Science Journal of Applied Mathematics and Statistics
2015; 3(3): 124-135
Published online May 16, 2015 (http://www.sciencepublishinggroup.com/j/sjams)
doi: 10.11648/j.sjams.20150303.15
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Application of Vector Autoregressive (VAR) Process in Modelling Reshaped Seasonal Univariate Time Series

Chepngetich Mercy[1], John Kihoro[2]

[1]Jomo Kenyatta University of Agriculture and Technology, School of Mathematical Sciences, Nairobi, Kenya
[2]Cooperative University College of Kenya, Department of Computing and e-learning, Nairobi, Kenya

Email address:
chepngetichmercy88@gmail.com (C. Mercy), Kihoro.jm@gmail.com (J. Kihoro)

To cite this article:
Chepngetich Mercy, John Kihoro. Application of Vector Autoregressive (VAR) Process in Modelling Reshaped Seasonal Univariate Time Series. *Science Journal of Applied Mathematics and Statistics.* Vol. 3, No. 3, 2015, pp. 124-135. doi: 10.11648/j.sjams.20150303.15

Abstract: Seasonal Autoregressive Integrated Moving Averages (SARIMA) model has been applied in most research work to forecast seasonal univariate data. Less has been done on Vector Autoregressive (VAR) process. In this research project, seasonal univariate time series data has been used to estimate a VAR model for a reshaped seasonal univariate time series for forecasting. This was done by modeling a reshaped seasonal univariate time series data using VAR. The quarterly data is reshaped to vector form and analyzed to vector form and analyzed using VAR for the year 1959 and 1997 to fit the model and the prediction for the year 1998 is used to evaluate the prediction performance. The performance measures used include; mean square error (MSE), root mean square error (RMSE), mean percentage error (MPE), mean absolute percentage error (MAPE) and Theil's U statistic. Forecasting future values from the fitted models in both SARIMA and VAR using Box Jenkins procedures, (Box and Jenkins; 1976) was done. The results showed that both models are appropriate in forecasting but VAR is more appropriate model than SARIMA model since its predictive performance was shown to be the best. Other data sets were also used for analysis and comparison purpose.

Keywords: Vector Autoregressive Process, Seasonal Autoregressive Integrated Moving Average Process, Vector Error Correction Model, Akaike Information Criterion

1. Introduction

1.1. Background of the Study

The goal of a time series analysis is to obtain the parameters of the underlying physical process that governs the observed time series and used it to forecast future values. The modelling and predictions is meant to determine the best model for forecasting. This is useful to many areas of study such as meteorological forecasting, electricity consumption, demographic forecasting and Economic growth. In these forecasts, both short term and long term seasonal trends can be forecasted. To model these variations for forecasting, usually a seasonal ARIMA model of Box and Jenkins (1976) is used. ARIMA time series is a popular models used for studying weather and market data

Seasonal ARIMA is a model of ARIMA class suitable for forecasting seasonal data variations and trends. SARIMA model exhibit some disadvantages that; constructing it is expensive. Also, adapting it for use requires expertise and consumes a lot of time. Most of the univariate data are seasonal. Hence, fitting the model that gives the best performance prediction helps a lot in future predictions and analysis. This will be of great importance to economic and planning purposes. For seasonal univariate time series data, SARIMA model is usually used to do forecasting, with its advantages and limitations. In this project SARIMA model is reshaped and forecast using Vector Autoregressive (VAR) model.

This Project applied VAR in the reshaped seasonal univariate time series data, test and distinguish between the two models in terms of its predictive performance results. The results show that although both models are appropriate in forecasting, VAR model gives the best performance compared to the SARIMA model.

1.2. Literature Review

Sims (1980) introduced VAR as a technique that could be used by microeconomics to characterize the mutual dynamic behavior of collection of variables without requiring strong restrictions of the kind needed to identify underlying structural parameters. Sims developed a VAR model with p lags, VAR (p)

for expressing a set of variables as a weighted linear combination of each variable's past values and the past values of the other variables in the set. However, multivariate data analysis in the context of VAR has evolved as standard instrument in econometrics.

The VAR (p) models are more flexible than AR models. Mei, Liu and Jing (2011) constructed a multi-factor dynamic system VAR forecast model of GDP by using six important economic indicators. This included the fiscal revenue, social retail goods, secondary industry output, and investment in fixed assets, employment rate and tertiary industry output which based on data from the shanghai region in china. In their analysis, the model showed high significance and the results show that the relative error forecast is quite small. The authors therefore conclude that the VAR model has a considerable practical value.

Clarida and Friedman (1984) use a VAR model to forecast the United States short-term interest rates during April 1979 to February 1983. A constant-coefficient, linear VAR model is generated to estimate the pre-October 1979 probability structure of the quarterly data , which takes six important united states macroeconomic factors into consideration. The result shows that short-term interest rates in the United States have been too high since October 1979. Based on their VAR model, the prediction results of conditional and unconditional forecast are both lower than the actual United States short-term interest rates during this period.

