

Outliers detection in state-space model using indicator saturation approach

Farid Zamani Che Rose¹, Mohd Tahir Ismail², Mohd Hanafi Tumin³

^{1,2}School of Mathematical Sciences, Universiti Sains Malaysia (USM), Malaysia

³School of Mathematics and Basic Sciences, Faculty of Science and Technology, Quest International University, Malaysia

Article Info

Article history:

Received Sep 10, 2020

Revised May 6, 2021

Accepted May 19, 2021

Keywords:

Additive outlier

General-to-specific

Indicator saturation

Local level

Monte Carlo simulation

State space

ABSTRACT

Structural changes that occur due to outliers may reduce the accuracy of an estimated time series model, shifting the mean distribution and causing forecast failure. This study used general-to-specific approach to detect outliers via indicator saturation approach in the local level model framework. Focusing on impulse indicator saturation, performance recorded by the suggested approach was evaluated using Monte Carlo simulations. To tackle the issue of higher number of regressors compared to the number of observations, this research utilized the split-half approach algorithm. We found that the impulse indicator saturation performance relies heavily on the size of outlier, location of outlier and number of splits in the series examined. Detection of outliers using sequential and non-sequential algorithms is the most crucial issue in this study. The sequential searching algorithm was able to outperform the non-sequential searching algorithm in eliminating the non-significant indicators based on potency and gauge. The outliers captured using impulse indicator saturation in financial times stock exchange (FTSE) United States of America (USA) shariah index correspond to the financial crisis in 2008-2009.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Farid Zamani Bin Che Rose

School of Mathematical Sciences

Universiti Sains Malaysia

11800 USM, Penang, Malaysia

Email: farid6908@gmail.com

1. INTRODUCTION

Observation data used in empirical models are assumed to originate from stationary process. In reality, most real-time series data are not stationary and their means and variances vary with the data. The changes may take various forms, magnitudes and numbers. Any structural change in time series may affect parameter estimation, especially estimation of economic and financial indicators. An outlier in time series is also known as an unusual large value of irregular disturbance at a specific time. The existence of outlying observations and structural changes always raise a big question on the accuracy and efficiency of the estimated model. Failing to model the structural changes may lead to misspecification of the empirical model and forecast accuracy [1]. These are the common problems when dealing with structural changes in time series data. Hence, this study proposed to solve the problem of distortion in the parameter due to presence of outliers using the indicator saturation approach in general-to-specific (GETS) modelling.

General-to-specific modelling is widely used to model economics and finance data. The idea of model selection originated with [2] by revisiting [3] work on data mining experiment. Hendry and Krolzig [4] improved the automated multipath GETS modelling using MATLAB code simulations by increasing the

computational power 1000-fold. The results attracted [4] to make further improvements of the selection algorithm in Autometrics. Then, they made the algorithm works in computer program viz PcGets. Another improvement made with the third implementation of GETS in Autometrics was by [5] that works using the same approach as in [2] and [4]. Indicator saturation method was embodied in Autometrics by [6] to capture outliers and structural breaks in the saturated regression model. In addition, recent study by [7] employed indicator saturation approach in Autometrics to model the location shifts and structural breaks in non-stationary big data.

General processes used to identify both outliers and structural breaks using state-space models are further discussed in [8]. It is important to capture the outliers in a model immediately to avoid misspecification of parameter estimation and improve forecast accuracy. Specifically, the indicator saturation approach is used to detect if the outlier is close to forecast origin. Hendry [9] started his work on impulse indicator saturation (IIS) to test the unknown amount of outlier, sign and magnitude occurring at unknown locations in time series data. The idea of IIS was first introduced to model US food demand in 1930s and 1940s [10]. To detect outliers, the method created a pulse dummy variable for each result obtained by the study. This is an active and burgeoning area of research. Numerous studies on the indicator saturation approach with economic applications can be found in [11]-[16]. The application of indicator saturation is not only limited to economic research, but also applicable to other areas of research such as climate change [17].

At present, research about the performance of indicator saturation especially from structural time series viewpoint is done by [14] using basic structural model (BSM) only. The work presented in this paper is motivated by the fact that outliers can be identified and modelled via the indicator saturation approach. Therefore, we aim to apply impulse indicator approach in the context of the local level model (LLM). The state-space framework chosen has an advantage over other time series models since the state-space method is able to handle non-stationary data. Typically, econometricians handle non-stationary data using differencing approach that reduces the integration order as in [18]. However, this approach is not able to model any long-run equilibria such as relationships in the data. Therefore, we propose to integrate the IIS in state-space model framework since it can handle non-stationary data. Apart from that, the suggested framework is able to handle missing or incomplete data, time-varying regression coefficients and multivariate extensions. Hence, we decide to take a new look on the local level model, which consists of a random disturbance around an underlying level which fluctuates without any specific direction. LLM is known as the simplest model in the state-space family and it may be described in state-space forms. We refer to [19] for detailed analysis treatment of state-space methods. The main benefit of LLM is its fully deterministic state-space models can also be addressed as classical linear regression models. The primary advantage of state-space methods is it fits better to data than classic linear regressions model [20].

