PROSIDING SEMINAR NASIONAL STATISTIKA
UNIVERSITAS DIPONEGORO 2013
ISBN: 978-602-14387-0-1

KOMBINASI PROSEDUR PEMODELAN SUBSET ARIMA DAN DETEKSI OUTLIER UNTUK PREDIKSI DATA RUNTUN WAKTU

Tarno
Program Studi Statistika FSM UNDIP
e-mail: tarno@undip.ac.id

Abstrak

Autoregressive Integrated Moving Average (ARIMA) merupakan salah satu model paling popular yang biasa digunakan untuk prediksi data runtun waktu. Tahapan yang paling krusial dalam pemodelan ARIMA adalah identifikasi dan pemilihan model terbaik berdasarkan karakteristik data. Tahapan-tahapan tersebut membutuhkan pemahaman yang mendalam tentang karakteristik data berdasarkan pola fungsi autokorelasi (FAK) dan fungsi autokorelasi parsial (FAKP). Tujuan dari tahap identifikasi adalah mencocokkan pola FAK dan FAKP sampel dengan pola FAK dan FAKP teoritis untuk menentukan order ARIMA yang tepat, termasuk order dari Subset ARIMA. Berdasarkan order yang ditentukan melalui tahapan identifikasi tersebut akan digunakan untuk penentuan model ARIMA atau Subset ARIMA yang tepat. Namun demikian apabila pada tahapan identifikasi ini dapat diketahui terdapat observasi yang secara mencolok berbeda dengan observasi lainnya, maka dapat diindikasikan bahwa dalam populasi terdapat data pencilan atau *outlier*. Pada kasus data runtun waktu, *outlier* dapat mempengaruhi kesesuaian model. Dalam tulisan ini, diusulkan prosedur pemodelan Subset ARIMA yang dikombinasikan dengan pendeteksian *outlier* untuk prediksi data runtun waktu. Proses tersebut dimulai dengan model ARIMA yang melibatkan lag yang signifikan berdasarkan pola FAK dan FAKP. Penambahan order AR atau MA didasarkan pada konsep *over-fitting*, yaitu berdasarkan pola FAK dan FAKP dari residual. Untuk menganalisis kesesuaian model salah satunya dilakukan dengan cara pendeteksian pengamatan *outlier*. Apabila terdapat outlier dalam data, maka perlu diatasi dengan cara memasukkan pengamatan *outlier* tersebut ke dalam model. *Outlier* diklasifikasikan menjadi *Additive Outlier* (AO), *Innovative Outlier* (IO), *Level Shift* (LS) dan *Transitory Change* (TC). Kombinasi prosedur tersebut diterapkan untuk mengkonstruksikan model inflasi di Indonesia.

Kata kunci: Runtun waktu; Subset ARIMA; FAK; FAKP; Outlier.

1. ` Pendahuluan

Autoregressive Integrated Moving Average (ARIMA) model merupakan metode yang dikenalkan oleh Box-Jenkins (1970). Sampai saat ini, ARIMA merupakan salah satu model yang paling populer untuk prediksi data runtun waktu univariat. Model-model stasioner non musiman terdiri dari AR, MA dan ARMA, sedangkan model non stasioner non musiman terdiri dari ARI, IMA dan ARIMA. Apabila komponen musiman dimasukkan ke dalam model tersebut menjadi model musiman (SARIMA).

ISBN: 978-602-14387-0-1

Metode Box-Jenkins untuk pemodelan ARIMA terdiri dari beberapa tahapan, yaitu identifikasi, estimasi parameter, verifikasi model dan *forecasting*.

Secara umum, model ARIMA(p,d,q) dapat ditulis sebagai (lihat Box et al. 1970, Makridakis et al. (1998) and Wei (2006))

$$\phi_p(B)(1-B)^d Z_t = \theta_q(B)a_t \quad (1)$$

dengan $\phi_p(B) = 1-\phi_1 B-\phi_2 B^2 -\cdots-\phi_p B^p$,

$\theta_q(B) = 1-\theta_1 B-\theta_2 B^2 -\cdots-\theta_q B^q$,

dengan B adalah operator *backward shift*, p dan q masing-masing menyatakan order dari autoregressive dan moving average dan d menyatakan order dari *difference*. Sedangkan model SARIMA$(P,D,Q)^S$ dapat dinyatakan sebagai

$$\Phi_P(B^S)(1-B^S)^D Z_t = \Theta_Q(B^S)a_t \quad (2)$$

dengan $\Phi_P(B^S) = 1-\Phi_1 B^S-\Phi_2 B^{2S} -\cdots-\Phi_P B^{PS}$

$\Theta_Q(B^S) = 1-\Theta_1 B^S-\Theta_2 B^{2S} -\cdots-\Theta_Q B^{QS}$,

dengan B merupakan operator *backward shift*, P dan Q masing-masing menyatakan order musiman dari autoregressive dan moving average, D menyatakan order musiman dari *difference* dan S menyatakan periodisitas musiman.

