International Journal of Natural Sciences Research
2014 Vol. 2, No. 9, pp. 156-164
ISSN(e): 2311-4746
ISSN(p): 2311-7435

ANOTHER LOOK AT THE SARIMA MODELLING OF THE NUMBER OF DENGUE CASES IN CAMPINAS, STATE OF SAO PAULO, BRAZIL

Ette Harrison Etuk[1] --- Nathaniel Ojekudo[2]

[1]*Department of Mathematics/Computer Science Rivers State University of Science and Technology, Port Harcourt, Nigeria*

[2]*Department of Computer Science, Ignatius Ajuru University of Education, Port Harcourt, Nigeria*

ABSTRACT

Martinez, et al. [1] analysed monthly numbers of dengue cases as reported in Campinas, southeast Brazil from 1998 to 2008, by SARIMA methods. Assuming X is the original series, they analysed the logarithm of X + 1. The models they proposed and compared are of orders $(2,1,2)x(1,1,1)_{12}$, $(2,1,1)x(1,1,1)_{12}$, $(1,1,2)x(1,1,1)_{12}$, $(1,1,1)x(1,1,1)_{12}$, $(2,1,3)x(1,1,1)_{12}$, and $(1,1,3)x(1,1,1)_{12}$. Using the R software, they chose the SARIMA$(2,1,2)x(1,1,1)_{12}$ model as the best on the basis of Akaike information criterion, AIC. The result in this work is different: the SARIMA$(2,1,1)x(1,1,1)_{12}$ model is herein adjudged as the best on the same minimum AIC grounds.

Keywords: Dengue, SARIMA, Time series analysis, Statistics, seasonal series, Eviews, AIC.

Contribution/ Originality

This paper's primary contribution is that monthly recorded dengue numbers in Campinas, South east Brazil, follow a SARIMA $(2, 1, 1) \times (1, 1, 1)_{12}$ model. It was previously believed that a SARIMA $(2, 1, 2)x(1, 1, 1)_{12}$ model was the better model. The Eviews software was used to do the analysis. Residual analysis of the chosen model shows that it is very adequate.

1. INTRODUCTION

Martinez, et al. [1] analyzed monthly recorded numbers of dengue in Campinas in Southeast Brazil using Box-Jenkins methods. They analyzed the logarithms of the increment of the raw data by 1. The approach they adopted was the seasonal autoregressive integrated moving average (SARIMA) approach. This was sequel to an observation of a seasonal tendency in the time series, the dengue numbers tending to increase during the rainy seasons and reduce during the dry seasons.

They considered six SARIMA models of orders: $(2,1,2)x(1,1,1)_{12}$, $(2,1,1)x(1,1,1)_{12}$, $(1,1,2)x(1,1,1)_{12}$, $(1,1,1)x(1,1,1)_{12}$, $(2,1,3)x(1,1,1)_{12}$ and $(1,1,3)x(1,1,1)_{12}$ and adjudged the $(2,1,2)x(1,1,1)_{12}$ model as the most adequate on the basis of the minimum value of Akaike information criterion, AIC. It is noteworthy that they used the R software.

However, using the Eviews software a different conclusion is reached in this work. The motivation of this paper is therefore to highlight the fact that a different model is chosen as the

best amongst the same set of selected models. It is surprising where the difference could have arisen.

2. LITERATURE REVIEW

Of recent there has been a growing interest in SARIMA modelling. Many real life time series have seasonal natures. Box and Jenkins [2] proposed that such series could be modelled by SARIMA models. A few of seasonal time series that have been modelled by SARIMA techniques are rainfall [3], inflation [4], microwave transmission [5], temperature [6], electricity [7] and foreign exchange rate [8]. It has been demonstrated that for intrinsically seasonal series SARIMA models outdo the ordinary autoregressive integrated moving average (ARIMA) models [8].