Prior works in [21] and [22] do not employ IS in Gaussian state-space model. Meanwhile, [14] applied impulse and step indicators to detect outliers and level shift in the basic structural model. Our study attempts to contribute to the literature by examining the performance of IIS in the context of the local level model by assessing the performance of IIS in LLM through Monte Carlo simulations replicated at $M = 1000$ times. Further, we apply the impulse indicator saturation to identify outliers in shariah-compliant stock price series, specifically financial times stock exchange (FTSE) hijrah shariah and FTSE all-world shariah. In addition, no study has made the effort to understand the effectiveness of indicator saturation integration in state-space model using gets package in R programming language. The gets package developed by [23] was able to automatically determine the required model based on general-to-specific modelling and indicator saturation. Hence, we aim to provide details of simulation analysis using gets package in the context of state-space model.

The rest of this paper is arranged as; section 2 elaborates the structure of the local level model for outlier detection and introduces the concepts of indicator saturation approach and how the approach can be applied into the local level model framework. Section 3 describes the simulation setting for the Monte Carlo experiment and shows the performance of Monte Carlo simulations on the detection power of IIS. Then, IIS is applied to real stock price data for detecting outliers in section 4. Finally, section 5 wraps up the whole paper.

2. RESEARCH METHOD

A simple example of the state-space model, the local level model's level component varies over time. In the classical regression model, this component can be viewed as the intercept, however, in the state-space model, the level component may differ according to specific time points. The local level model can be formulated as (1).

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2) \quad (1)$$

$$\mu_{t+1} = \mu_t + \psi_t, \quad \psi_t \sim NID(0, \sigma_\psi^2) \quad (2)$$

For $t = 1, \dots, T$ where μ_t is the unobserved level at time t , ε_t is the observation disturbance or irregular component at time t and ψ_t is the level disturbance at time t . Supposed to be serially and mutually independent, the observation and level disturbances are normally distributed with zero mean and variances ε_t and ψ_t , respectively. We denote (1) as the observation equation, while (2) is known as the state equation. The transition equation shows the fundamental values based on a random walk. ε_t is the noise from the fundamental that is expected to be unrelated from any change experienced by μ_t . The signal-to-noise ratio of variances, $q = \sigma_\varepsilon^2 / \sigma_\psi^2$, measures the potency of fundamental value signal versus random deviation. Due to this, the local level model can also be defined as random walk plus noise model [20].

Indicator saturation methods introduced by [9] is able to identify the presence of several outliers happening at various locations, and unknown magnitudes. Moreover, a slight alteration to the model may result to identification of various deterministic structural change formats [1]. In this section, we discuss IIS significantly. IIS is able to generate a full indicator variables set, e.g., for T number of observations, IIS will create T indicators in the set of candidate variables. However, this method would not incorporate all indicators as regressors to avoid perfect fit in general-to-specific modelling process. If all the indicators are included, the model will have more regressors as compared to number of observations, T . This will lead to degree of freedom deficiency [24]. The impulse indicators defined as $\{1_{\{h=t\}}\}$ where $\{1_{\{h=t\}}\}$ equals to unity when $h=t$ and equals to zero otherwise for $h = 1, \dots, T$.

To illustrate the procedure, [25] and [26] used the split-half approach to integrate IIS into a generated model $y_t = \mu_t + \varepsilon_t$ for $t = 1, \dots, T$ where ε_t is normally and independently distributed with mean zero and variance σ_ε^2 . Specifically, in the first half of the sample, $T/2$ impulse indicators are included in the model which results to:

$$y_t = \mu_t + \sum_{h=1}^{(T/2)} \beta_h I_{h,t}(h) + \varepsilon_t, \quad \text{for } t = 1, \dots, T/2 \quad (3)$$

where $I_{h,t}(h)$ is impulse indicator vector and ε_t is the error term which $\varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$. Meanwhile the other half of the sample result to:

$$y_t = \mu_t + \sum_{h=1}^{(T-T/2)} \beta_h I_{h,t}(h) + \varepsilon_t, \quad \text{for } t = T/2, \dots, T \quad (4)$$

