Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 28

5-1-2010

On the Appropriate Transformation Technique and Model Selection in Forecasting Economic Time Series: An Application to Botswana GDP Data

D. K. Shangodoyin
University of Botswana, shangodoyink@mopipi.ub.bw

K. Setlhare
University of Botswana, setlharek@mopipi.ub.bw

K. K. Moseki
University of Botswana, mpsekikk@mopipi.ub.bw

K. Sediakgotla
University of Botswana, sediakgotla@mopipi.ub.bw

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 287-295

1538 – 9472/10/$95.00

On the Appropriate Transformation Technique and Model Selection in Forecasting Economic Time Series: An Application to Botswana GDP Data

D. K. Shangodoyin K. Setlhare K. K. Moseki K. Sediakgotla
University of Botswana,
Botswana

Selected data transformation techniques in time series modeling are evaluated using real-life data on Botswana Gross Domestic Product (GDP). The transformation techniques considered were modified, although reasonable estimates of the original with no significant difference at $\alpha = 0.05$ level were obtained: minimizing square of first difference (MFD) and minimizing square of second difference (MSD) provided the best transformation for GDP, whereas the Goldstein and Khan (GKM) method had a deficiency of losing data points. The Box-Jenkins procedure was adapted to fit suitable ARIMA (p, d, q) models to both the original and transformed series, with AIC and SIC as model order criteria. ARIMA (3, 1, 0) and ARIMA (1, 0, 0) were identified, respectively, to the original and log of the transformed series. All estimates of the fitted stationary series were significant and provided a reliable forecast.

Key words: Data transformation technique, autoregressive integrated moving average, model order criteria, forecast, gross domestic product.

Introduction

The foremost difficulty with economic research in developing countries is the dearth of data. Much of the available economic time series data are constructed out of bits and pieces that must be shaped and arranged to yield a final series that is useable for model building. One way to circumvent this problem is to estimate some components for dates for which time series is not readily available from known values of that component for other dates For example, the US real Gross Domestic Product (GDP) and German real GDP are produced and publicly released at quarterly intervals, although both US and German economic analysts and business-decision makers often need monthly GDP forecasts (Stum & Wollmershauser, 2005), quarterly figures may be required only when the series of annual data are available. This problem has led to several transformations of the data to the form required by researchers for particular research objectives. Economists use many transformations of time series data to help extract economically relevant information (Cohen, 2001).

A facet of the research conducted focuses on the interpolation of some values of a series at a given time period by a related series (Friedman, 1962). The problem with this technique is that it assumes that a related series - as well as some values of the series to be interpolated - are readily available: this may not be the case in developing countries. Various studies have been concerned with the derivation of quarterly figures from annual data, including Lisman and Sandee (1964), Boot, et al. (1967) and Goldstein & Khan (1976); in each of these examples the value of a quarterly figure for each year t, is considered as a weighted average of the totals of the years. A system of equations is built from which weighted coefficients were calculated subject to some criteria.

D. K. Shangodoyin is Associate Professor at the Department of Statistics. Email: shangodoyink@mopipi.ub.bw. K. Setlhare is Senior Lecturer at the Department of Statistics. Email: setlharek@mopipi.ub.bw. K. K. Moseki is a Lecturer at the Department of Statistics. Email: mpsekikk@mopipi.ub.bw. K. Sediakgotla is a Lecturer at the Department of Statistics. Email: sediakgotla@mopipi.ub.bw.

The challenge, therefore, is to explore the efficiency of the transformation techniques and analyze their prediction potential. Some transformation techniques can be found in Boot, et al. (1967) which introduced two methods that involve minimizing the squared first differences (MSFD) and minimizing the squared second differences (MSSD). Goldstein and Khan (1976) proposed an interpolation technique based on the quadratic function: the transformed data could be modeled appropriately by checking the order of the fitted model using model order selection criteria as discussed by Shibata (1976).

In this article, the focus is to evaluate the efficacy of data transformation techniques with the aim of using two known models' order determination criteria to produce the best model order-transformation technique for forecasting economic time series with application to Botswana GDP data. This is considered a challenge to analysts in view of the dearth of quarterly economic series data in some sectors of a nation's economy where only annual data is available.