3. MATERIALS AND METHODS

The data for this work as published in Martinez, et al. [1] are the reported monthly numbers of dengue from January 1998 to December 2009. As in [1], only the 132 numbers from 1998 to 2008 are analysed. The 2009 values were used to validate the fitted model in [1].

3.1. Sarima Models

A stationary time series $\{X_t\}$ is said to follow an autoregressive moving average model of orders p and q denoted by ARMA(p,q) if it satisfies the following difference equation

$$X_t - \alpha_1 X_{t-1} - \alpha_2 X_{t-2} - \ldots - \alpha_p X_{t-p} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q} \quad (1)$$

where $\{\varepsilon_t\}$ is a white noise process and the α's and the β's are constants such that the model is both stationary and invertible. The model could be written as

$$A(L)X_t = B(L)\varepsilon_t \quad (2)$$

where $A(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_p L^p$ and $B(L) = 1 + \beta_1 L + \beta_2 L^2 + \ldots + \beta_q L^q$ and L is the backshift operator defined by $B^k X_t = X_{t-k}$. For stationarity and invertibility it is well known that the zeros of A(L) and B(L) must be outside the unit circle respectively.

Most real-life time series are non-stationary. For such a series, Box and Jenkins [2] proposed that differencing up to an appropriate order could make the series stationary. Supposed is such an appropriate order. If the d^{th} difference of $\{X_t\}$, denoted by $\{\nabla^d X_t\}$, satisfies model (1), then $\{X_t\}$ is said to follow an *autoregressive integrated moving average model of orders p, d and q*, denoted by ARIMA(p, d, q). Here ∇ is the difference operator defined by $\nabla = 1 - L$.

If $\{X_t\}$ is seasonal of period s, Box and Jenkins [2] proposed that it may be modelled by

$$A(L)\Phi(L^s)\nabla^d \nabla_s^D X_t = B(L)\Theta(L^s)\varepsilon_t \quad (3)$$

where $\Phi(L)$ and $\Theta(L)$ are polynomials in L with coefficients such that the entire model (3) is both stationary and invertible. The seasonal difference operator ∇_s is defined by $\nabla_s = 1 - L^s$. Suppose the degrees of the polynomials $\Phi(L)$ and $\Theta(L)$ are P and Q respectively, the model (3) is

called a *seasonal autoregressive integrated moving average model or orders p, d, q, P, D, Q and s* denoted by SARIMA(p, d, q)x(P, D, Q)$_s$.

3.2. Sarima Model Estimation

Model estimation invariably begins with order determination. The orders p, d, q, P, D, Q and s must first of all be estimated. The seasonal period s is often obvious from experience or observation. Where the time-plot fails to clearly indicate the period, the autocorrelation function (ACF) could better do so by a significant spike at the seasonal lag. The ACF of a seasonal series of period s should exhibit oscillatory movements of the same period such that at integral multiple lags of s the correlations be positive and midway between such lags the correlations be negative. The autoregressive orders p and P may be suggestive by the non-seasonal and the seasonal cut-off lags of the partial autocorrelation function (PACF) respectively. Similarly the moving average orders q and Q may be determined respectively by the non-seasonal and the seasonal cut-off lags of the ACF.

Often it is enough to put d = D = 1. Before and after differencing, stationarity is tested by the Augmented Dickey Fuller (ADF) test.

After order determination estimation of the parameters may be done by a non-linear optimization technique because of the presence of items of the white noise process in the model. In this work the Eviews software shall be used. It employs the least error sum of squares criterion for model estimation.

A fitted model must be subjected to some residual analysis for confirmation of its goodness-of-fit to the data. The residuals of an adequate model are expected to be uncorrelated as well as follow a Gausian distribution of zero mean.

4. RESULTS AND DISCUSSION

The analyzed series is given by $\{Z_t\}$ where $Z = \log(X_t + 1)$ just as in Martinez, et al. [1]. The time-plot of Z, the realization of $\{Z_t\}$ that is actually analyzed, in Figure 1 shows some periodic movements of considerable regularity. However with a statistic value of -4.2 and the 1%, 5% and 10% critical values of -3.5, -2.9 and -2.6 respectively, the ADF test adjudges Z as stationary. The correlogram of Z in Figure 2 shows the ACF of a seasonal series of period 12. Z therefore cannot be stationary since the ACF exhibits oscillatory movements of period 12.