The selected indicators at the chosen significance value, α are determined using the t-statistics value in the first half of the sample. Location of significant indicators will be recorded. Then, $T/2$ impulse indicator is added in the second half of the sample, $T - T/2$, after which the selection procedure is performed repetitively to obtain significant indicators under the null hypothesis of no outlier. Finally, a terminal model is obtained from joining together two sets of significant dummies. The selection of significant indicators discussed above adheres to sequential selection where the non-significant indicator is dropped one at a time at chosen significant level. Alternatively, there is another selection algorithm called non-sequential selection. It works by dropping all non-significant indicators simultaneously at chosen α in every partition. The retained indicators are significant indicators. This approach is always feasible if the number of regressors, N equals to the number of observations, T . As mentioned, if the total number of regressors is higher than the number of observations, $N > T$, we consider a cross block algorithm proposed by [27]. This algorithm segregates all indicators into m blocks, and the selection algorithm is repeated. Hence, this gives a total of $m - (m - 1)/2$ runs of the selection procedure. This approach was employed by [14] on the BSM context. Inspired by their work, we integrated the IS approach in the local level model as in (5). We have defined m as total blocks into which indicators are split, and T is multiples of m . Supposed that the size of the blocks are similar, hence, the observation equation in (1) is extended to:

$$y_t = \mu_t + \sum_{h=(T/m)(h+1)+1}^{(T/m)j} \beta_h I_{h,t}(h) + \varepsilon_t, \quad \text{for } t = 1, \dots, T \quad (5)$$

where $I_{h,t}(h)$ denotes IIS. Next, (4) is converted into state-space form together with (2). Hence, the data generating process (DGP) in (5) can be represented in matrix notation as:

$$\mathbf{y} = \mathbf{I}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (6)$$

where \mathbf{y} and $\boldsymbol{\varepsilon}$ are vectors of $T \times 1$. The term \mathbf{I} is the IIS in matrix form with dimension $T \times T$ and $\boldsymbol{\beta}$ is the matrix coefficient with dimension $T \times 1$.

Representation of IIS matrix for univariate time series as in [17] can be shown as:

$$\begin{aligned}
 \tau_{1,t} &= (\tau(1), 0, 0, \dots, 0)^T \\
 \tau_{2,t} &= (0, \tau(1), 0, \dots, 0)^T \\
 \tau_{3,t} &= (0, 0, \tau(1), \dots, 0)^T \\
 &\vdots \\
 &\vdots \\
 \tau_{T,t} &= (0, 0, 0, \dots, \tau(1))^T
 \end{aligned}
 \tag{7}$$

where $\tau_{h,t} = \tau(t)$ when $h = t$ and zero for $h \neq t$. Assume that there is an outlier in the first block, b_1 with unknown location, L_1 and magnitude of unknown outlier denoted as λ . The DGP in matrix notation will be as follows:

$$\mathbf{y} = \mathbf{I}_{b_1} \boldsymbol{\beta}_{b_1} + \mathbf{v}
 \tag{8}$$

where $\mathbf{I}_{b_1} = (\tau_1, \tau_1, \dots, \tau_{T/2})$ is denoted as IIS matrix and $\boldsymbol{\beta}_{b_1} = (\beta_1, \beta_1, \dots, \beta_{T/2})^T$ is the coefficient vector for the IIS. Thus, the estimated coefficient for IIS vector is obtained through ordinary least square (OLS) estimation formulated as:

$$\begin{aligned}
 \boldsymbol{\beta}_{b_1} &= \lambda (\mathbf{I}_{b_1}^T \mathbf{I}_{b_1})^{-1} \mathbf{I}_{b_1}^T \boldsymbol{\tau}_{L_1} + (\mathbf{I}_{b_1}^T \mathbf{I}_{b_1})^{-1} \mathbf{I}_{b_1}^T \boldsymbol{\varepsilon} \\
 &= \lambda \mathbf{b} + (\mathbf{I}_{b_1}^T \mathbf{I}_{b_1})^{-1} \mathbf{I}_{b_1}^T \boldsymbol{\varepsilon}
 \end{aligned}
 \tag{9}$$

where \mathbf{b} is the IIS vector of size $T/2 \times 1$ which consists of value one at the location of outlier, $t = L_1$ and zero elsewhere. The vector \mathbf{b} is obtained when solving the expression $(\mathbf{I}_{b_1}^T \mathbf{I}_{b_1})^{-1} \mathbf{I}_{b_1}^T \boldsymbol{\tau}_{L_1}$ showing the element in \mathbf{b} at L_1 -th reflects the magnitude of outlier, λ . The expected value and variance of estimator $\hat{\boldsymbol{\beta}}_{b_1}$ are $\lambda \mathbf{b}$ and $\sigma_\varepsilon^2 (\mathbf{I}_{b_1}^T \mathbf{I}_{b_1})^{-1}$ are respectively obtained as shown in [17]. Then, the distribution of the OLS estimator in the first block is:

$$\hat{\boldsymbol{\beta}} - \lambda \mathbf{b} \sim N(\mathbf{0}, \sigma_\varepsilon^2 (\mathbf{I}_{b_1}^T \mathbf{I}_{b_1})^{-1})
 \tag{10}$$

Therefore, the significance testing in GETS modelling can be employed with standard t-test statistics as shown in [16]. The significance indicator of IIS will retain in the model if the absolute value of t-statistics is greater or equal to the critical value of the chosen significance level, $|t_{\hat{\boldsymbol{\beta}}}| \geq c_\alpha$. On the other hand, the impulse indicator will be eliminated if $|t_{\hat{\boldsymbol{\beta}}}| < c_\alpha$ or the t-statistics is close to zero. Then, the IIS procedure in GETS modelling described in this section is repeated for the second block, b_2 and final terminal model.