Model ARIMA telah digunakan untuk prediksi di berbagai bidang terapan. Sebagai contoh, Al-Fattah (2006) menerapkan model ARIMA untuk prediksi gas alam A.S.; Aston (2007) menggunakan model SARIMA untuk prediksi; Chang et al. (2011) menggunakan ARIMA untuk prediksi arus lalu lintas jangka pendek; Ghosh (2004) melakukan prediksi arus lalu lintas di Dublin; Meyler (1998) memprediksi inflasi Irish; Ojo et al. (2009) menganalisis estimasi dan performa dari subset ARIMA; Spreen et al. (1979) menerapkan model subset AR untuk prediksi harga sapi bulanan; Suhartono et al. (2011) menggunakan subset, *multiplicative* atau model SARIMA *additive* untuk prediksi kedatangan turis.

Hampir semua tulisan sebelumnya terfokus untuk membahas model ARIMA, namun yang membahas subset ARIMA masih sangat terbatas. Perbedaan krusial antara model ARIMA dan Subset ARIMA adalah terletak pada penentuan order dari model.

ISBN: 978-602-14387-0-1

Penentuan order dari model ARIMA atau Subset ARIMA ditentukan berdasarkan konsep *over-fitting*.

Salah satu tahapan yang sangat penting dalam pemodelan ARIMA adalah identifikasi berdasarkan karakteristik data. Tahapan identifikasi ini bertujuan untuk menentukan order ARIMA atau Subset ARIMA yang tepat, yang akhirnya dapat menghasilkan model terbaik. Order dari suatu model ARIMA dapat ditentukan berdasarkan pola FAK dan FAKP. Dalam praktek, jika diberikan data runtun waktu $Z_1, Z_2, \cdots, Z_n$, FAK dan FAKP teoritis ρ_k dan ϕ_{kk} diestimasi menggunakan FAK sampel $\hat{\rho}_k$ dan FAKP sampel $\hat{\phi}_{kk}$. Penambahan order juga dapat ditentukan berdasarkan pola FAK dan FAKP dari residual.

Apabila telah diperoleh estimasi modelnya, maka salah satu cara untuk menentukan kesesuaian model adalah dengan melakukan deteksi *outlier* dalam data pengamatan. Jika terdapat *outlier* dalam data harus diatasi dengan cara memasukkan pengamatan *outlier* tersebut dalam model. *Outlier* dalam data tersebut dapat diklasifikasikan menjadi *Additive Outlier* (AO), *Innovative Outlier* (IO), *Level Shift* (LS) dan *Transitory Change* (TC). Dalam tulisan ini dibahas tentang prosedur pemodelan Subset ARIMA yang dikombinasikan dengan deteksi *outlier* untuk prediksi inflasi di Indonesia sebagai studi kasus.

2. Model Arima

Berdasarkan persamaan (1) dan (2) dapat dirumuskan model *multiplicative, additive* atau subset ARIMA non-musiman dan model ARIMA musiman.

Model Arima Multiplicative

Secara umum, model SARIMA *multiplicative* dapat ditulis sebagai:

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D Z_t = \theta_q(B)\Theta_Q(B^S)a_t \qquad (3)$$

Persamaan (3) biasa dikenal sebagai model SARIMA(p,d,q) $(P,D,Q)^S$. Model SARIMA *multiplicative* akan tereduksi menjadi model ARIMA(p,d,q) ketika tidak ada efek musiman, serta menjadi ARMA(p,q) ketika runtun waktu tersebut stasioner.

ISBN: 978-602-14387-0-1

Model Arima Additive

model sarima additive yang digeneralisasi dapat ditulis sebagai:

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p-\Phi_1 B^S-\Phi_2 B^{2S}-\cdots-\Phi_P B^{PS})(1-B)^d(1-B^S)^D Z_t = (1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q-\Theta_1 B^S-\Theta_2 B^{2S}-\cdots-\Theta_Q B^{QS})a_t \quad (4)$$

Model ini merupakan jumlahan antara model non-musiman dan musiman tanpa parameter *multiplicative*.