A seasonal (i.e. 12-month) differencing of Z produces the series herein called SDZ. The time-plot of SDZ in Figure 3 shows two peaks, one between 1999 to 2003 and the other between 2005 to 2007. Between the peaks is a trough. With a statistic value of -2.5 and the same critical values as mentioned above, the ADF test adjudges SDF as non-stationary.

A non-seasonal differencing of SDZ yields the series DSDZ which, with a statistic value of -5.6 and the same respective critical values as mentioned above, is adjudged as stationary. Its time-plot is shown in Figure 4. The correlogram of DSDZ of Figure 5 supports the stationarity hypothesis. With a negative significant spike at lag 12 in the ACF, there is indication of a 12-monthly seasonality as expected and the involvement of a seasonal moving average component of order 1. The comparable spikes at lags 11 and 13 suggests a $(0, 1, 1)x(0, 1, 1)_{12}$ component.

Similarly the spikes at lags 11, 12 and 13 in the PACF suggests a (1, 1, 0) x $(1, 1, 0)_{12}$ component. Combining these components, a (1, 1, 1)x$(1, 1, 1)_{12}$ model is suggestive. No wonder that this (1, 1, 1)x$(1, 1, 1)_{12}$ model is one of the chosen models. Though not the best of them in the AIC sense, it is still adequate in terms of uncorrelated and normally distributed residuals.

Martinez, et al. [1] compared the SARIMA models (p, 1, q)x$(1, 1, 1)_{12}$ with (p,q) equal to (2,2), (2,1), (1,2), (1,1), (2,3) and (1,3). They chose the first model (2, 1, 2)x$(1, 1, 1)_{12}$ on the grounds of minimum AIC. However the result of this work is different: the second model (2, 1, 1)x$(1, 1, 1)_{12}$ is the best (See Tables 1, 2 and 3).

Table 2 and table 3 show details of the estimation of the two competing models: (2, 1, 2)x(1, 1, 1) and (2, 1, 1)x(1, 1, 1) respectively. Figure 6 shows that the residuals of the chosen model (2, 1, 1)x$(1, 1, 1)_{12}$ are uncorrelated and Figure 7 shows that they follow a normal distribution with zero mean.

5. CONCLUSION

It is concluded that the dengue numbers follow a SARIMA (2, 1, 1)x$(1, 1, 1)_{12}$ and not a SARIMA(2, 1, 2)x(1, 1, 1) model amongst the six chosen models as earlier believed. This model has been shown to be the most adequate of all the proposed models. It is not certain where the difference could have emanated from. It is therefore recommended that further research be done to ascertain the source of the difference in the results obtained herein and in Martinez, et al. [1].

REFERENCES

[1] E. Z. Martinez, E. A. Soares Da Silva, and A. L. D. Fabbro, "A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of Sao Paulo, Brazil," *Rev. Soc. Bras. Med. Trop.*, vol. 44, pp. 436 – 440, 2011.

[2] G. E. P. Box and G. M. Jenkins, *Time series analysis, forecasting and control.* San Francisco: Holden-Day, 1976.

[3] E. H. Etuk, I. U. Moffat, and B. E. Chims, "Modelling monthly rainfall data of Port Harcourt, Nigeria by seasonal box-jenkins methods," *International Journal of Sciences*, vol. 2, pp. 60 – 67, 2013.

[4] O. A. Otu, G. A. Osuji, J. Opara, H. J. Machu, and A. J. Iheagwara, "Application of Sarima models in modelling and forecasting Nigeria's inflation rates," *American Journal of Applied Mathematics and Statistics*, vol. 2, pp. 16 – 28, 2014.