3. RESULTS AND DISCUSSION

3.1. Monte Carlo simulations settings

Performance of the indicator saturation approach is measured using Monte Carlo experiments. A time series is generated based on the DGP in (5) with initial values of components $\sigma_\psi^2 = 0.0563$ and $\sigma_\varepsilon^2 = 1$. The generated series were then added with additive outliers (AO). Firstly, we come out with a benchmark simulation setting for the DGP and outlier detection procedure. Then, we consider various alternative settings to investigate the robustness of the procedure. Every experiment involves 1000 replications. The following are specifications for simulation settings for a reference DGP:

- Sample size $T=240$ observations, reflecting 20 years of monthly data.
- Locations of the AO are as: A single AO was positioned right in the middle of the sample, while double AO were predetermined at the [0.25, 0.75] as a share of length T .
- Target size or significance level, $\alpha=0.001, 0.01$ and 0.025 [13] defined α as the statistical tolerance of the procedure to control the risk of inadvertently retain any irrelevant indicator. For example, a target of 0.01 for IIS indicates that on average, for every 100 observations, we accept a maximum of single impulse dummy that is not included in the data generating process.
- Size of an outlier is given as $s\sigma$ where s is a positive integer and σ is the prediction error standard deviation (PESD) of the series. As a reference to [14], we set 7σ as benchmark value that determines the size of an outlier. However, the size of outlier varies from $3\sigma, 5\sigma, 9\sigma$ and 12σ .

- There are two searching algorithms applied in the simulations which are denoted as non-sequential and sequential algorithms in selecting significant indicators. The former leads to 1-cut model selection, while the latter leads to multi-path model selection [10]. An additional number of paths can be added by increasing the number of m blocks of regressors as suggested by [17] to reduce the variance estimator and increase the detection power of IIS.
- With regard to outlier detection procedure using IIS, the indicator variables are divided into two, four and six blocks.
- The locations of AO also vary based on the share of the sample.

We decided to determine the appropriate size of AO in our Monte Carlo experiments using PESD since we deal with multiple sources of disturbances in the structural time series. As mentioned, we denote σ as PESD of the steady-state innovation. It can be formulated as:

$$v_t = y_t - E(y_t|Y_{t-1}) = \mu_t - E(y_t|Y_{t-1}) + \varepsilon_t \quad (11)$$

where $Y_t = \{y_{t-1}, y_{t-2}, \dots, y_1\}$. Moreover, this approach is also consistent with [28] and [29]. Overall, we measured the robustness of model based on few aspects which are number of observations, T, number of AO added, values of target size, magnitude of AO, number of blocks estimation and locations of AO in the series.

3.2. Evaluating the performance of Monte Carlo experiment

The main goal of the Monte Carlo experiment is to gauge the efficiency of IIS in the local level model. Therefore, we apply the concepts of potency and gauge to assess the performance of the outlier detection procedure. Potency can be defined as the proportion of relevant indicator that remained in the final model, while gauge is the proportion of irrelevant indicator that remained in the final model. Both concepts are computed based on the retention rate denoted by:

$$\tilde{r} = \frac{1}{M} \sum_{i=1}^M 1[\hat{\beta}_{ij} \neq 0], \quad j = 1, \dots, T \quad (12)$$

$$potency = \frac{1}{n} \sum_j \hat{r}_j, \quad j \in R_n \quad (13)$$

$$gauge = \frac{1}{T-n} \sum_j \hat{r}_j, \quad j \in R_{T-n} \quad (14)$$

where M denotes the number of replication and n is the number of true outliers in a particular time series of length T. Hence, let R_n and R_{T-n} be sets of time indices for relevant and irrelevant indicators in the model respectively. Meanwhile, $\hat{\beta}_{ij}$ denotes estimated coefficient in the impulse indicator and if $I_{h,t}(h)$ is selected, then the variable $1[\hat{\beta}_{ij} \neq 0]$ will take value one, where one means true and zero otherwise. We followed the rule of thumb suggested by [30] to determine the value of target size $\alpha = \min[0.05, 1/T]$ in stock return series. This will ensure the low gauge value below 5% of the sample, T or only one irrelevant indicator variable is maintained in the final model. On the other hand, the concept of potency and gauge used in this study can be illustrated as a confusion matrix as in [14]. However, the following confusion matrix only summarizes the result of one Monte Carlo experiment shown in Table 1.