Model Subset Arima

Model Subset ARIMA merupakan bagian dari model ARIMA tergeneralisasi, sehingga tidak dapat dinyatakan dalam bentuk umum. Model subset ARIMA ini merupakan himpunan bagian dari model ARIMA. Sebagai contoh subset ARIMA([1,5],0,[1,12]) dapat ditulis sebagai:

$$(1-\phi_1 B-\phi_5 B^5)Z_t=(1-\theta_1 B-\theta_{12}B^{12})a_t \ .$$

Dengan demikian model subset ARIMA merupakan model ARIMA dengan beberapa parameternya sama dengan nol.

Prosedur Pemodelan Subset ARIMA

Identifikasi Model

Jika diberikan data runtun waktu $Z_1, Z_2, \cdots, Z_n$, FAK ρ_k dan FAKP ϕ_{kk} diestimasi dengan FAK sampel $\hat{\rho}_k$ dan FAKP sampel $\hat{\phi}_{kk}$. Menurut Bartlett bahwa FAK ρ_k berdistribusi normal dengan mean nol dan varinasi $(1/n)(1+2\sum_{i=1}^{k-1} r_i^2)$, r_i: estimasi FAK pada lag-i; dan FAKP ϕ_{kk} berdistribusi normal dengan mean nol dan variansi (1/n), n: banyaknya obsservasi (lihat Box et al. (1970), Makridakis et al. (1998), Wei (2006)). Menurut Wei (2006), karakteristik dari FAK dan FAKP teoritis untuk proses stasioner AR(p), MA(q) dan ARMA(p,q) ditunjukkan seperti Tabel I.

TABEL 1. KARAKTERISTIK FAK DAN FAKP TEORITIS UNTUK PROSES STASIONER

Proses	**FAK**	**FAKP**
AR(*p*)	Turun secara ekponensial atau membentuk gelombang sinus	Terputus setelah lag *p*
MA(*q*)	Terputus setelah lag *q*	Turun secara ekponensial atau membentuk gelombang sinus
ARMA(*p*,*q*)	Terputus setelah lag (*q*-*p*)	Terputus setelah lag (*p*-*q*)

ISBN: 978-602-14387-0-1

ARMA (p,q) ditetapkan sebagai model awal, jika $\hat{\phi}_{kk}$ terputus setelah lag-p dan $\hat{\rho}_k$ terputus setelah lag-q untuk suatu bilangan bulat non-negatif p, q. Jika terdapat sebarang k himpunan bagian dari {1,2,3,…,p} sedemikian hingga $\hat{\phi}_{kk}$ lebih besar dari selang kepercayaan FAKP dan nol untuk yang lain , atau untuk sebarang k himpunan bagian dari {1,2,3,…,q} sedemikian hingga $\hat{\rho}_k$ lebih besar dari selang kepercayaan FAK dan nol untuk yang lain, maka model awalnya adalah subset ARIMA dengan order k dengan k merupakan himpunan bagian dari {1,2,3,…,p} atau k merupakan himpunan bagian dari {1,2,3,…,q}. Proses identifikasi ini akan digunakan untuk menentukan estimasi awal parameter dalam model.

Estimasi Model

Model-model yang telah teridentifikasi pada tahapan sebelumnya, parameter-parameter modelnya dapat diestimasi berdasarkan data. Untuk estimasi parameter model dapat digunakan metode *Maximum Likelihood* (ML), metode *Unconditional Least Squares* (ULS) atau metode *Conditional Least Squares* (CLS). Estimasi awal yang telah diperoleh dapat digunakan sebagai nilai awal dari metode estimasi secara iterative.

Verifikasi Model

Pada tahapan ini, model tentative diverifikasi dengan cara melakukan uji signifikansi parameter yang diestimasi dan mengevaluasi kesesuaian model (asumsi white noise dan residual berdistribusi normal dengan mean nol variansi konstan) .

Proses penambahan order dilakukan apabila:

- Parameter yang diestimasi semuanya signifikan, tetapi berdasrkan uji Ljung-Box mengindikasikan residual tidak memenuhi syarat white noise.
- Tidak semua pamameter yang diestimasi tidak signifikan, khususnya parameter yang berada di antara order-order yang lain dan residual tidak memenuhi syarat white noise.

