Mathematics and Statistics 10(1): 222-232, 2022
DOI: 10.13189/ms.2022.100121 http://www.hrpub.org

Expectation-Maximization Algorithm Estimation Method in Automated Model Selection Procedure for Seemingly Unrelated Regression Equations Models

Nur Azulia Kamarudin*, Suzilah Ismail, Norhayati Yusof

Department of Mathematics & Statistics, School of Quantitative Sciences, Universiti Utara Malaysia,
06010 UUM Sintok, Kedah, Malaysia

Received August 27, 2021; Revised December 13, 2021; Accepted December 21, 2021

Cite This Paper in the following Citation Styles

(a): *[1] Nur Azulia Kamarudin, Suzilah Ismail, Norhayati Yusof , "Expectation-Maximization Algorithm Estimation Method in Automated Model Selection Procedure for Seemingly Unrelated Regression Equations Models," Mathematics and Statistics, Vol. 10, No. 1, pp. 222 - 232, 2022. DOI: 10.13189/ms.2022.100121.*

(b): *Nur Azulia Kamarudin, Suzilah Ismail, Norhayati Yusof (2022). Expectation-Maximization Algorithm Estimation Method in Automated Model Selection Procedure for Seemingly Unrelated Regression Equations Models. Mathematics and Statistics, 10(1), 222 - 232. DOI: 10.13189/ms.2022.100121.*

Abstract Model selection is the process of choosing a model from a set of possible models. The model's ability to generalise means it can fit both current and future data. Despite numerous emergences of procedures in selecting models automatically, there has been a lack of studies on procedures in selecting multiple equations models, particularly seemingly unrelated regression equations (SURE) models. Hence, this study concentrates on an automated model selection procedure for the SURE model by integrating the expectation-maximization (EM) algorithm estimation method, named *SURE(EM)-Autometrics.* This extension procedure was originally initiated from *Autometrics,* which is only applicable for a single equation. To assess the performance of *SURE(EM)-Autometrics,* simulation analysis was conducted under two strengths of correlation among equations and two levels of significance for a two-equation model with up to 18 variables in the initial general unrestricted model (GUM). Three econometric models have been utilised as a testbed for true specification search. The results were divided into four categories where a tight significance level of 1% had contributed a high percentage of all equations in the model containing variables precisely comparable to the true specifications. Then, an empirical comparison of four model selection techniques was conducted using water quality index (WQI) data. System selection to select all equations in the model simultaneously proved to be more efficient than single equation selection. *SURE(EM)-Autometrics* dominated the comparison by being at the top of the rankings for most of the error measures. Hence, the integration of EM algorithm estimation is appropriate in improving the performance of automated model selection procedures for multiple equations models.

Keywords Expectation-Maximization Algorithm, Automated Model Selection, Multiple Equations, Seemingly Unrelated Regression Equations

1. Introduction

To make scientific discoveries or anticipate future outcomes, data analysts use a variety of statistical models and methodologies to analyse observable data. Regardless of the data and fitting processes used, selecting the best acceptable model or approach from a pool of candidates is an important step. An essential part of data analysis for scientific investigations is model selection, which is essential for obtaining accurate statistical inferences or predictions [1], [2]. The model selection procedure begins with an estimation of a model, which the researcher first stated. Afterwards, the results of hypothesis tests of the single parameters are used to identify significant variables or to conduct diagnostic checking for the assumptions of

the model [3]. This entire procedure may be done automatically or manually. However, it is quite difficult to perfect this intuitive judgement in manual model selection. Repetitive manual retraining and recalibrating of models are frequently prohibitively expensive, time-consuming, and in some circumstances impossible [4].

During the model development process, different researchers may develop distinct modelling paradigms or strategies. Therefore, a number of alternative models for the same data set are created even though they are based on the same methodology. The different models that arise illustrate that variability may occur when the model specification procedure is performed manually. Finally, this circumstance contributes to the creation of a gap between specialists and novice users, where novice users may include beginners in statistical or economic modelling who struggle to understand the model itself [1]. This difference in understanding has inspired the demands for an automated technique for a more convenient and definitive answer. With the advancement and spreading of data modelling, the necessity of automated model specification has grown exponentially. In bridging the gap between the experts and novices in model selection, the employment of expert systems approaches seems an optimal solution. The specification of multiple equation models is an example of this challenging task [5].

1.1. Seemingly Unrelated Regression Equations (SURE) Models

Multiple equation models apply to a group of related variables in some situations. With these complex factors present, it is reasonable to evaluate the models simultaneously as a system of equations. To view the equations together rather than individually, the word "system" is used. The advantages may be appreciated further by combining all the equations to describe the dynamic composition of the real operation. Additional information may be provided via a group of equations rather than by the addition of single equations. The more knowledge about the causal linkages and structures included are known, for instance, the more accurate the forecasts will be obtained [6].

Zellner [7] proposed the SURE model, which uses multiple equations and is an extension of the standard linear regression model. Every equation can be estimated independently, even though the error terms are considered to be correlated across equations. Because each equation must be able to stand on its own, each equation has a dependent variable as well as potentially separate sets of regressors. The results proved that these equations were unconnected because of their individuality but were nonetheless linked because of the error terms.