[5] F. K. Oduro-Gyimah, E. Harris, and K. F. Darkwah, "Sarima time series model application to microwave transmission of Yeji-Salaga (Ghana) line-of sight link," *International Journal of Applied Science and Technology*, vol. 2, pp. 40 – 51, 2012.

[6] K. Helman, "Sarima models for temperature and precipitation time series in the Czech Republic for the period 1961 – 2008," *Aplimat-Journal of Applied Mathematics*, vol. 4, pp. 281 – 290, 2011.

[7] Z. H. Ismail and K. A. Mahpol, "Sarima model for forecasting Malaysian electricity generated," *Matematika*, vol. 21, pp. 143 – 152, 2005.

[8] E. H. Etuk, "Modelling of daily Nigerian Naira-British pound exchange rates using Sarima methods," *British Journal of Applied Science and Technology*, vol. 4, pp. 222 – 234, 2014.

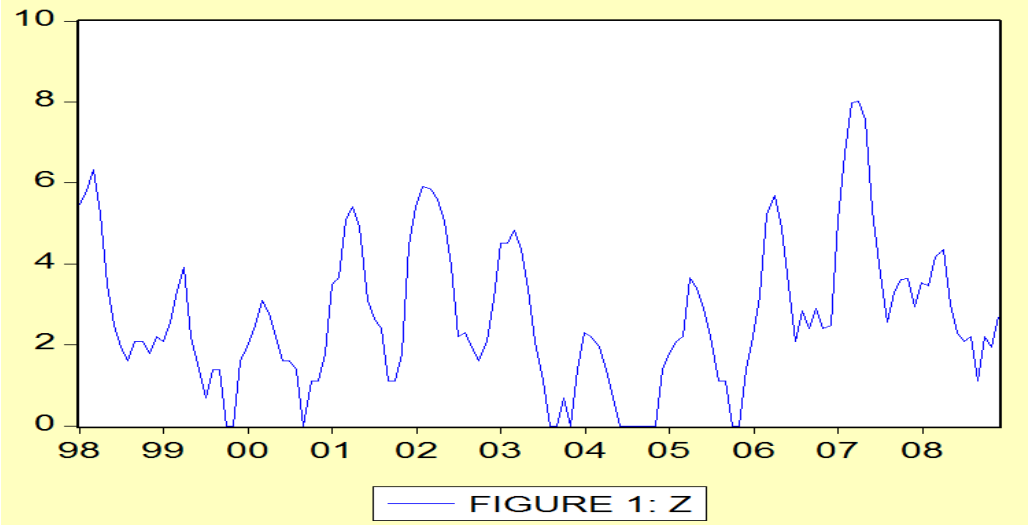
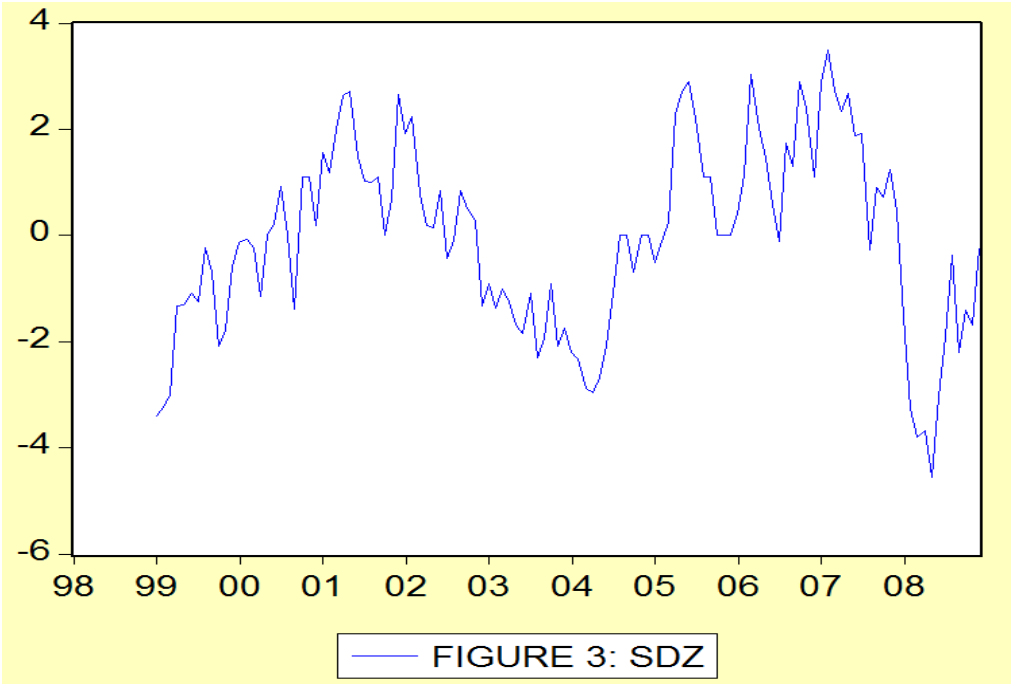