Analisis *Outlier*

Outlier adalah pengamatan yang secara jelas berbeda dengan pengamatan lainnya. Dalam kasus runtun waktu, o*utlier* diklasifikasikan menjadi *Additive Outlier* (AO), *Innovative Outlier* (IO), *Level Shift* (LS) dan *Transitory Change* (TC). *Additive Outlier* (AO) hanya berpengaruh pada pengamatan ke-T, sedangkan tiga jenis *outlier* lainnya yaitu *Innovative Outlier* (IO), *Level Shift* (LS) dan *Transitory Change* (TC) berpengaruh

ISBN: 978-602-14387-0-1

pada pengamatan ke-T, T+1, …. . Menurut Wei [11], secara umum model dengan *outlier* ditulis sebagai:

$$Z_t = \sum_{j=1}^{k} \varpi_j v_j(B) I_j^{T_j} + \frac{\theta(B)}{\phi(B)} a_t \quad (5)$$

dengan

$v_j(B) = 1$ untuk AO,

$v_j(B) = \frac{\theta(B)}{\phi(B)}$ untuk IO

$v_j(B) = \frac{1}{1-B}$ untuk LS

$v_j(B) = \frac{1}{(1-\delta B)}$; $0 < \delta < 1$ untuk TC, dan pada TC nilai δ yang sering digunakan adalah 0,7. Salah satu cara untuk penanganan *outlier* adalah dengan cara memasukkan pengamatan *outlier* ke dalam model.

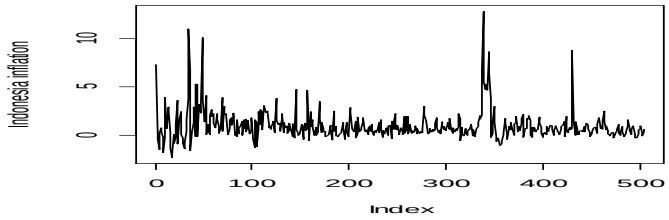
3. Hasil dan Pembahasan

Untuk mengimplementasikan prosedur pemodelan Subset ARIMA dan pendeteksian *outlier*, digunakan data inflasi Indonesia sebagai studi kasus. Data pengamatan merupakan data inflasi bulanan dari Januari 1970 sampai dengan Februari 2012 dan diperoleh dari Badan Pusat Statistik (BPS) (lihat www.bps.go.id). Prosedur pemodelan yang diusulkan adalah sebagai berikut.

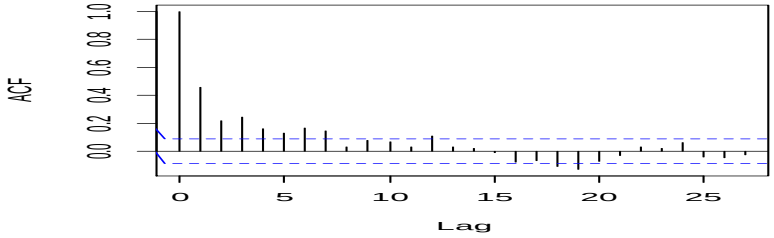
Identifikasi Model

Berdasarkan plot data runtun waktu menunjukkan bahwa data inflasi Indonesia bersifat stasioner dalam mean, namun terdapat beberapa pengamatan yang berbeda secara jelas dengan pengamatan lainnya. Hal ini menunjukkan bahwa dalam data ada *outlier*. Menurut plot FAK lag-lag yang signifikan adalah lag-1 sampai lag-7, yang berarti bahwa data stasioner dalam mean. Sedangkan menurut plot FAKP lag-lag yang signifikan adalah lag-1 dan lag-3. Dengan demikian dapat diidentifikasi model AR ([1,3],0,0) sebagai model awal.