The SURE model is frequently used in the fields of economics and financial modelling [8]–[10]. It may, however, be used in other disciplines as well, such as transportation [11]–[13], agricultural [14], management [15] and medical [16]. More studies had also been spurred by improvements and expansions to the original equations [17]–[19]. As a result, the SURE model is applicable in nearly every element of life.

SURE modelling is presented to gain estimation efficiency by integrating information on several equations and imposing or testing restrictions that include parameters in separate equations [20]. Assume that the equations are in series,

$$\begin{aligned} y_{1t} &= \beta_{11}x_{1t,1} + \beta_{12}x_{1t,2} + \cdots + \beta_{1k_1}x_{1t,k_1} + u_{1t} \\ y_{2t} &= \beta_{21}x_{2t,1} + \beta_{22}x_{2t,2} + \cdots + \beta_{2k_1}x_{2t,k_2} + u_{2t} \\ &\vdots \\ y_{mt} &= \beta_{m1}x_{mt,1} + \beta_{m2}x_{mt,2} + \cdots + \beta_{mk_1}x_{mt,k_m} + u_{mt} \end{aligned}$$

which can generally be written as,

$$\underset{T\times 1}{\mathbf{y}_i} = \underset{T\times k_i}{\mathbf{X}_i}\ \underset{k_i\times 1}{\beta_i} + \underset{T\times 1}{\mathbf{u}_i}$$

where $\mathbf{y}_i$ is the vector of T identically distributed observations for each random variable, $\mathbf{X}_i$ is a nonstochastic matrix of fixed variables of rank k_i, $\boldsymbol{\beta}_i$ is the vector of unknown coefficients, and u_i is a vector of disturbances.

1.2. Estimation Methods

In economic analysis, the SURE model may be identified in several instances [7]. One example is when the equations have uncorrelated disturbances occurring at the same time. If all of the response variables are believed to be connected by the same regressors, this is a typical multivariate linear regression model. Ordinary least squares (OLS) provide the most effective and best linear unbiased estimators. Another example is when the equations have the same regressor, but the disturbances are contemporaneously linked. Because the OLS ignores the correlations of the disturbances, it is commonly accepted that the generalised least squares (GLS) estimate technique is more efficient. Nevertheless, in the majority of cases, the covariance of disturbances Ω is unidentified, making GLS impractical. As a result, feasible generalised least squares (FGLS) using a consistent estimator in place of Ω was presented [7].

Srivastava and Giles [20] found that most studies comparing FGLS to OLS used models with two equations. However, when sample sizes are limited and/or disturbance correlation coefficients in SURE models are near zero, FGLS can be less efficient than OLS. Moreover, OLS outperforms FGLS when disturbance correlations are weak. The effectiveness of FGLS may be lost in cases when all equations have identical regressors, as in a vector regressive system. FGLS can potentially be risky when missing variables are specified. Because the equations are treated as a system, one error in one equation can affect the estimations in others. Unlike OLS estimation, which only applies to one equation at a time. Other equations with no specification errors will be unbiasedly calculated. As a

result, some researchers still oppose FGLS estimation since single equation misspecification might affect all estimates in the system Additionally, the FGLS does not exist in the high-dimensional SURE model [21]. Thus, FGLS is not always an effective estimator, and therefore another estimation method for the SURE model is being investigated in this study, which is the EM algorithm.

The EM algorithm is also applicable to the SURE model, which is considered a repeated measures analysis. The SURE model's regression parameters and variance-covariance matrix may be estimated using a two-stage Aitken estimation procedure. Additionally, the EM method is considered simple to implement because it does not entail any evaluation of the likelihood or its derivatives, besides being able to estimate the 'missing' data values [22].

McLachlan and Krishnan [22] and Ng et al. [23] listed some attractive properties of the EM algorithm in relation to other iterative algorithms such as Newton-Raphson and Fisher's scoring method for searching estimators including:

i. The EM algorithm is numerically stable.
ii. The EM algorithm normally has consistent global convergence.
iii. The EM algorithm is easy as it relies on calculations of complete data.
iv. The EM algorithm is normally easy to programme due to no evaluation of the likelihood or its derivatives are involved.
v. Only small storage is needed for the EM algorithm and thus, can be done on a small computer.
vi. Low cost for each iteration offsets the bigger number of iterations required for the EM algorithm compared to other procedures.
vii. Estimated values of the 'missing' data can be computed using the EM algorithm.

Sohn [24] discovered numerous useful characteristics of the EM method when used to estimate integrated choice and latent variable models. The EM algorithm significantly reduced the calculation time since it does not involve any time-consuming numerical computation of derivatives or the Hessian of the simulated likelihood function. Additionally, it avoids a lack of empirical identification. Additionally, the EM algorithm excelled at decreasing sampling errors when the sample size was small (250 and 500). Even when large samples of 1000 and 2000 were used, the EM algorithm outperformed its competitor. Furthermore, Ryzin and Vulcano [25] demonstrated that the convergence time for the EM algorithm is between twice and six times faster than that of the direct ML estimator, while the quality of the estimates is comparable. Additionally, their findings were validated in experiments including many parameters or a high degree of censoring.

Due to the extensive advantages of the EM algorithm, many studies did employ EM in a wide range of applications, for example, pattern recognition [26], biomedical and health [27]–[29], nature [30] and crime [31]. Overall, the EM algorithm is an outstanding tool to be studied.