Methodology

The Technique and Model Determination

Boot, et al. (1967) considered two procedures for the interpolation of quarterly figures given only annual data; the basis of their research is the work of Lisman & Sandel (1964). The first approach is based on the criterion that minimizes the sum of square of the first difference (MFD) between the successive quarterly values, which are subject to the constraint that, each year, the sum of the quarterly total should equal the yearly totals.

Mathematically, consider n years for which it is necessary to minimize $\sum_{i=2}^{4n}(x_i - x_{i-1})^2$ subject to $\sum_{i=4t-3}^{4k} x_i = y_t$ for $t = 1, 2, ...n$ where x_i is the i^{th} quarterly total and y_t is the given yearly total in year t. The problem can be solved by using the Langrangean expression:

$$\sum_{i=2}^{4n}(x_i - x_{i-1})^2 - \sum_{k=1}^{n}\lambda_t\left(\sum_{i=4t-3}^{4t} x_i - y_t\right).$$

The MFD derived formula for calculating the estimated quarterly total within three successive years is given as:

$$\begin{pmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \end{pmatrix} = \frac{1}{1836}\begin{pmatrix} \mathbf{X_{11}} \\ \cdots\cdots \\ \mathbf{X_{12}} \\ \cdots\cdots \\ \mathbf{X_{13}} \end{pmatrix}\begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_t \\ \mathbf{y}_{t+1} \end{pmatrix} \qquad \text{(i)}$$

where x_{ki} is the estimated quarterly value in years $k = 1, 2, 3$ and quarters $i = 1, 2, 3, 4$, $t = 1, 2, 3, 4$ and $\mathbf{y}_{t-1}, \mathbf{y}_t, y_{t+1}$ are the totals for the three successive years, and

$$\mathbf{X_{11}} = \begin{pmatrix} 569 & -135 & 25 \\ 525 & -87 & 15 \\ 437 & 27 & -5 \\ 305 & 189 & -35 \end{pmatrix},$$

$$\mathbf{X_{12}} = \begin{pmatrix} 129 & 405 & -75 \\ 7 & 513 & -61 \\ -61 & 513 & 7 \\ -75 & 405 & 129 \end{pmatrix},$$

$$\mathbf{X_{13}} = \begin{pmatrix} -35 & 189 & 305 \\ -5 & 27 & 437 \\ 15 & -187 & 525 \\ 25 & -135 & 569 \end{pmatrix}.$$

The second approach is the minimization of the sum of square of the second difference (MSD) in which $\sum_{i=2}^{4n}(\Delta x_i - \Delta x_{i-1})^2$, is minimized, where $\Delta x_i = x_{i+1} - x_i$, is subject to the constraint $\sum_{i=4t-3}^{4t} x_i = y_t \quad t = 1, 2, ...n,$ $t = 1, 2, ...n$, and the $x_i's$ are as defined above. Similar to the MSFD, the problem is solved by considering the Lagrangean expression

$\sum_{i=2}^{4n}(\Delta x_i - \Delta x_{i-1})^2 - \sum_{t=1}^{n}\lambda_k(\sum_{i=4t-3}^{4t} x_i - y_t)$, which - when solved routinely for $n=3$ - was shown to give the solution:

$$\begin{pmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \end{pmatrix} = \frac{1}{9280}\begin{pmatrix} \mathbf{X_{21}} \\ \cdots\cdots \\ \mathbf{X_{22}} \\ \cdots\cdots \\ \mathbf{X_{23}} \end{pmatrix}\begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_t \\ \mathbf{y}_{t+1} \end{pmatrix} \qquad \text{(ii)}$$

where the x_{ki} and the $y's$ are defined as previously and

$$\mathbf{X_{21}} = \begin{pmatrix} 3499 & -1488 & 309 \\ 2697 & -464 & 87 \\ 1911 & 528 & -119 \\ 1173 & 1424 & -227 \end{pmatrix},$$

$$\mathbf{X_{22}} = \begin{pmatrix} 531 & 2128 & -338 \\ 49 & 2512 & -241 \\ -241 & 2512 & 49 \\ -339 & 2128 & 531 \end{pmatrix},$$

$$\mathbf{X_{23}} = \begin{pmatrix} -277 & 1424 & 1173 \\ -119 & 528 & 1911 \\ 87 & -464 & 2697 \\ 309 & -1488 & 3499 \end{pmatrix}.$$