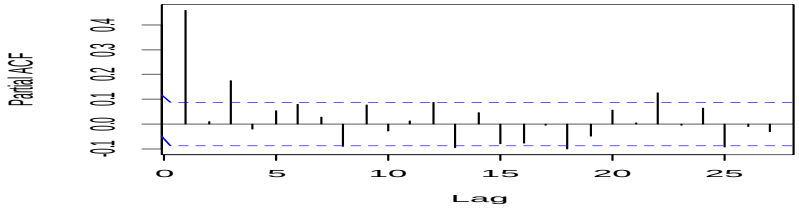
ISBN: 978-602-14387-0-1



Gambar 1. Plot data runtun waktu inflasi Indonesia



Gambar 2. Plot FAK dari data inflasi Indonesia



Gambar 3. Plot FAKP data inflasi Indonesia

Untuk mengidentifikasi model MA, dapat juga ditentukan dengan cara mencermati pola FAK dari residual berdasarkan model paling sederhana, dalam hal ini model subset AR([1,3],0,0). Dari plot FAK residual, lag-lag yang signifikan adalah lag-2, lag-7, lag-12, lag-19 dan lag-24. Beberapa model yang berhasil diidentifikasi adalah ARIMA ([1,3],0,[7,12,19]), ARIMA ([1,3],0,[7,12,24]), ARIMA ([1,3],0,[7,12,19,24]), ARIMA ([1,3],0,[2,7,12,19]), ARIMA ([1,3],0,[2,7,19,24]), ARIMA ([1,3],0,[2,7,12,19,24]), ARIMA ([1,3],0,[2,7,12,24]), ARIMA ([1,3],0,[2,7,19])$(1,0,0)^{12}$, ARIMA ([1,3,12],0,[2,7,19,24]).

Estimasi Model

Dengan memperhatikan lag-lag signifikan yang telah diidentifikasi pada tahapan sebelumnya, diperoleh tiga estimasi model subset ARIMA yang signifikan yaitu:

1. ARIMA ([1,3],0,[2,7,12,24])

Parameter model diestimasi dengan menggunakan metode conditional least squares (CLS). Berdasarkan metode CLS, parameter lag-1, lag-3 dari suku AR adalah signifikan

ISBN: 978-602-14387-0-1

dan parameter lag-2, lag-7, lag-12, lag-24 dari suku MA juga signifikan. Sehingga diperoleh estimasi model ARIMA ([1,3],0,[2,7,12,24]) sebagai berikut:

$$(1-0.50494B-0.12516B^3)Z_t = 0.351496 + (1-0.13642B^2+0.13990B^7+0.15964B^{12}+0.11831B^{24})a_t$$

2. ARIMA ([1,3],0,[2,7,19])$(1,0,0)^{12}$

Dengan menggunakan metode CLS diperoleh estimasi model musiman multiplikatif ARIMA([1,3],0,[2,7,19])$(1,0,0)^{12}$ yang dapat dituliskan sebagai:

$$(1-0.47780B-0.13437B^3)(1-0.14029B^{12})Z_t = 0.330211 + (1-0.10776B^2+0.11783B^7+0.09743B^{24})a_t$$

3. ARIMA ([1,3,12],0,[2,7,19,24])

Model ketiga yang signifikan adalah model ARIMA ([1,3,12],0,[2,7,19,24]) dengan estimasi modelnya dapat ditulis sebagai:

$$(1-0.45804B-0.13576B^3-0.08475B^{12})Z_t = 0.317442 + (1-0.10123B^2+0.11481B^7-0.11421B^{19}+0.10025B^{24})a_t$$

Dengan demikian diperoleh tiga model signifikan dengan nilai AIC dan SBC seperti ditunjukkan pada Tabel 2.

TABEL 2. TIGA MODEL YANG SIGNIFIKAN

Model	MSE	RMSE	AIC	SBC
ARIMA ([1,3],0,[2,7,12,24])	1,66498	1,29034	1695,36	1724,92
ARIMA ([1,3],0,[2,7,19])$(1,0,0)^{12}$	1,69413	1,30159	1702,94	1732,49
ARIMA ([1,3,12],0,[2,7,19,24])	1,69486	1,30187	1704,14	1737,91

Berdasarkan nilai AIC dan SBC dari ketiga model tersebut dipilih satu calon model terbaik yaitu ARIMA ([1,3],0,[2,7,12,24]). Namun hal ini belum cukup karena model tersebut masih perlu dilakukan pengujian asumsi untuk mendapatkan model yang sesuai.

Verifikasi Model

Model yang telah diestimasi pada tahapan sebelumnya, perlu dilakukan uji kesesuaian model. Residual dari model harus memenuhi asumsi white noise, berdistribusi normal dan tidak terjadi heteroskedastisitas. Berdasarkan uji Ljung-Box semua model pada Tabel 5 memenuhi asumsi independensi residual, tetapi residual tidak berdistribusi normal dan berdasarkan uji LM terdapat efek ARCH. Selain itu berdasarkan pendeteksian adanya *outlier* ditemukan 13 *Additive Outlier* (AO).