1.3. Automated Model Selections

The work by Hoover and Perez [32] was a pioneer in the field of automated model selection. Hendry and Krolzig [33] built on Hoover Perez's work by improving data mining algorithms and developed *PcGets*, a software for empirical modellers. Multiple equations models can be selected by using one equation at a time or all equations at the same time. The automated approach has facilitated these selections in a fraction of the time required by the human approach. *SURE-PcGets* by Ismail [34] is one such method in which *PcGets* selection stages are integrated with the SURE model. *SURE-PcGets* updated testing for contemporaneous correlation disturbances, and the model formulation of PcGets and showed that the simulation results demonstrated the algorithm's efficacy in finding the true model.

Doornik and Hendry [35] then created *Autometrics* by utilising the notion of the tree search strategy. *Autometrics* algorithm is comparable to *PcGet*s. *Autometrics*, on the other hand, employs a tree search approach with adjustments to the pre-search simplification and goal function. Yusof [36] created *SURE-Autometrics* in response to the success of *SURE-PcGets* by utilising *Autometrics*, another single equation selection technique. This method also makes use of the SURE model. The tree search strategy incorporated in *Autometrics* enables extensive exploration of pathways and enhances the likelihood of discovering terminal models using pruning, bunching, and chopping reduction principles. This would finally result in the final model being the 'best' of all.

SURE-Autometrics was constructed by adapting Ismail's algorithm [34]. In this study, two approaches were considered: (i) a model selected by OLS compared a model selected by FGLS; and (ii) a model selected by OLS and subsequently estimated by FGLS, and against a model selected and estimated by FGLS concurrently. The former strategy was to establish how substantially estimation of single-equation and system varied, whilst the latter sought to illustrate the distinctions between single and simultaneous selection processes. Yusof [36] demonstrated that when the correlation strength was strong, *SURE-Autometrics* was able to remove more irrelevant variables when the model was selected and estimated concurrently.

This had prompted another approach in utilizing iterative feasible generalized least squares (IFGLS) estimation method instead of FGLS by using an extended version of *SURE-Autometrics*, known as *SURE(IFGLS)-Autometrics* [37], [38]. The results were found to support the earlier findings of *SURE-Autometrics* but showed slightly better performance than *SURE-Autometrics*. As a consequence of these initial

attempts to use *Autometrics* to integrate the joint estimation and model selection of the whole SURE system, a new avenue of study into the SURE system has been discovered further. Because of this, the purpose of this study was to evaluate the performance of *SURE-Autometrics*, but with the use of the EM algorithm estimation method. The algorithm has been renamed *SURE(EM)-Autometrics* to better reflect its capabilities.

2. Methodology

2.1. *SURE(EM)-Autometrics*

The development of *SURE(EM)-Autometrics* continues to take on original *SURE-Autometrics* with five phases described in this section. The original *SURE-Autometrics* algorithm was changed by the usage of the EM in contrast to the FGLS estimation method for SURE models. Therefore, the EM algorithm estimation method is applied in each phase of *SURE(EM)-Autometrics*.

Phase 1: Specification of initial GUMS

The algorithm begins with Phase 1 where every equation in the SURE model is specified by the modeller for its initial specification. This includes the number of equations, the main level of significance and the variables with their lags. Meanwhile, OLS is used to estimate the single equations in the SURE model separately. Any misspecification of the equations is also checked through diagnostic testing such as tests for normality errors, parameters constancy, autocorrelation, unconditional homoscedasticity, and conditional homoscedasticity. This phase also consists of testing of contemporaneous correlation of disturbances among equations and EM algorithm estimation.

Phase 2: Pre-search Reduction

The purpose of this phase is to lessen the computational effort where the highly insignificant variables are removed. Nevertheless, the whole algorithm can still function even with the absence of Phase 2. There are three types of reductions used to remove insignificant variables: closed, common, and common-X lag reductions. The closed lag reduction is designed to test a set of lags starting with the biggest lag and ending when a lag cannot be removed. Meanwhile, the common lag reduction is to test grouped lags with the same lag number and organize the joint (common) significances beginning with the least significant lag group. Lastly, the common–X lag reduction follows similar steps as common lag reduction, but the lag of Y is excluded from the procedure. The three lag reductions are employed in chronological order and again in reverse chronological order. Finally, the encompassing test is implemented to make certain that the reduced model is a legitimate reduction of the initial system of GUM.

Phase 3: Variable Reduction over Root Branches

Phase 3 involves a tree search in which the whole space of models created by variables in the initial model is searched. In the tree search, three main principles are involved; (i) pruning is executed when a single variable is being evaluated for removal (ii) bunching is used when variables are combined for deletion rather than deleting one variable at a time (iii) chopping happens when the search process permanently removes a highly insignificant brunch.

Phase 4: Search for Nested Terminal

Phase 4 deals with additional inspections on the terminal with the aim of finding variables that ought to be in the GUMS system. In order to find various terminal models, a minimal bunch is eliminated along the present path.

Phase 5: Selections of Final Model

In this phase, the algorithm finishes iterating when the new GUMS is the same as the previous GUMS. The model with the lowest information criterion values is chosen as the final model if there are numerous terminal candidate models.