Goldstein and Khan (1976) (GKM) proposed an interpolation technique for converting annual totals to quarterly series by using the quadratic functions passing through three successive points y_{t-1}, y_t and y_t the expressions for these interpolations are:

$$\begin{aligned} &0.0548y_{t-1} + 0.2343y_t - 0.0390y_{t+1} \\ &0.0077y_{t-1} + 0.2657y_t - 0.0235y_{t+1} \\ &0.0100y_{t-1} + 0.2500y_t - 0.01500y_{t+1} \\ &0.0400y_{t-1} + 0.2400y_t - 0.0110y_{t+1} \end{aligned} \qquad \text{(iii)}$$

In the expressions, the first year will have $y_{t-1} = y_t = 0$ and the second year will have $y_{t-1} = 0$ in the computation of the quarterly total for the years, assuming y_{t-1}, y_t and y_{t+1} are independent aggregates. Lisman and Sandel (1964) assumed that the quarterly data, for example, $\mathbf{Z_J}$, was linearly dependent on three successive annual totals and proposed the computation of quarterly data from the following:

$$\begin{pmatrix} 0.0729 & 0.1982 & -0.0211 \\ -0.0103 & 0.3018 & -0.0415 \\ -0.0415 & 0.3018 & -0.0103 \\ -0.0211 & 0.1982 & 0.0729 \end{pmatrix}\begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_t \\ \mathbf{y}_{t+1} \end{pmatrix} \qquad \text{(iv)}$$

All of these methods are known to have limitations (Boot, et al., 1967), thus other mathematical methods of interpolation have been developed by researchers such as Glejer (1966), Boots and Feibes (1967) and Vangrevelinghe (1966). The choice of method as described in (i)-(iv) is based on the similarity in their computation. It would be of tremendous assistance to analysts if the various methods are subjected to real-life data experimentation, while the transformed data are modeled with an appropriate check on the models order to ascertain their suitability in forecasting.

In this article it is assumed that the **y's** are moving by 3 points, models are run up to **n-2**, and the identified (or fitted) model is used to compute **n-1** and **n** so that no year is omitted and the model provides a reasonable degree appropriateness for the transformed data. The Box-Jenkins modeling was performed on both the original and transformed data with a view to forecast. However, the unknown value of the model order, **P**, may constitute a casualty in modeling as attempts to under fit increases the residual variance, while over fitting results in too many parameters which eventually causes unreliability (Jones, 1975; Shibata, 1976). Various selection criteria have been advanced for model order selection (Box, Jenkins & Reinsel, 1994), in this article, three similar

criteria were employed vis-à-vis the Akaike information criteria (AIC) $\{\mathbf{N}\ln\sigma_{\mathbf{p}}^2 + 2p\}$, final predictor error (FPE) $\{\left(\frac{N+p}{N-p}\right)\sigma_p^2\}$ and Schwarz's criterion (SIC) $\{\mathbf{N}\ln\sigma_{\mathbf{p}}^2 + \mathbf{P}\ln\mathbf{N}\}$. The order in which two of these criteria agree shall be considered to be the best order for the data.

Results

Data Analysis: Transformation and Modeling of Botswana GDP Data

Data presented in Appendix I shows that no significant variation exists between the average values of data computed by the three techniques and the original data. The test of difference conducted between the original series and the transformed series indicates that there is no significant difference between the means of the GDP, MFD, MSD and GKM. It was observed (see Appendix II), that the MFD and MSD provided the best transformation for the Botswana GDP data while the GKM had a deficiency of losing data points. The proposed method of moving point incorporated into the selected techniques is shown to be worthwhile because neither the MFD nor the MSD lost any data.

Model Selection and Order Determination

The original GDP series is made stationary by taking the first difference (see Appendix II) - an autoregressive process of order 3 is identified as the most suitable model. Based on AIC and SIC criteria, the fitted values (Appendix II) are adequate as indicated in Figure 2 and the bounds placed on the fitted values appear to have accommodated the original values adequately.

The MFD, MSD and GKM series became stationary only when the log-transformation was taken, the AIC, SIC and model RESIDUALS were the criteria used in selecting the best order for the model and these identified the AR (1) models to MFD, MSD and GKM. The behavior of the fitted values (see Appendix III, Figures 1-4 and Tables 1--4) indicate the appropriateness of the model as confirmed by the Portmanteaux test for model adequacy.

