based on the GRMSE in one and two step-ahead forecast (Table 3), multiple models selection algorithm *SURE-PcGets*, outperformed the individual selection approach for single model but in three-step ahead forecast *Autometrics* and *Stepwise* (single model) performed the 'best'. It is also noticeable, the automated model selection approach performed better than manual (*MINE*, *SURE-MINE* and *MINE-SURE*).

Table 1. Algorithmic Approaches and countries

Approaches		**Germany**	**Sweden**	**Italy**	**Japan**	**US**	**Canada**
1. *Stepwise*	$\bar{R}^2$	0.370	0.132	0.516	0.131	0.191	0.650
	SE	0.071	0.095	0.061	**0.152**	0.082	0.047
2.*Autometrics*	$\bar{R}^2$	0.370	0.132	0.516	0.131	0.191	0.650
	SE	0.071	0.095	0.061	**0.152**	0.082	0.047
3.*MINE*	$\bar{R}^2$	0.370	0.132	0.516	0.264	0.302	0.650
	SE	0.071	0.095	0.061	**0.140**	0.076	0.047
4.*SURE-Autometrics*	$\bar{R}^2$	0.376	0.198	0.512	0.206	0.247	0.581
	SE	0.067	0.084	0.058	**0.137**	0.076	0.076
5.*SURE-PcGets*	$\bar{R}^2$	0.349	0.102	0.230	0.068	0.165	0.576
	SE	0.068	0.091	0.075	**0.150**	0.081	0.048
6.*SURE-MINE*	$\bar{R}^2$	0.403	0.132	0.552	0.277	0.298	0.553
	SE	0.064	0.092	0.054	**0.128**	0.072	0.048
7. *Autometrics-SURE*	$\bar{R}^2$	0.367	0.132	0.512	0.130	0.191	0.636
	SE	0.068	0.092	0.058	**0.148**	0.080	0.043
8. *Stepwise-SURE*	$\bar{R}^2$	0.367	0.132	0.512	0.130	0.191	0.636
	SE	0.068	0.092	0.058	**0.148**	0.080	0.043
9. *MINE-SURE*	$\bar{R}^2$	0.365	0.132	0.515	0.243	0.301	0.637
	SE	0.068	0.092	0.058	**0.131**	0.072	0.043

Table 2. Forecasting Performances Based on RMSE

	One-Step		**Two-Step**		**Three-Step**	
Approaches	**RMSE**	**Rank**	**RMSE**	**Rank**	**RMSE**	**Rank**
1. *Stepwise*	8.60	**1**	9.37	**1**	10.19	3
2. *Autometrics*	8.60	**1**	9.37	**1**	10.19	3
3. *MINE*	8.60	**1**	9.37	**1**	10.19	3
4. *SURE-Autometrics*	8.77	8	9.65	9	9.35	2
5. *SURE-PcGets*	8.71	7	9.42	7	9.06	**1**
6. *SURE-MINE*	8.84	9	9.47	8	10.55	9
7. *Autometrics-SURE*	8.63	5	9.38	5	10.21	7
8. *Stepwise-SURE*	8.63	5	9.38	5	10.21	7
9. *MINE-SURE*	8.61	4	9.37	**1**	10.20	6

*Tied elements are assigned to the lowest rank.

Table 3. Forecasting Performances Based on GRMSE

	One-Step		Two-Step		Three-Step	
Approaches	**GRMSE**	**Rank**	**GRMSE**	**Rank**	**GRMSE**	**Rank**
1. *Stepwise*	4.99	4	7.13	5	6.33	**1**
2. *Autometrics*	4.99	4	7.13	5	6.33	**1**
3. *MINE*	5.78	9	7.50	9	8.58	8
4. *SURE-Autometrics*	5.15	6	7.18	7	8.25	6
5. *SURE-PcGets*	4.30	**1**	6.87	**1**	6.69	5
6. *SURE-MINE*	5.64	8	6.98	2	8.70	9
7. *Autometrics-SURE*	4.92	2	7.09	3	6.67	3
8. *Stepwise-SURE*	4.92	2	7.09	3	6.67	3
9. *MINE-SURE*	5.48	7	7.48	8	8.53	7

CONCLUSION

Multiple models selection algorithms performed well in one and two step-ahead forecast but in three step-ahead individual selection approach for single model (*Stepwise* and *Autometrics*) is the 'best'. Perhaps this is due to large error in Japan model where pooling the models in *SURE* affected the performance of multiple models selection. Based on this study, automated model selection outperformed manual model selection.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Ministry of Higher Education for the financial support under the Fundamental Research Grant Scheme and to Universiti Utara Malaysia for facilitating the management of the research.

REFERENCES

Doornik, J. A. (2009). Autometrics. In J. L. Castle & N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry* (pp. 88–121). New York: Oxford University Press.

Doornik, J. A., & Hendry, D. F. (2007). *Empirical Econometric Modelling using PcGive 12: Volume 1*. London: Timberlake Consultants Ltd.

Hendry, D. F., & Doornik, J. A. (2014). *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*. MIT Press.

Hendry, D. F., & Krolzig, H.-M. (2001). *Automatic Econometric Model Selection Using PcGets 1.0*. London: Timberlake Consultans Press.

Hendry, D. F., & Krolzig, H.-M. (2003). New Developments in Automatic General-to-Specific Modeling. In *Econometrics and the Philosophy of Economics: Theory-data confrontations in economics* (pp. 379–419). Princeton: Princeton University Press.

Hoover, K. D., & Perez, S. J. (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, *2*, 167–191.

Ismail, S. (2005). *Algorithmic approaches to multiple time series forecasting*. *Department of Management Science*. University of Lancaster, Lancaster.

Lazim, M. A. (2011). *Introductory Business Forecasting. A Practical Approach*. Shah Alam: Pusat Penerbitan Universiti, Universiti Teknologi Mara.

Lovell, M. C. (1983). Data mining. *The Review of Economics and Statistics*, *65*(1), 1–12.

Magnus, J. R., & Morgan, M. S. (1999). Methodology and tacit knowledge: Two experiments in econometrics. New York: John Wiley.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182–1189. doi:10.1111/j.1365-2656.2006.01141.x

Yusof, N., & Ismail, S. (2014). Lag variable reduction in multiple models selection. In *International Conference on the Analysis and Mathematical Applications in Engineering and Science* (pp. 169–173). Curtin University Sarawak, Malaysia.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, *57*(298), 348–368.