2.2. Expectation-Maximization (EM) Algorithm

Dempster et al. [39] extensively described the EM algorithm method. Each algorithm iteration includes two steps: the E- or expectation step and the M- or maximisation phase. θ_0 is assumed to be an initial set of parameter estimations. The E-step entails calculating the conditional expectation of the log-likelihood for all of the data. This is where the expectation relevant to the distribution of the ‘missing' data, conditional on θ0 and on the observed data is gained. The M-step is the next step. The ‘expected log-likelihood generated in the E-step is maximised in relation to θ, yielding a new estimate θ_1. Continue by returning to the E-step, substituting θ_0 with θ_1, and cycling through the E- and M-steps until convergence is achieved.

Mclachlan and Krishnan [22] agreed that computing the ML estimator using the EM method is frequently made achievable by intentionally presenting it as an incomplete data problem, even if it does not appear to be an incomplete data problem at first. Given that this is an early attempt at using an EM algorithm to select the SURE model of time series missing data, this research was constructed with only one missing rate of 30%, in accordance with the rates used in previous simulation designs such as in Emura and Shiu [40]. The following EM algorithm was used accordingly:

Step 1: Start from initial values based on the means calculated from the available data.

Step 2: Perform the FGLS estimates of SURE model parameters for the full data set.

Step 3: Forecast missing values using the estimates found above.

Step 4: Replace forecasted values for missing values.

Step 5: Go to Step 2 and continue until convergence of parameter estimates.

3. Analysis

3.1. Simulation Analysis

The simulation study was designed to evaluate the overall performance of the suggested algorithm in searching for the true model specification based on experiments in [34], [41] and [42]. The first model, M1 was also denoted as 'empty model' since it was purely random errors, while M2 comprised the first lag of the dependent variable. In the meantime, M3 was an augmentation of M2 by inserting one independent variable (x_t) and its first lag (x_{t-1}). The three econometric models in Table 1 have been utilised as a testbed for true specification search.

Table 1. True Specification Models of SURE(EM)-Autometrics

Notation	Model
M1	$y_{1,t} = 61.229 + 6.709\varepsilon_{1,t}$ $y_{2,t} = 61.584 + 7.377\varepsilon_{2,t}$
M2	$y_{1,t} = 48.239 + 0.211y_{1,t-1} + 6.444\varepsilon_{1,t}$ $y_{2,t} = 49.271 + 0.199y_{2,t-1} + 7.091\varepsilon_{2,t}$
M3	$y_{1,t} = 71.492 + 0.060y_{1,t-1} - 0.414x_{1,4,t} - 0.034x_{1,4,t-1} + 4.210\varepsilon_{1,t}$ $y_{2,t} = 53.052 + 0.292y_{2,t-1} - 0.419x_{2,4,t} + 0.115x_{2,4,t-1} + 4.705\varepsilon_{2,t}$

During Phase 1 of the procedure, additional unnecessary variables were then added to these true models. As the data-generating process was known, the performances of *SURE(EM)-Autometrics* were evaluated by computing the percentages of the final models selected comparable to the true models. The goal was to achieve a high percentage of these outcomes. Table 2 summarises all the conditions that were put up for the simulation study:

Table 2. Conditions of Simulation Analysis

No.	Condition	Level
1	Sample size	$n = 550$
2	Correlation strength among equations	Weak, $\rho = 0.2$ Strong, $\rho = 0.9$
3	Initial GUMS	$k = 18$ total variables
4	Number of equations	$m = 2$
5	Main significance level	$\alpha = 1\%$ $\alpha = 5\%$
6	True specification model	M1 = no relevant variable M2 = one relevant variable M3 = three relevant variables

The simulation results are divided into four categories, with the criteria adopted from [36]. Table 3 lists the various categories. If all the equations in the model met the requirements, the results fall into the assigned category.

Table 3. Simulation Outcomes Categories

Category	Criteria	Description
1	TRUE = FINAL	The true specification is chosen.
2	TRUE ⊂ FINAL	The true specification is nested in the final specification.
3	TRUE ⊄ FINAL	An incorrect specification is chosen, the true specification is not nested in the final specification.
4	Nil	At least one equation failed to fall under the same category.

Category 1 refers to all equations in the model that contain variables precisely comparable to the true specification, which is what this simulation aimed for. The high number of Category 1 results showed strong algorithm performance. Category 2 applies to models consisting of all equations with more variables specified than with true specifications. To put it another way, true specifications are layered within final models. On the other hand, Category 3 classifies all equations comprising only some or no variables as in true specification. Furthermore, there are certain cases where different categories resulted for each equation in the model. For example, Equation 1 belongs to Category 1, whereas Equation 2 belongs to Category 2. As a result, the final category is specifically designed for this type of circumstance, Category 4.

3.2. Empirical Analysis

On top of the *SURE(EM)-Autometrics* procedures, this empirical study considered three other selections: *Autometrics-SURE, Autometrics-SURE(EM), and SURE-Autometrics*. These selections were categorised according to how the equations were selected individually or collectively, and the estimation method was utilised in the final models, as listed in Table 4. *Autometrics* is a single equation model algorithm that is built into the *PcGive* programme.