ISBN: 978-602-14387-0-1

Karena model yang diestimasi belum sepenuhnya memenuhi asumsi yang disyaratkan, maka perlu dilakukan tindakan perbaikan yaitu dengan cara memasukkan 13 pengamatan AO tersebut ke dalam model serta dengan memperhatikan efek ARCH. Setelah dilakukan estimasi model dengan memasukkan AO dan memperhatikan efek ARCH maka diperoleh model sebagai berikut.

$$\begin{aligned} Z_t = {} & 0{,}280774 + 6{,}94542\,AOJAN1970 + 3{,}71636\,AONOV1970 + 3{,}49698\,AONOV1971 + \\ & 8{,}82120\,AONOV1972 + 5{,}94907\,AODEC1972 + 7{,}07233\,AOJAN1974 + \\ & 2{,}82703\,AOAPR1974 + 3{,}18730\,AOSEP1975 + 4{,}11414\,AOJAN1982 + \\ & 3{,}1821\,AOJAN1983 + 7{,}24787\,AOFEB1998 + 4{,}48714\,AOJUL1998 + \\ & 7{,}34447\,AOOCT2005 + \frac{(1 + 0{,}11909B^{7} + 0{,}14296B^{12})}{(1 - 0{,}48471B - 0{,}15928B^{3})} a_t . \end{aligned} \quad (6)$$

dengan $a_t \sim N(0, \sigma_t^2)$ dan $\sigma_t^2 = 0{,}0997 + 0{,}2431 a_{t-1}^2 + 0{,}07405 \sigma_{t-1}^2$.

Nilai AIC dan SBC dari model masing-masing adalah 1354,364 dan 1430,7. Dengan demikian model inilah yang akan digunakan untuk prediksi data inflasi Indonesia.

4. Kesimpulan

Order dari model subset ARIMA ditentukan berdasarkan pola FAK dan FAKP dari data runtun waktu yang dikombinasikan dengan pola FAK dan FAKP residual model awal yang signifikan. Penentuan order yang tepat akan mempengaruhi akurasi model. Prosedur pemodelan subset ARIMA yang digabungkan dengan deteksi *outlier* yang diterapkan pada studi kasus data inflasi di Indonesia dapat meningkatkan akurasi model. Hal ini didasarkan pada menurunnya nilai AIC dan SBC bila dibandingkan dengan nilai AIC dan SBC sebelum memasukkan *outlier* dalam model.

DAFTAR PUSTAKA

A. Meyler, G. Kenny and T. Quinn. Forecasting Iris inflation using ARIMA models, Technical Paper, Economics Analysis, Research and Publications Department, Central Bank of Ireland, 1998.

B. Ghosh, B. Basu and M. O'Mahony. Time series forecasting for vehicular traffic flow in Dublin, 2004.

F.J. Ojo and T.O. Olatayo. On the Estimation and Performance of Subset Autoregressive Integrated Moving Average Models, European Journal of Scientific Research, Vol.28 No.2, , 2009, pp.287-293.

G. Chang, Y. Zhang, D. Yao and Y. Yue. Short-term traffic flow forecasting methods, ICCTP2011, 2011.

G.E.P. Box and G.M. Jenkins. Time series analysis: forecasting and control, Holden-Day, San Francisco, 1970.

ISBN: 978-602-14387-0-1

J.A.D. Aston, D.F. Findley, T.S. Mcelroy, K.C. Wills and D.E.K. Marten. New ARIMA models for seasonal time series and their application to adjustment and forecasting, Research Report Series, Statistics, No.14, 2007.

S. M. Al-Fattah. Time Series Modeling for U.S. natural Gas forecasting, E-Journal of Petroleum Management and Economics, 2006. http://www.petroleumjournals.com/

S. Makridakis, S. C. Wheelwright and R.J. Hyndman. Forecasting: Methods and Applications, John Wiley & Sons Inc., New York, 1998.

Suhartono and Muhammad Hisyam Lee. Forecasting of tourist arrivals using subset, multiplicative or additive seasonal ARIMA Model, MATEMATIKA, Volume 27, Number 2, 2011, 169-182.

T.H. Spreen, R.E. Mayer, J.R. Simpson and J.T. McClave. Forecasting monthly slaughter cow prices with subset autoregressive model, Southern Journal of Agricultural Economics, No. 1751, 1979, pp.126-131.

W.W.S. Wei. Time Series Analysis: Univariate and Multivariate Methods, Second Edition, Pearson Education Inc. Boston, 2006.