Figure-2. CORRELOGRAM OF Z

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.865	0.865	100.94	0.000
		2	0.631	-0.462	155.12	0.000
		3	0.361	-0.188	172.98	0.000
		4	0.121	0.003	175.01	0.000
		5	-0.046	0.051	175.31	0.000
		6	-0.121	0.084	177.37	0.000
		7	-0.086	0.200	178.43	0.000
		8	0.038	0.173	178.63	0.000
		9	0.209	0.121	184.90	0.000
		10	0.368	0.048	204.50	0.000
		11	0.489	0.124	239.43	0.000
		12	0.505	-0.170	276.98	0.000
		13	0.405	-0.166	301.40	0.000
		14	0.198	-0.241	307.25	0.000
		15	-0.046	-0.074	307.57	0.000
		16	-0.264	-0.080	318.19	0.000
		17	-0.414	-0.076	344.61	0.000
		18	-0.478	-0.108	380.08	0.000
		19	-0.442	-0.018	410.67	0.000
		20	-0.335	-0.111	428.41	0.000
		21	-0.187	-0.005	433.96	0.000
		22	-0.033	0.030	434.13	0.000
		23	0.080	0.038	435.16	0.000
		24	0.115	-0.043	437.31	0.000
		25	0.052	-0.024	437.77	0.000
		26	-0.082	0.017	438.87	0.000
		27	-0.245	0.032	448.98	0.000
		28	-0.402	-0.081	476.40	0.000
		29	-0.523	-0.107	523.32	0.000
		30	-0.580	-0.125	581.68	0.000
		31	-0.552	-0.027	635.07	0.000
		32	-0.449	-0.059	670.74	0.000
		33	-0.299	-0.016	686.70	0.000
		34	-0.148	-0.091	690.63	0.000
		35	-0.009	0.079	690.64	0.000
		36	0.083	0.064	691.91	0.000