Given that the SURE model contains multiple equations, each one was estimated independently using OLS and selected several times. Nevertheless, the FGLS was used to estimate the final model in *Autometrics-SURE*. In the meantime, *Autometrics-SURE(EM)* is similar to Autometrics-SURE, but it uses the EM algorithm to estimate the final model. *SURE-Autometrics* and *SURE(EM)-Autometrics* are automated model selection algorithms that concentrate on the multiple equations model. In *SURE-Autometrics*, model selection was done concurrently using the FGLS estimation, whereas in *SURE(EM)-Autometrics, EM* estimation was integrated.

Table 4. Models Selection Procedures

No.	Model selection procedures	Final models estimations	Equation model
1	*Autometrics-SURE*	FGLS	Single
2	*Autometrics-SURE(EM)*	EM	Single
3	*SURE-Autometrics*	FGLS	Multiple
4	*SURE(EM)-Autometrics*	EM	Multiple

The dependent variable used in this study was the weekly data of the water quality index (WQI) of a river in Malaysia of 80 observations. Meanwhile, the independent variables were Dissolved Oxygen (DO) (% saturation), Dissolved Oxygen (DO) (mg/L), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Suspended Solids (SS), pH and Ammoniacal Nitrogen (NH_3N). The sub-indices in the analysis were created using these independent variables. Data sets were collected from two sampling stations, namely S7 and S8 which yielded a two-equation model.

4. Results and Discussion

4.1. Simulation Analysis

This simulation started with all WQI data available in hand. Two monitoring stations represent two equations in models. True models of M1, M2 and M3 were used as initial GUMS consisting of 18 independent variables at 1% and 5% levels of significance in a sample of 550 observations with strong (0.9) and weak (0.2) disturbance correlations. Table 5 presents percentages gained for each category by *SURE(EM)-Autometrics* of two equations with strong disturbance correlation.

The results reveal that *SURE(EM)-Autometrics* displayed a gradual increment from M1 to M3 at a 1% level. Category 1 still managed to provide considerably high percentages of more than 80% while these values continued to surge up to 90% in M2 and slightly more than 90% in M3. The highest percentage in Category 1 of M3 could be caused by the inclusion of the most correlated variable with the dependent variable, x_{i4t} and strong disturbance correlation. Nevertheless, only less than 10% of final models were grouped in each Category 2 and 4 but none in Category 3. The tight significance level of 1% had possibly successfully avoided any significant irrelevant variables to be in the final model.

Performance of *SURE(EM)-Autometrics* for 5% level of significance showed that percentages in Category 1 fell drastically compared to the same category for 1% level and thus triggered more final models to be classed in Category 2. This is followed by M2 with moderate accomplishment and finally, M1 with the least scores. All of the final models included true specifications; thus, no models were found under Category 3. Finally, there were still models which had different equations under different categories and therefore still contributed percentages under Category 4.

As for the results of final models for *SURE(EM)-Autometrics* with a weak correlation of disturbance ($\rho = 0.2$) at the 1% level, M3 still covered most models in Category 1 compared to the other two initial GUMS, while M2 was not far behind M3. The strength of disturbance correlation did not influence the model selection processes in this case, unlike the consequence of x_{i4t} inclusion in the model. In an empty model as in M1, the algorithm produced 12% of final models exactly as in true specifications.

On the other hand, at a 5% level of significance, none of the final models was found in Category 1 for M1, but close to 10% for M2 and M3 under weak correlation of disturbance. Overall, most of the true specifications were found to be nestled in the final models, Category 2. No model fell in Category 3, but mixed results were attained for Category 4. Consequently, a weak correlation had made the SURE model less efficient. The overall findings were found to be consistent with the results by Yusof [36] where the algorithm performed well when the number of equations and number of predictors in the true specification models was as minimal as possible.

4.2. Empirical Analysis

Stations S7 and S8 are represented by a system of two equations. The estimated models using different model selection procedures are displayed in Table 6 and Table 7.

Table 5. Percentages of Simulation Results for *SURE(EM)-Autometrics* with $m = 2$, k=18 and n=550

True model	Category															
	$\rho = 0.9$								$\rho = 0.2$							
	$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 1\%$				$\alpha = 5\%$			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
M1	86	7	-	7	14	70	-	16	12	68	-	20	-	89	-	11
M2	90	5	-	5	50	37	-	13	35	27	-	38	9	46	-	45
M3	93	5	-	2	53	27	-	20	56	10	-	34	10	40	-	50

Table 6. Estimated Models of WQI for S7

Models selection procedures	Constant	Δy_{it-1}	Δy_{it-2}	Δy_{it-3}	Δx_{i1t}	$\Delta x_{i1(t-1)}$	Δx_{i2t}	$\Delta x_{i2(t-1)}$	Δx_{i3t}	$\Delta x_{i3(t-1)}$
Autometrics-SURE	66.663***	-	-	-	0.245***	-	-	-	-0.674***	-
Autometrics-SURE(EM)	66.897***	-	-	-	0.245***	-	-	-	-0.670***	-
SURE-Autometrics	66.621***	-	-	-	0.242***	-	-	-	-0.683***	-
SURE(EM)-Autometrics	65.715***	0.088**	-0.064**	-	0.224***	-	-	-	-0.645***	-

Table 6. Estimated Models of WQI for S7 (cont.)