Table 1. Summary result of one Monte Carlo experiment

Actual	Predicted		Total
	No outlier	Outlier	
No outlier	W	X	M(T-n)
Outlier	Y	Z	Mn
Total	W+Y	X+Z	MT

W and Z are known as true positive and true negative, denote numbers of correct decisions made. On the other hand, X and Y, also known as a false positive and false negative, denote false decisions when there is no outlier or one outlier, respectively. Hence, the potency is defined as the ratio of Z/Mn. Meanwhile, the gauge is given by the ratio of X/[M(T-n)].

The potency and gauge values were tabulated in Tables 2 and 3 as a result of a non-sequential searching algorithm adopted. It is evident that a variety of factors significantly contribute to retention rate, \tilde{r} when IIS approach is implemented. AO size plays a dominant role in retention rate. As examined closely, changes in potency value about 50% is found to be more noticeable when the size of AO increased from 3σ to 5σ . Meanwhile, the difference slowly reduces when the target size increases. However, as expected, the

potency reaches over 90% as the size of AO increases more than 7, although with a small inaccuracy. On the other hand, the gauge values clustered around the chosen α . This proves that the significance level is able to control the risk of irrelevant indicators inadvertently retained in the final terminal model.

Table 2. Potency and gauge values when contaminated with single AO with different significant values using non-sequential selection

		2 blocks estimation					4 blocks estimation					6 blocks estimation				
T=240	α	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ
Potency	0.001	18.7	71.9	94.2	98.8	100	18.7	63.5	94.2	93.5	99.8	13.7	71.9	93.6	98.2	99.9
	0.01	45.6	88.6	98.5	99.9	100	43.6	73.7	98.5	100	100	46.6	88.2	97.7	99.6	100
	0.025	58.1	78.7	96.1	100	100	56.6	82.7	99.2	99.9	100	56.4	92.8	99.1	99.6	100
Gauge	0.001	0.13	0.35	0.44	0.45	0.46	0.18	0.25	0.45	0.11	0.04	0.33	0.49	0.07	0.05	0.01
	0.01	1.13	1.24	1.18	1.11	1.02	1.03	0.99	0.98	0.14	0.58	1.06	1.07	0.64	0.35	0.33
	0.025	0.27	0.46	0.52	2.88	2.58	3.09	2.79	2.33	1.04	1.39	2.79	2.57	1.99	1.97	1.02

Table 3. Potency and gauge values when contaminated with double AO with different significant values using non-sequential selection

		2 blocks estimation					4 blocks estimation					6 blocks estimation				
T=240	α	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ
Potency	0.001	13.3	70.8	93.7	97.3	99.9	21.2	70.8	93.7	98.5	99.9	21.2	70.7	93.7	99.0	99.9
	0.01	46.3	86.8	98.1	99.3	100	45.7	86.5	98.1	98.8	100	45.3	86.4	98.1	99.8	100
	0.025	61.1	92.5	99.0	99.8	100	59.7	92.3	99.0	99.9	100	58.7	92.1	99.0	99.9	100
Gauge	0.001	0.08	0.04	0.02	0.01	0.00	0.05	0.08	0.04	0.03	0.04	0.21	0.12	0.06	0.02	0.09
	0.01	0.84	0.69	0.51	0.49	0.22	0.77	0.58	0.38	0.42	0.32	0.78	0.53	0.29	0.15	0.33
	0.025	3.16	2.54	1.85	1.50	1.94	2.64	2.02	1.36	1.17	1.09	2.31	1.69	1.11	0.67	0.74

IIS also performed well in sequential selection, as shown in Tables 4 and 5 with the potency almost reaches 100% for benchmark settings. When the size of AO increased from 3 to 5, the method is able to improve its instant identification rate circa 50%. This shows that the IIS efficiency is highly related to the magnitude of AO, as seen in [14], [17], [23]. Aside from that, total number of blocks, m is also a vital criterion that affects the performance of IIS to detect outlier even though there is a sequential selection algorithm in the IS approach. Thus, we decided to simulate using two, four and six blocks for both series. The number of blocks selected is based on [14] to minimize the risk of missing any essential structural change if there are too many blocks used. We also discovered that one limitation of IIS is any additional blocks, m more than ten will not lead to the detection of outliers in the sample observation.