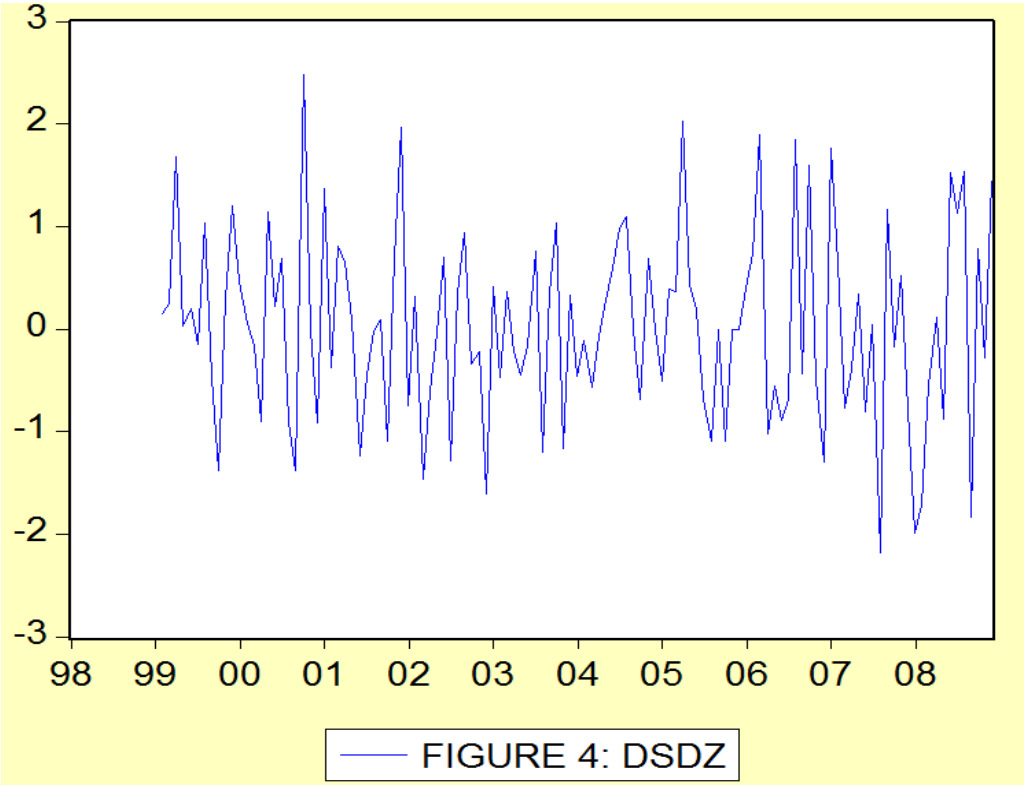
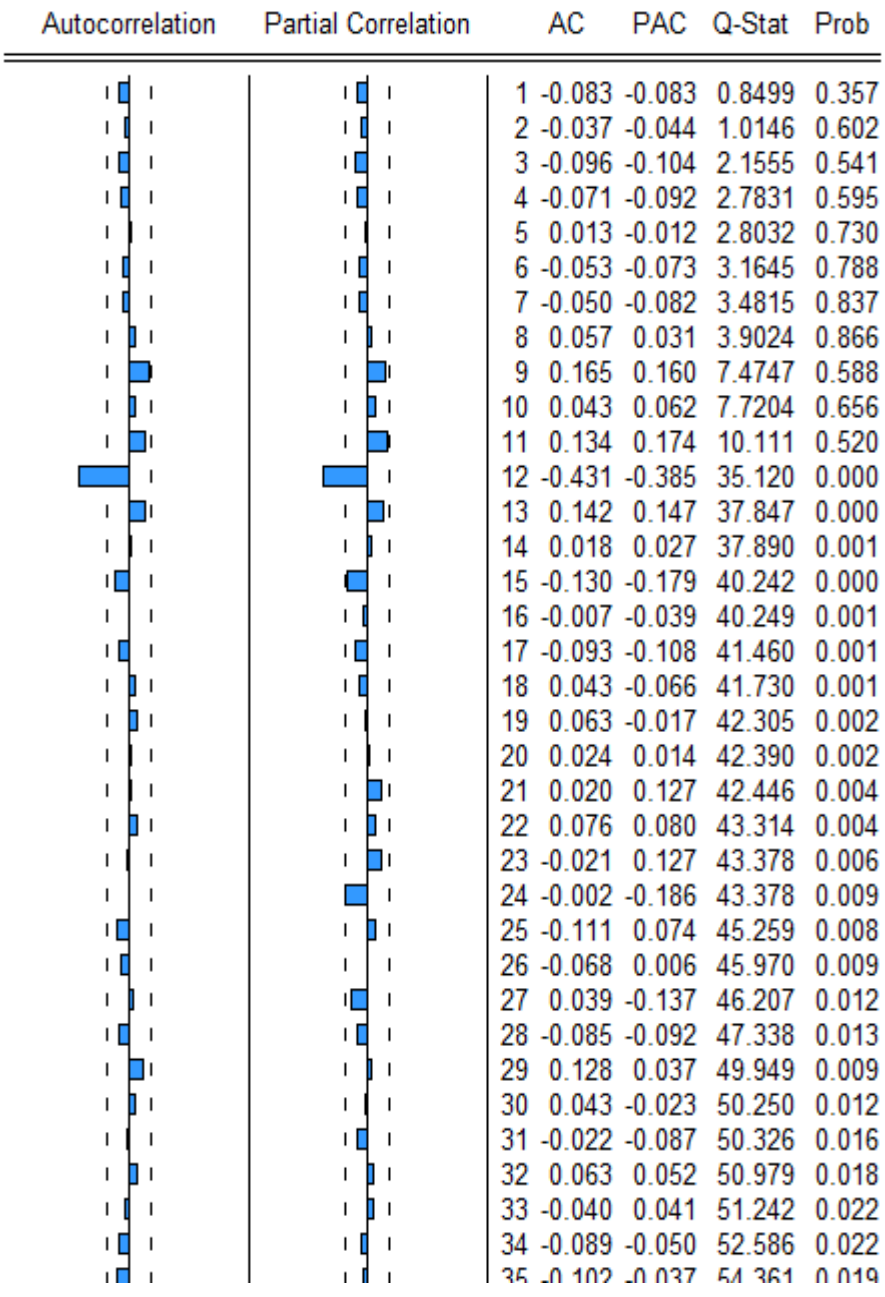