Models selection procedures	Δx_{i4t}	$\Delta x_{i4(t-1)}$	Δx_{i5t}	$\Delta x_{i5(t-1)}$	Δx_{i6t}	$\Delta x_{i6(t-1)}$	Δx_{i7t}	$\Delta x_{i7(t-1)}$	$\overline{R}^2$	Std. errors
Autometrics-SURE	-0.164***	-	-0.079***	-	-	-	-1.250***	-	0.943	1.545
Autometrics-SURE(EM)	-0.161***	-	-0.080***	-	-	-	-1.324***	-	0.942	1.564
SURE-Autometrics	-0.161***	-	-0.080***	0.007	-	-	-1.250***	-	0.945	1.512
SURE(EM)-Autometrics	-0.168***	-	-0.078***	0.011**	-	-	-1.468***	0.254*	0.943	1.507

***Significant at 1%, ** Significant at 5%, * Significant at 10%

Table 7. Estimated Models of WQI for S8

Models selection procedures	Constant	Δy_{it-1}	Δy_{it-2}	Δy_{it-3}	Δx_{i1t}	$\Delta x_{i1(t-1)}$	Δx_{i2t}	$\Delta x_{i2(t-1)}$	Δx_{i3t}	$\Delta x_{i3(t-1)}$
Autometrics-SURE	61.831***	-	-	-	0.293***	-	-	-	-0.724***	-
Autometrics-SURE(EM)	62.219***	-	-	-	0.291***	-	-	-	-0.714***	-
SURE-Autometrics	56.977***	0.037	0.038	-	0.287	-	-	-	-0.719	-
SURE(EM)-Autometrics	58.142***	0.071*	-	-	0.279***	-	-	-	-0.705***	-

Table 7. Estimated Models of WQI for S8 (cont.)

Models selection procedures	Δx_{i4t}	$\Delta x_{i4(t-1)}$	Δx_{i5t}	$\Delta x_{i5(t-1)}$	Δx_{i6t}	$\Delta x_{i6(t-1)}$	Δx_{i7t}	$\Delta x_{i7(t-1)}$	$\overline{R}^2$	Std. errors
Autometrics-SURE	-0.118***	-	-0.060***	-	-	-	-1.387***	0.232**	0.949	1.603
Autometrics-SURE(EM)	-0.120***	-	-0.058***				-1.459***	0.218**	0.947	1.624
SURE-Autometrics	-0.114***	-	-0.058***	-	-	-	-1.353***	0.301**	0.950	1.557
SURE(EM)-Autometrics	-0.119**	-	-0.061***	-	-	-	-1.490***	0.342*	0.948	1.603

***Significant at 1%, ** Significant at 5%, * Significant at 10%

Table 8. Forecasting Performances based on RMSE and GRMSE

Model selection procedure	One-Step		Two-Step		Three-Step		One-Step		Two-Step		Three-Step	
	RMSE	Rank	RMSE	Rank	RMSE	Rank	GRMSE	Rank	GRMSE	Rank	GRMSE	Rank
Autometrics-SURE	1.6980	3	1.8551	3	2.1069	3	1.2874	3	1.4083	3	1.8082	3
Autometrics-SURE (EM)	1.7066	4	1.8616	4	2.1284	4	1.2673	2	1.3551	2	1.8307	4
SURE-Autometrics	1.6902	2	1.7318	1	1.8903	1	1.4108	4	1.4391	4	1.6790	2
SURE(EM)- Autometrics	1.5704	1	1.7612	2	1.9892	2	1.1005	1	1.2783	1	1.6522	1

Following the experiments by Suzilah [34] and Yusof [36], two types of error measurements were used for this study: root mean square error (RMSE) and geometric root mean square error (GRMSE). The presence of low values for these indicators indicates good forecasting performance. When it comes to evaluating the effectiveness of forecasting models, the RMSE is one of the most used measures among practitioners. When forecasts are performed across successive periods with the same forecast horizon and the cost function is quadratic, the RMSE is an appropriate measure of accuracy. Between now and then, Fildes [13] proposed that the GRMSE be used in situations in which there are infrequent outliers in the data (and errors), as well as when dealing with a significantly large error term due to an exceptionally terrible forecast.

After computing the two error measures for each procedure up to three steps ahead, the medians for all equations in the SURE model were determined and ranked from 1 (the best) to 4 (the worst). This provided an assessment of the performance of each selection procedure. The results for one, two, and three-step-ahead forecasts using a two-equation model are shown in Table 8.

SURE(EM)-Autometrics has consistently performed at the top in this empirical analysis ranking first or second for all RMSEs and GRMSEs. This was followed by *SURE-Autometrics,* which rated first in two circumstances. Both *Autometrics-SURE(EM)* and *Autometrics-SURE* received worse rankings than their competitors. These findings revealed that system selections for multiple equations systems are more competent than individual selections. These results support the earlier findings by Ismail, Yusof and Muda [43] and Yusof [36] where multiple models selection algorithms performed well in one and two step-ahead forecasts. In addition, the EM algorithm proved to be more superior to FGLS in selecting the equations simultaneously.