Table 4. Potency and gauge values when contaminated with single AO with different significant values using sequential selection

		2 blocks estimation					4 blocks estimation					6 blocks estimation				
T=240	α	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ
Potency	0.001	19.7	75.6	97.6	99.8	100	21.1	73.9	97.6	99.8	100	23.4	77.8	98.0	99.6	100
	0.01	50.1	92.5	99.7	99.9	100	47.0	92.4	99.7	100	100	48.0	92.7	99.5	100	100
	0.025	64.2	96.7	100	100	100	59.5	95.5	99.9	100	100	60.9	96.4	99.9	100	100
Gauge	0.001	0.05	0.10	0.03	0.04	0.02	0.12	0.08	0.02	0.01	0.01	0.09	0.09	0.04	0.01	0.01
	0.01	0.87	0.73	0.70	0.59	0.56	0.78	0.61	0.33	0.26	0.16	0.72	0.59	0.32	0.20	0.12
	0.025	3.12	2.63	2.28	2.02	1.91	2.56	1.91	1.34	0.88	0.62	2.34	1.69	1.18	0.76	0.44

Table 5. Potency and gauge values when contaminated with double AO with different significant values using sequential selection

		2 blocks estimation					4 blocks estimation					6 blocks estimation				
T=240	α	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ	3σ	5σ	7σ	9σ	12σ
Potency	0.001	20.2	71.8	95.0	98.8	100	19.1	70.3	94.6	99.5	100	19.6	71.6	95.1	99.3	100
	0.01	48.7	90.1	98.7	100	100	47.3	89.0	98.8	99.9	100	47.1	90.6	98.6	99.9	100
	0.025	64.0	94.5	99.3	99.9	100	61.7	93.3	99.4	99.9	100	60.5	93.1	99.5	100	100
Gauge	0.001	0.03	0.02	0.01	0.01	0.00	0.11	0.01	0.03	0.00	0.00	0.09	0.05	0.02	0.00	0.00
	0.01	0.78	0.51	0.28	0.12	0.02	0.68	0.34	0.18	0.08	0.01	0.64	0.30	0.11	0.03	0.01
	0.025	2.83	1.78	1.07	0.57	0.17	2.24	1.38	0.74	0.33	0.08	1.93	1.09	0.52	0.22	0.05

As examined closely in Table 6, we found symmetry pattern for potency values when the AO is located in the sample observations. Performance of IIS is affected by the location of AO where the potency value is the highest when AO is located in the middle of the observation. The two outliers were located together in the same sample half as suggested in [14] to allow immediate outliers detection. The idea behind this is that the impulse indicators in the half sample cover all the outliers. Table 6 shows that overall gauge values obtained are satisfactory because their values are less than 3%, which indicates that they are tightly clustered around the chosen significance level. Higher target size will lead to too many retention dummies [22]. Thus, this explains higher gauge values if we set $\alpha=2.5\%$.

Table 6. Potency and gauge values at different locations.

	Location of single AO					Locations of double AO				
	0.1	0.3	0.5	0.7	0.9	[0.1,0.2]	[0.3,0.4]	[0.5,0.6]	[0.7,0.8]	[0.9,1]
Potency	95.2	97.2	97	96.1	94.4	95.3	96.6	97.0	95.8	94.3
Gauge	0.66	0.67	0.67	0.64	0.63	0.23	0.23	0.13	0.20	0.20

*Location of AO is given as a share of sample length, T

4. EMPIRICAL APPLICATION

This subsection will demonstrate the utilization of IIS to the stock return of the FTSE USA Shariah index. The series consists of 142 closing monthly prices from October 2007 until July 2019. We aimed to characterize any structural change such as recessionary period triggered by any financial crisis via indicator saturation approach. The reference model used is the local level model, as mentioned in section 2. The selection of target size or significance level is determined by $1/T = 0.0071$. Section 3 shows the significance of total number of blocks, m , on the study output or outliers detected. Thus, we decided to generate the results using two, four, six and ten blocks, before combining all significant indicators in final terminal model. An interesting pattern emerged when we examine the results of IIS tabulated in Table 7 and Figure 1.

Table 7. Outliers detected in FTSE USA Shariah stock return using IIS

	Number of blocks, m				
	2	4	6	8	10
Jan-08	Jan-08	Jan-08	Jan-08	Jan-08	Jan-08
Sep-08	Sep-08	Sep-08	Sep-08	Sep-08	Sep-08
Jan-09	Jan-09	Jan-09	Jan-09	Jan-09	Jan-09
Apr-10					