Figure-5. CORRELOGRAM OF DSDZ



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.083	-0.083	0.8499	0.357
		2	-0.037	-0.044	1.0146	0.602
		3	-0.096	-0.104	2.1555	0.541
		4	-0.071	-0.092	2.7831	0.595
		5	0.013	-0.012	2.8032	0.730
		6	-0.053	-0.073	3.1645	0.788
		7	-0.050	-0.082	3.4815	0.837
		8	0.057	0.031	3.9024	0.866
		9	0.165	0.160	7.4747	0.588
		10	0.043	0.062	7.7204	0.656
		11	0.134	0.174	10.111	0.520
		12	-0.431	-0.385	35.120	0.000
		13	0.142	0.147	37.847	0.000
		14	0.018	0.027	37.890	0.001
		15	-0.130	-0.179	40.242	0.000
		16	-0.007	-0.039	40.249	0.001
		17	-0.093	-0.108	41.460	0.001
		18	0.043	-0.066	41.730	0.001
		19	0.063	-0.017	42.305	0.002
		20	0.024	0.014	42.390	0.002
		21	0.020	0.127	42.446	0.004
		22	0.076	0.080	43.314	0.004
		23	-0.021	0.127	43.378	0.006
		24	-0.002	-0.186	43.378	0.009
		25	-0.111	0.074	45.259	0.008
		26	-0.068	0.006	45.970	0.009
		27	0.039	-0.137	46.207	0.012
		28	-0.085	-0.092	47.338	0.013
		29	0.128	0.037	49.949	0.009
		30	0.043	-0.023	50.250	0.012
		31	-0.022	-0.087	50.326	0.016
		32	0.063	0.052	50.979	0.018
		33	-0.040	0.041	51.242	0.022
		34	-0.089	-0.050	52.586	0.022
		35	-0.102	-0.037	54.361	0.019

Table-1. Comparison of the Proposed Models

Model	Status	AIC value	S. E. of regression
$(2,1,2)x(1,1,1)_{12}$	Invertible	2.172557	0.685329
$(2,1,1)x(1,1,1)_{12}$	Invertible	2.165231	0.688742
$(1,1,2)x(1,1,1)_{12}$	Invertible	2.290075	0.733347
$(1,1,1)x(1,1,1)_{12}$	Invertible	2.231494	0.718449
$(2,1,3)x(1,1,1)_{12}$	Noninvertible	2.390756	0.757929
$(1,1,3)x(1,1,1)_{12}$	Invertible	2.407217	0.770956