5. Conclusions

SURE(EM)-Autometrics has successfully demonstrated excellent performance in selecting a multiple equations model, in particular two-equation SURE models. Because the model is made up of many equations, the performance of the algorithm is reflected by the percentages calculated when all equations are comparable to the true model in the simulation analysis. The high percentages in searching the final models similar to true specifications were mainly found when the correlation of disturbance was strong at a 1% level of significance.

On top of the simulation analysis, an empirical analysis was also conducted, and it was discovered that *SURE(EM)-Autometrics* proved to be the most successful selection procedure with the first or second ranking relative to other procedures. This suggested model selection has shown that selection of the system is considerably more efficient than individual selections.

The 'best' model may be selected concurrently from several equations of the SURE model by using the EM algorithm estimation in the algorithm. This indicates that in this study, the EM algorithm estimation method has been verified to be more trustworthy rather than FGLS. The findings of this study offer a valuable understanding of the EM algorithm estimation method in multiple equations model selection, notably in the SURE model. Consequently, this new estimation method also gives meaningful outcomes in identifying the 'best' parsimonious model from a general model and finally able to improve the model's forecast ability.

Acknowledgement

The authors would like to express utmost gratitude and appreciation to Universiti Utara Malaysia for the financial support (Grant S/O Code: 14925). We also appreciate the helpful recommendations provided by the anonymous reviewers.

REFERENCES

[1] D. Laredo, S. F. Ma, G. Leylaz, O. Schütze, and J. Q. Sun, "Automatic model selection for fully connected neural networks," *International Journal of Dynamics and Control*, vol. 8, no. 4. pp. 1063–1079, 2020, doi: 10.1007/s40435-020-00708-w.

[2] H. Chai, J. F. Ton, M. A. Osborne, and R. Garnett, "Automated model selection with Bayesian quadrature," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 1555–1564, 2019.

[3] W. H. Greene, *Econometric analysis*, 8th ed. Harlow: Pearson Education Limited, 2012.

[4] R. Bakirov, D. Fay, and B. Gabrys, "Automated adaptation strategies for stream learning," *Mach. Learn.*, vol. 110, no. 6, pp. 1429–1462, 2021, doi: 10.1007/s10994-021-05992-x.

[5] C. Epprecht, D. Guégan, Á. Veiga, and J. Correa da Rosa, "Variable selection and forecasting via automated methods for linear models: LASSO/adaLASSO and Autometrics," *Commun. Stat. Simul. Comput.*, vol. 50, no. 1, pp. 103–122, 2021, doi: 10.1080/03610918.2018.1554104.

[6] R. S. Pindyck and D. L. Rubinfeld, *Econometric models and economic forecasts*, 4th ed. Boston, Massachusetts: Irwin/McGraw-Hill, 1998.

[7] A. Zellner, "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," *J. Am. Stat. Assoc.*, vol. 57, no. 298, pp. 348–368, 1962.

[8] A. Khan, S. Baloch, K. Arif, and J. Alvi, "Does Capital Asset Pricing hold in Pakistan Stock Exchange? An Application of Seemingly Unrelated Regression," *Int. J. Psychosoc. Rehabil.*, vol. 24, no. 09, 2020.

[9] L. Yang, P. Hui, R. Yasmeen, S. Ullah, and M. Hafeez, "Energy consumption and financial development indicators

nexuses in Asian economies: a dynamic seemingly unrelated regression approach," *Environ. Sci. Pollut. Res.*, vol. 27, no. 14, pp. 16472–16483, 2020.

[10] X. Pan, S. Guo, C. Han, M. Wang, J. Song, and X. Liao, "Influence of FDI quality on energy efficiency in China based on seemingly unrelated regression method," *Energy*, vol. 192, p. 116463, 2020.

[11] M. Sheng and B. Sharp, "Aggregate road passenger travel demand in New Zealand: A seemingly unrelated regression approach," *Transp. Res. part A policy Pract.*, vol. 124, pp. 55–68, 2019.

[12] X. Xu, Ž. Šarić, F. Zhu, and D. Babić, "Accident severity levels and traffic signs interactions in state roads: A seemingly unrelated regression model in unbalanced panel data approach," *Accid. Anal. Prev.*, vol. 120, pp. 122–129, 2018.

[13] R. Fildes, Y. Wei, and S. Ismail, "Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures," *Int. J. Forecast.*, vol. 27, no. 3, pp. 902–922, Jul. 2011.

[14] H. Rohani, M. Ghorbani, and M. Kohansal, "Analysis of the effective factors on dimensions of sustainable agricultural development in Khorasan Razavi province, using seemingly unrelated regression equations," *Iran. J. Agric. Econ. Dev. Res.*, vol. 52, no. 1, pp. 33–52, 2021.

[15] W. T. Lin and J. Shi, "Chief executive officer compensation, firm performance, and strategic coopetition: A seemingly unrelated regression approach," *Manag. Decis. Econ.*, vol. 41, no. 1, pp. 130–144, 2020.

[16] C. Chen *et al.*, "Altered metabolite levels and correlations in patients with colorectal cancer and polyps detected using seemingly unrelated regression analysis," *Metabolomics*, vol. 13, no. 11, pp. 1–10, 2017.

[17] H. pin Lai, "Maximum simulated likelihood estimation of the seemingly unrelated stochastic frontier regressions," *Empir. Econ.*, vol. 60, no. 6, pp. 2943–2968, 2021, doi: 10.1007/s00181-020-01962-9.