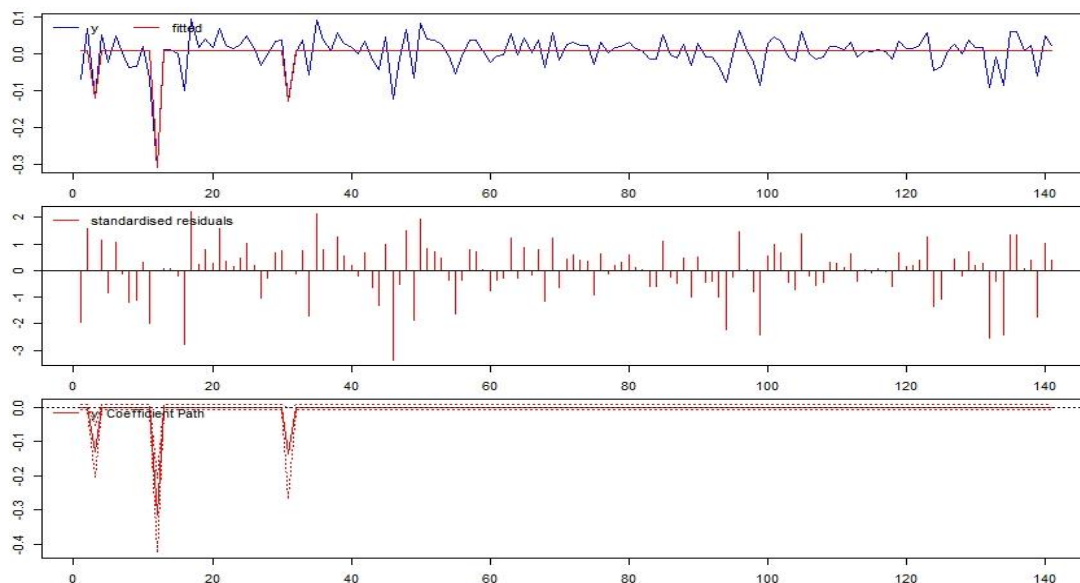


Figure 1. Results obtained from the impulse indicator saturation model. The uppermost figure portrays observed (blue) and fits (red) time series, respectively. The central figure portrays the standardized residuals; while the bottommost figure portrays the coefficient path together with the intercept and estimated 95% confidence interval

The dummy variables retained in the model are the same even though we vary the number of blocks, m using sequential selection algorithm. Thus, we concluded that the outliers detected in January 2008, September 2008 and January 2009 correspond to the recessionary period due to the world financial crisis. This finding demonstrates that IIS is a useful approach to identify outliers in financial data as compared to recent study using shariah compliant stock prices in Malaysia by [31]. This study differs from [31] research that was prone to methodological inadequacies due to inability of Box-Whisker plot to handle non-stationary data and to model the outliers detected. In addition, the box-whisker plot approach is not able to perform significance testing to the outliers captured compared to IIS in GETS modelling.

5. CONCLUSION

The presence of outliers in economic data may have pernicious effects on model estimation and forecast accuracy. Hence, this study aimed to study the efficiency of IIS in identifying outliers using gets package through Monte Carlo simulations. This study has shown that IIS is very useful in detecting outliers in the state-space framework even though the level component varies over time. We explored several variables that may influence the efficiency of the proposed method using Monte Carlo simulations. First, the performance of IIS relies heavily on the size of outliers. Then, the location of AO added have a significant impact on the potency values. IIS procedure can easily identify the significant indicators if the outliers are added in the same sample. Next, the number of blocks, m also plays a vital role to drive the performance of IIS. However, too many splits in the sample did lead to any detection of significant indicators. It is essential to locate and quantify the economic shock that characterizes and model any structural break in the series examined. Aside from that, the future works can also utilize the step indicator saturation (SIS) approach in order to capture multiple structural breaks.

ACKNOWLEDGEMENTS

The authors would like to extend their sincere gratitude to the Ministry of Higher Education Malaysia (MOHE) for the financial supports received for this work under FRGS grant (203/PMATHS/6711604).

REFERENCES

- [1] J. L. Castle and D. F. Hendry, "Modelling our Changing World," Palgrave Pivot, Cham, 2019.
- [2] David J. Hand, "Discussion contribution on 'Data mining reconsidered: encompassing and the general-to-specific approach to specification search' by Hoover and Perez," *The Econometrics Journal*, vol. 2, no. 2, pp. 241–243, 1999, doi: 10.1111/1368-423x.00030.
- [3] M. C. Lovell, "Data Mining," *The Review of Economics and Statistics*, vol. 65, no. 1, pp. 1–12, 1983.
- [4] D. F. Hendry and H. Krolzig, "Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez," *The Econometrics Journal*, vol. 2, no. 2, pp. 202–219, 1999, doi: 10.1111/1368-423x.00027.
- [5] J. Castle and N. Shephard, "The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry," OUP Oxford, pp. 1–464, 2009, doi: 10.1093/acprof:oso/9780199237197.001.0001.
- [6] C. Santos, D. F. Hendry, and S. Johansen, "Automatic selection of indicators in a fully saturated regression," *Computational Statistics*, vol. 23, no. 2, pp. 317–335, 2008, doi: 10.1007/s00180-007-0054-z.
- [7] J. L. Castle, J. A. Doornik, and D. F. Hendry, "Modelling non-stationary 'Big Data,'" *International Journal of Forecasting*, 2020, doi: 10.1016/j.ijforecast.2020.08.002.
- [8] Andrew Harvey, Siem Jan Koopman, and Jeremy Penzer, "Messy Time Series: a Unified Approach," *Advances in Econometrics*, vol. 13, pp. 103-143, 1998.
- [9] D. F. Hendry, "An econometric analysis of US food expenditure, 1931-1989," in J. R. Magnus and M. S. Morgan, *Methodology and tacit knowledge: two experiments in econometrics*, Chichester: John Wiley and Sons, 1999, pp. 341-361.
- [10] D. F. Hendry and J. A. Doornik, "Empirical Model Discovery and Theory evaluation," London: The MIT Press, 2014.
- [11] S. Johansen and B. Nielsen, "An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator," *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, 2009, doi: 10.1093/acprof:oso/9780199237197.003.0001.
- [12] N. R. Ericsson, "How biased are U.S. government forecasts of the federal debt?," *International Journal of Forecasting*, vol. 33, no. 2, pp. 543–559, 2017, doi: 10.1016/j.ijforecast.2016.09.001.
- [13] R. Mariscal and A. Powell, "Commodity Price Booms and Breaks: Detection, Magnitude and Implications for Developing Countries," *Ssrn*, no. January, 2014, doi: 10.2139/ssrn.2384422.
- [14] M. Marczak and T. Proietti, "Outlier detection in structural time series models: The indicator saturation approach," *International Journal of Forecasting*, vol. 32, no. 1, pp. 180–202, 2016, doi: 10.1016/j.ijforecast.2015.04.005.
- [15] J. R. Stillwagon, "Non-linear exchange rate relationships: An automated model selection approach with indicator saturation," *North American Journal of Economics and Finance*, vol. 37, pp. 84–109, 2016, doi: 10.1016/j.najef.2016.03.009.