Conclusion

The moving point method introduced into the transformation techniques utilized in this research has shown a tremendous improvement over the MFD and MSD. It was observed that both MFD and MSD give nearly the same fitted values as the original series; thus confirming the findings of Shangodoyin and Adubi (2000) who used Nigeria GDP data. The choice of the model order should not, however, be limited to the order determination criteria but also to the model residual variance.

References

Boot, J. C. G., et al. (1967). Further method of derivation of quarterly figures from annual data. *J.R.S.S. Series C*, *16*(*1*), Vo.16. 65-75.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control*, (*3rd Ed.*). New Jersey: Prentice Hall.

Cohen, D. (2001). *Linear transformation used in economics*. Washington DC: Federal Reserve Board.

Friedman, M. (1962). The interpolation of time series by related series. *J.A.S.A.*, *57*(*300*), 729-757.

Glejser, H. (1966). Une methode d'evaluation de donnees menuelles a partir d'indices trimestriels ou annuels. *Cahiers Economiques de Bruxelles*, *19*(*1*), 45.

Goldstein, & Khan. (1976). An analysis of transformation revisited. *J.A.S.A,.76*, 296-311.

Jones, R. H. (1975). Fitting autoregressions. *J.A.S.A.*, 70, 590-592.

Lisman, J. H., & Sandee, J. (1964). Derivation of quarterly figures from annual data. *Applied Statistics*, *13*, 87-90.

Shangodoyin, D. K., & Adubi, A. A. (2000). *Appropriate data transformation techniques in forecasting economic series: A case study of Nigeria GDP*. Nigerian Statistical Association Conference, Floria Nigeria.

Shibata, R. C. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*, 117-126.

Stum, J-E., & Wollmershauser, T. (2005). *IFO survey data in business cycle and monetary policy analysis*. Springer International handbook of education, Volume. 13.

Vangrevelinghe, G. (1966). Levolution a court terme de la consummation des ménages. *Etudes et Conjoncture* (*INSEE*), *9*, 54-102.

Appendix I: One-Way Analysis of Variance

The results indicate that no significant difference exists between the means of the four series (original, FMD, MSD, and GKM), at $\alpha = 0.05$.

Analysis of Variance

Source	DF	SS	MS	F	P
Factor	3	497763	165921	0.06	0.981
Error	132	372383781	2821089		
Total	135	372881545			

Individual 95% CIs for Mean Based on Pooled StDev

Level	N	Mean	StDev	---+---------+---------+---------+---
ORIGINAL	36	5068	1765	(----------*----------)
MFD	36	5068	1742	(----------*----------)
MSD	36	5068	1745	(----------*----------)
GKM	28	4918	1368	(----------*----------)
				---+---------+---------+---------+--