[18] J. Liu and Q. Xia, "Some finite sample results for a system of seemingly unrelated regression equations," *Commun. Stat. Methods*, pp. 1–16, 2020.

[19] R. B. Afolayan, A. W. Banjoko, M. K. Garba, and W. B. Yahya, "On Seemingly Unrelated Regression and Single Equation Estimators Under Heteroscedastic Error and Non-Gaussian Responses," *FUOYE J. Eng. Technol.*, vol. 5, no. 2, 2020, doi: 10.46792/fuoyejet.v5i2.469.

[20] V. K. Srivastava and D. E. A. Giles, *Seemingly unrelated regression equations models: Estimation and inference*. CRC press, 2020.

[21] L. Zhao and X. Xu, "Generalized canonical correlation variables improved estimation in high dimensional seemingly unrelated regression models," *Stat. Probab. Lett.*, vol. 126, pp. 119–126, 2017.

[22] G. J. Mclachlan and T. Krishnan, *The EM algorithm and extensions*, 2nd ed. New Jersey: John Wiley & Sons, Inc., 2008.

[23] S. K. Ng, T. Krishnan, and G. J. McLachlan, "The EM algorithm," in *Handbook of computational statistics*, 2nd ed., J. E. Gentle, W. K. Hardle, and Y. Mori, Eds. Berlin, Heidelberg: Springer, 2012, pp. 139–172.

[24] K. Sohn, "An expectation-maximization algorithm to estimate the integrated choice and latent variable model," *Transp. Sci.*, vol. 51, no. 3, Aug. 2016.

[25] G. van Ryzin and G. Vulcano, "An expectation-maximization method to estimate a rank-based choice model of demand," *Oper. Res.*, vol. 65, no. 2, pp. 396–407, 2017, doi: 10.1287/opre.2016.1559.

[26] W. Bailer, M. Winter, J. Ebert, J. Flavio, and K. Plimon, "Expectation-Maximization for Scheduling Problems in Satellite Communication," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3459–3466.

[27] A. Subudhi, M. Dash, and S. Sabut, "Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 277–289, 2020.

[28] B.-R. Sah *et al.*, "Clinical evaluation of a block sequential regularized expectation maximization reconstruction algorithm in 18F-FDG PET/CT studies," *Nucl. Med. Commun.*, vol. 38, no. 1, pp. 57–66, 2017.

[29] L. Malan, C. M. Smuts, J. Baumgartner, and C. Ricci, "Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns," *Nutr. Res.*, vol. 75, pp. 67–76, 2020.

[30] L. Shenhav *et al.*, "FEAST: fast expectation-maximization for microbial source tracking," *Nat. Methods*, vol. 16, no. 7, pp. 627–632, 2019.

[31] R. B. Babu, G. Snehal, and P. A. S. Kiran, "Detection of Crimes Using Unsupervised Learning Techniques," *Aptikom J. Comput. Sci. Inf. Technol.*, vol. 2, no. 1, pp. 8–11, 2017.

[32] K. Hoover and S. Perez, "Data mining reconsidered: encompassing and the general ‐ to ‐ specific approach to specification search," *Econom. J.*, vol. 2, pp. 167–191, 1999.

[33] H. Krolzig and D. F. Hendry, "Computer automation of general-to-specific model selection procedures," *J. Econ. Dyn. Control*, vol. 25, pp. 831–866, 2001.

[34] S. Ismail, "Algorithmic approaches to multiple time series forecasting," Unpublished doctoral dissertation, University of Lancaster, Lancaster, 2005.

[35] J. A. Doornik and D. F. Hendry, *Empirical Econometric Modelling using PcGive 12: Volume 1*. London: Timberlake Consultants Ltd, 2007.

[36] N. Yusof, "SURE-Autometrics algorithm for model selection in multiple equations," Unpublished doctoral dissertation, Universiti Utara Malaysia, Sintok, 2016.

[37] N. A. Kamarudin and S. Ismail, "Model Selection Approaches of Water Quality Index Data," *Glob. J. Pure Appl. Math.*, vol. 12, no. 2, pp. 1821–1829, 2016.

[38] N. A. Kamarudin and S. Ismail, "Manual and Automated Model Selection Procedures for Seemingly Unrelated Regression Equations with Different Estimation Methods," *Far East J. Math. Sci.*, vol. 101, no. 8, pp. 1655–1670, 2017.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via EM algorithm," *J. R. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[40] T. Emura and S.-K. Shiu, "Estimation and model selection for left-truncated and right-censored lifetime data with application to electric power transformers analysis," 57528, 2014. Accessed: May 17, 2017. [Online]. Available: http://mpra.ub.uni-muenchen.de/57528/.

[41] N. Yusof, "A model selection algorithm for multiple equations within the general-to-specific approach," Northern University of Malaysia, 2016.

[42] J. Doornik, "Autometrics," in *The methodology and practice of econometrics*, J. Castle and N. Shephard, Eds. Oxford: Oxford University Press, 2009.

[43] S. Ismail, N. Yusof, and T.-Z. T-Muda, "Algorithmic approaches in model selection of the air passengers flows data," *Proc. 5th Int. Conf. Comput. Informatics*, no. 218, pp. 32–37, 2015.