- [16] J. L. Castle, J. A. Doornik, and D. F. Hendry, "Evaluating Automatic Model Selection," *Journal of Time Series Econometrics*, vol. 3, no. 1, 2011, doi: 10.2202/1941-1928.1097.
- [17] F. Pretis, L. Schneider, J. E. Smerdon, and D. F. Hendry, "Detecting Volcanic Eruptions in Temperature Reconstructions By Designed Break-Indicator Saturation," *Journal of Economic Surveys*, vol. 30, no. 3, pp. 403–429, 2016, doi: 10.1111/joes.12148.
- [18] J. H. Stock and M. W. Watson, "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1167–1179, 2002, doi: 10.1198/016214502388618960.
- [19] S. J. Commandeur, Jacques J. F., and Koopman, *An Introduction to State Space Time Series Analysis*, First. New York: Oxford University Press, 2007.
- [20] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space methods*, Second. Oxford: Oxford University Press, 2012.
- [21] A. C. Atkinson, S. J. Koopman, and N. Shephard, "Detecting shocks: Outliers and breaks in time series," *Journal of Econometrics*, vol. 80, no. 2, pp. 387–422, 1997, doi: 10.1016/S0304-4076(97)00050-X.
- [22] M. Bergamelli and G. Urga, "Detecting Multiple Structural Breaks: Dummy Saturation vs Sequential Bootstrapping. With an Application to the Fisher Effect for US," pp. 1-45, 2 April 2014.
- [23] J. Reade, and G. Sucarrat, "General-to-Specific (GETS) Modelling and Indicator Saturation with the R Package gets," Economics Series Working Papers 794, University of Oxford, Department of Economics, pp. 1–30, 2016.
- [24] J. L. Castle, J. A. Doornik, and D. F. Hendry, "Robust Discovery of Regression Models," *Econometrics and Statistics*, Available online 1 June 2021, doi: 10.1016/j.ecosta.2021.05.004.
- [25] C. Santos, D. F. Hendry, and S. Johansen, "Automatic selection of indicators in a fully saturated regression," *Computational Statistics*, vol. 23, no. 2, pp. 317–335, 2008, doi: 10.1007/s00180-007-0054-z.
- [26] S. Johansen and B. Nielsen, "An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator," *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, 2009, doi: 10.1093/acprof:oso/9780199237197.003.0001.
- [27] D. F. Hendry and H. M. Krolzig, "The properties of automatic gets modelling," *Economic Journal*, vol. 115, no. 502, pp. C32–C61, 2005, doi: 10.1111/j.0013-0133.2005.00979.x.
- [28] C. Chang, Ih, Tiao, George C. and Chen, "Estimation of time series parameters in the presence of outliers," *Technometrics*, vol. 30, no. 2, pp. 193–204, 1988.
- [29] C. Chen and L.-M. Liu, "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 284–297, 1993, doi: 10.2307/2290724.
- [30] F. Pretis, J. J. Reade, and G. Sucarrat, "Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks," *Journal of Statistical Software*, vol. 86, no. 3, 2018, doi: 10.18637/jss.v086.i03.
- [31] N. A. Bakar, "Monte Carlo Simulation for Data Volatility Analysis of Stock Prices in Islamic Finance for Malaysia Composite Index," *International Journal of Advanced Engineering Research and Science*, vol. 6, no. 3, pp. 6-12, 2019, doi: 10.22161/ijaers.6.3.2.