Pooled StDev = 1680 4400 4800 5200 5600

Appendix II: Original and Transformed Series with Forecast Values

Actual GDP	MFD	MSD	GKM	Fitted GDP Values	Fitted MFD Values	Fitted MSD Values	Fitted GKM Values
2918.5	2713.68	2669.984	2924.033	NA	NA	NA	NA
2633	2732.35	2728.983	3009.161	NA	2812.433	2766.022	3049.796
2667.8	2769.68	2789.227	2962.78	NA	2913.945	2864.974	3178.91
2822	2825.68	2853.205	3228.221	NA	3018.262	2966.907	3311.378
2964.8	2900.35	2924.651	3308.047	3310.34	3125.433	3071.894	3447.203
2959.3	2996.24	3008.543	3451.497	3004.924	3235.503	3180.005	3586.381
3074.2	3113.37	3107.371	3407.489	2939.061	3348.52	3291.311	3728.91
3263.4	3251.74	3221.135	3704.262	3311.223	3464.531	3405.888	3874.781
3469.9	3411.33	3347.346	4148.561	3598.809	3583.582	3523.809	4023.983
3330.3	3531.03	3481.024	4349.12	3559.041	3705.721	3645.149	4176.504
3325.3	3610.82	3618.437	4274.65	3369.796	3830.992	3769.985	4332.328
4078.4	3650.72	3757.094	4604.015	3554.994	3959.443	3898.395	4491.436
4540.1	4308.45	4172.611	4856.836	4346.209	4091.118	4030.458	4653.806
4280.4	4359.09	4349.4	4987.995	4649.145	4226.064	4166.252	4819.415
4385	4460.37	4524.359	4895.197	4367.403	4364.324	4305.858	4988.234
4534.7	4612.29	4693.829	5311.871	4600.11	4505.945	4449.359	5160.236
4894.5	4814.85	4852.319	5175.132	4762.759	4650.971	4596.837	5335.387
5107.8	4986.19	4992.512	5287.936	5018.153	4799.445	4748.376	5513.653
4861.9	5126.32	5110.746	5208.404	5225.658	4951.41	4904.06	5694.997
5298.3	5235.24	5207.023	5697.818	5064.246	5106.911	5063.975	5879.38
5614.1	5312.94	5285.001	5906.844	5501.775	5265.99	5228.209	6066.761
5937.3	5371.21	5352	6120.156	5749.231	5428.688	5396.848	6257.095
4578.2	5410.06	5413.51	6021.465	6074.464	5595.048	5569.981	6450.337
5394.1	5429.49	5473.19	6532.291	4770.482	5765.11	5747.699	6646.438
6144.7	6059.24	5875.95	6831.44	5582.387	5938.915	5930.091	6845.349
6444.7	6129.85	6116.536	7050.604	6288.624	6116.503	6117.25	7047.018
5856.1	6271.08	6356.419	6929.72	6593.752	6297.914	6309.268	7251.391
6497.6	6482.93	6594.195	7519.338	6040.774	6483.185	6506.238	7458.412
7144.8	6765.39	6827.755	NA	6676.607	6672.355	6708.256	7668.025
7009.6	7035.87	7054.289	NA	7295.269	6865.461	6915.416	NA
6906.9	7294.36	7272.392	NA	7165.869	7062.54	7127.815	NA
7575.2	7540.87	7482.064	NA	7085.977	7263.628	7345.549	NA
7794.8	7775.39	7684.71	NA	7748.615	7468.76	7568.718	NA
7269.2	7951.28	7883.141	NA	7950.03	7677.97	7797.419	NA
7924.1	8068.55	8079.464	NA	7429.783	7891.292	8031.753	NA
8934.4	8127.18	8275.084	NA	8099.146	8108.759	8271.82	NA

Appendix III: Graphs and Tables Results

Table 1: GDP at Constant 1993/94 Prices in P'000 000

Stationary	First Difference		
Identified Model		ARIMA(p,1,0)	
Order of Model	1	2	3
c	179.8878(68.48002)	165.8992(33.4594)	166.0609(21.4485)
a1	-0.1669(0.1812)	-0.3379(0.1551)	-0.6071(0.1753)
a2	na	-0.6494(0.1560)	-0.8321(0.1583)
a3	na	na	-0.4559(0.1802)
AIC	15.1775	14.808	14.6707
SIC	15.2673	14.944	14.8539
Residual Var	6909122	4346325	3431744
Best Model: $ARIMA(3,1,0)$	$D(GDP) = 166.0609 - 0.6071X_{t-1} - 0.8321X_{t-2} - 0.45589X_{t-3}$		
Forecasting Model	$\widehat{X}_{t+m=} = 166.0609 + GDP(-1) - 0.6071GDP(-1) - 0.8321GPD(-2) - 0.45589GDP(-3)$		

Table 2: Table 2: Results of Fitted Model on MFD Series

	MFD		
Stationary	Logarithm Transformation		
Identified Model		ARIMA (p, 0, 0)	
Order of Model	1	2	3
c	12.3638(7.2757)	11.6372(4.2104)	11.1797(2.8020)
a1	0.992(0.015)	0.8208(0.1756)	0.7673(0.1828)
a2	na	0.1674(0.1753)	0.0979(0.2312)
a3	na	na	0.11859(0.1811)
AIC	-4.098	-4.066	-4.02
SIC	-4.009	-3.932	-3.839
Residual Var	0.03066	0.028598	0.027209
Best Model: $ARIMA(1,1,0)$	$Log(MFD) = 12.3636 + 0.992MFD(-1)$		
Forecasting Model	$\widehat{X}_{t+m} = \exp(12.3636 + 0.992MFD(-1))$		

Appendix III: Graphs and Tables Results (continued)