Table-2. Estimation of the Sarima $(2, 1,2)X(1, 1, 1)_{12}$ Model

Dependent Variable: DSDZ
Method: Least Squares
Date: 03/24/14 Time: 17:52
Sample(adjusted): 2000:04 2008:12
Included observations: 105 after adjusting endpoints
Convergence achieved after 64 iterations
Backcast: 1999:02 2000:03

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.666614	0.125409	5.315520	0.0000
AR(2)	-0.171081	0.175957	-0.972289	0.3334
AR(12)	-0.160365	0.099378	-1.613685	0.1099
AR(13)	0.333005	0.112307	2.965133	0.0038
AR(14)	-0.348619	0.100182	-3.479869	0.0008
MA(1)	-0.595921	0.117437	-5.074371	0.0000
MA(2)	-0.018409	0.180892	-0.101770	0.9192
MA(12)	-0.797828	0.081125	-9.834548	0.0000
MA(13)	0.375634	0.127449	2.947332	0.0040
MA(14)	0.092484	0.169085	0.546964	0.5857

R-squared	0.537243	Mean dependent var	4.55E-05
Adjusted R-squared	0.493403	S.D. dependent var	0.962870
S.E. of regression	0.685329	Akaike info criterion	2.172557
Sum squared resid	44.61921	Schwarz criterion	2.425315
Log likelihood	-104.0593	F-statistic	12.25459
Durbin-Watson stat	2.158195	Prob(F-statistic)	0.000000

Inverted AR Roots	.91 -.19i	.91+.19i	.75+.53i	.75 -.53i
	.48 -.77i	.48+.77i	.11 -.92i	.11+.92i
	-.31 -.88i	-.31+.88i	-.69+.63i	-.69 -.63i
	-.91 -.23i	-.91+.23i		
Inverted MA Roots	.99	.86+.49i	.86 -.49i	.65
	.50+.85i	.50 -.85i	.01 -.98i	.01+.98i
	-.18	-.48 -.85i	-.48+.85i	-.84+.49i
	-.84 -.49i	-.97		

Table-3. Estimation of the Sarima $(2, 1, 1)X(1, 1, 1)_{12}$ Model

Dependent Variable: DSDZ
Method: Least Squares
Date: 03/24/14 Time: 17:58
Sample(adjusted): 2000:04 2008:12
Included observations: 105 after adjusting endpoints
Convergence achieved after 60 iterations
Backcast: 1999:03 2000:03

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AR(1)	0.445020	0.118790	3.746259	0.0003
AR(2)	-0.149584	0.085114	-1.757456	0.0820
AR(12)	-0.146818	0.090093	-1.629635	0.1064
AR(13)	0.160347	0.098164	1.633463	0.1056
AR(14)	-0.201312	0.089071	-2.260118	0.0261
MA(1)	-0.348015	0.114328	-3.044005	0.0030
MA(12)	-0.868345	0.037438	-23.19394	0.0000
MA(13)	0.312118	0.104962	2.973628	0.0037

R-squared	0.522783	Mean dependent var	4.55E-05
Adjusted R-squared	0.488345	S.D. dependent var	0.962870
S.E. of regression	0.688742	Akaike info criterion	2.165231
Sum squared resid	46.01345	Schwarz criterion	2.367438
Log likelihood	-105.6746	F-statistic	15.18026
Durbin-Watson stat	2.170861	Prob(F-statistic)	0.000000

Inverted AR Roots	.88+.21i	.88 -.21i	.69+.55i	.69 -.55i
	.41 -.75i	.41+.75i	.10 -.87i	.10+.87i
	-.30 -.85i	-.30+.85i	-.67+.61i	-.67 -.61i
	-.89+.22i	-.89 -.22i		
Inverted MA Roots	.99	.85 -.49i	.85+.49i	.49 -.86i
	.49+.86i	.36	-.00 -.99i	-.00+.99i
	-.49+.86i	-.49 -.86i	-.86 -.49i	-.86+.49i
	-.99			