Figure 1: Forecast and MFD Values

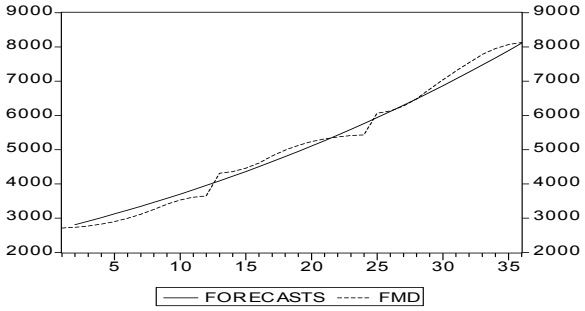


Figure 2: Forecast And GDP Values

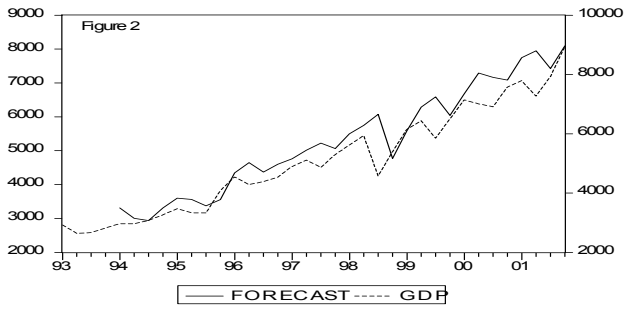


Table 3: Results of Fitted Model on MSD Series

	MSD		
Stationary	Logarithm Transformation		
Identified Model		ARIMA(p,0,0)	
Order of Model	1	2	3
c	14.502(9.4691)	11.844(4.2843)	10.7887(2.4542)
a1	0.9947(0.0084)	1.2938(0.1696)	1.2326(0.1822)
a2	na	-0.3006(0.1690)	-0.0995(0.2917)
a3	na	na	-0.1419(0.1803)
AIC	-5.252	-5.2796	-5.222
SIC	-5.163	-5.1449	-5.0412
Residual Variance	0.009471	0.08501	0.008718
Best Model	$Log(SMD) = 14.502 + 0.9947SMD(-1))$		
Forecasting Model	$\widehat{X}_{t+m} = \exp(14.502 + 0.9947SMD(-1))$		

Figure 3: Forecast and MSD Values

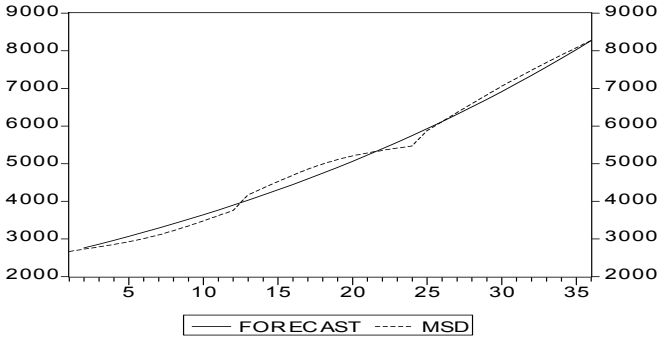
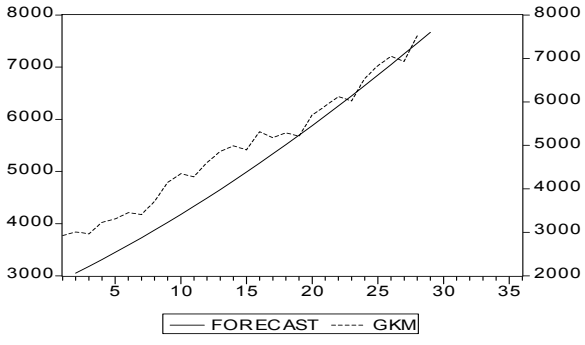


Figure 4: Forecast and Values Of GKM



Appendix III: Graphs and Tables Results (continued)

Table 4: Results of Fitted Model on GKM Series

	GKM		
Stationary	Logarithm Transformation		
Identified Model		ARIMA(p,0,0)	
Order of Model	1	2	3
c	10.7201(4.4355)	10.7123(3.4997)	9.9587(1.2735)
a1	0.9846(0.0298)	0.6679(0.2054)	0.5865(0.2058)
a2	na	0.3118(0.2022)	0.1671(0.2498)
a3	na	na	0.2078(0.2084)
AIC	-3.4161	-3.399	-3.433
SIC	-3.3201	-3.2537	-3.2381
Residual Variance	0.044768	0.040382	0.034315
Best Model	$Log(GKM) = 10.7201 + 0.9846GKM(-1))$		
	$\widehat{X}_{t+m} = \exp(10.7201 + 0.9846GKM(-1))$		