The VAR model extends univariate autoregressive models to dynamic multivariate and provides better forecasts than univariate time series models (Zivot and Wang, 2006). VAR models are used to describe and forecast multivariate time series for stationary time series. For non-stationary time series a vector error correction term is added to form a vector error correction model (VECM) and it is necessary to test for the existence of a stationary linear combination of the non-stationary terms (co integration). It must be transformed into vector error correction model (VECM) by taking the first difference.

Autoregressive integrated moving-average (ARIMA) models are simple time series models that can be used to fit and forecast univariate data such as fisheries landings. With ARIMA models, data are assumed to be the output of a stochastic process, generated by unknown causes, from which future values can be predicted as a linear combination of past observations and estimates of current and past random shocks to the system (box et al., 2008).

Xiao Han Cai, (2008), uses both VAR model and SARIMA model to analyze time series data of air pollution co in California south coast area. Their results showed that VAR model is a better model to forecast multiple variables data sets though not easy to find the order to fit an accurate VAR model. On the other hand, SARIMA model presents how the current month air pollutant concentrations depend upon the previous month's air pollutant concentrations.

1.3. Statement of the Problem

Modeling a seasonal univariate data, usually a SARIMA model is applied. This is because of the seasonal variations exhibited. This seasonal variation could either be re-shaped into vector form, making it a multivariate time series then forecasted using VAR

VAR has been seen as a more advanced most successful, flexible and easy to use model for forecasting (Zivot and Wang, 2006). This calls for a research on this to examine if VAR can be a better tool to use in the re-shaped time series than the SARIMA model for forecasting a seasonal univariate time series.

2. Methodology

In this study we will first examine the appropriateness of VAR model in forecasting a reshaped seasonal univariate time series by predicting one step ahead vector observation. Then compare the forecast with those obtained with univariate SARIMA model.

2.1. Vector Autoregressive Process

In order to build a VAR model, (box and Jenkins; 1976) steps can be followed. This includes model identification, estimation of constants, diagnostic check and finally forecasting. Conditional heteroscedasticity and outliers of the residual series is also checked. The existence of co integration is used to check the presence of any common trends. Error Correction Model (ECM) can be developed in case of co integration. This improves the forecasting of long term.

2.1.1. Model Specification and Identification

Just like any other univariate time series, the first step in building a VAR (p) model involves model identification. This helps in identifying the order of the appropriate model. The most common information criterion used to identify the model include Akaike information criterion (AIC) (box and Jenkins; 1976), Schwarz-Baysian (BIC) and Hannan-quinn (HQC).

The time series y_t , where $y_t=(y_{1t},y_{2t},\ldots,y_{nt})$ denote an (n×1) vector variables, follows a VAR(p) model if it satisfies

$$Y_t = \pi + \Phi_1 y_{t-1} + \ldots + \Phi_{t-p} y_{t-p} + e_t \,, \; t=1,2,3,\ldots t, \qquad (1)$$

In matrix form, this can be expressed as;

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{bmatrix} = \begin{bmatrix} \Pi 1 \\ \Pi 2 \\ \vdots \\ \Pi n \end{bmatrix} + \begin{bmatrix} \Phi_{11}^1 & \Phi_{12}^1 & \ldots & \Phi_{1n}^1 \\ \Phi_{21}^1 & \Phi_{21}^1 & \ldots & \Phi_{2n}^1 \\ \vdots & \vdots & \ldots & \vdots \\ \Phi_{n1}^1 & \Phi_{n2}^1 & \ldots & \Phi_{nn}^1 \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ \vdots \\ y_{nt-1} \end{bmatrix} + \begin{bmatrix} \Phi_{11}^2 & \Phi_{12}^2 & \ldots & \Phi_{1n}^2 \\ \Phi_{21}^2 & \Phi_{21}^2 & \ldots & \Phi_{2n}^2 \\ \vdots & \vdots & \ldots & \vdots \\ \Phi_{n1}^2 & \Phi_{n2}^2 & \ldots & \Phi_{nn}^2 \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ \vdots \\ y_{nt-1} \end{bmatrix} + \ldots + \begin{bmatrix} \Phi_{11}^p & \Phi_{12}^p & \ldots & \Phi_{1n}^p \\ \Phi_{21}^2 & \Phi_{21}^p & \ldots & \Phi_{2n}^p \\ \vdots & \vdots & \ldots & \vdots \\ \Phi_{n1}^p & \Phi_{n2}^p & \ldots & \Phi_{nn}^p \end{bmatrix} \begin{bmatrix} y_{1t-p} \\ y_{2t-p} \\ \vdots \\ y_{nt-p} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ \vdots \\ e_{nt} \end{bmatrix}$$

Where π is a k-dimensional vector, $\Phi_1, \Phi_2, \ldots, \Phi_p$ are k×k parameter matrices and e_t is a sequence of independently and identically distributed error vectors.

Assumptions of the errors

The error term e_t is a multivariate normal $k \times 1$ vector of error satisfying the following assumptions;

1. $E(e_t)=0$ every error term has mean zero;
2. $E(e_t e'_t)=s$ the contemporaneous covariance matrix of error terms is ω (a $n \times n$ positive-semi-definite matrix);
3. $E(e_t e'_{t-k})=0$ for any non-zero k. There is no correlation across time; in particular, no serial correlation in individual error terms.

and Φ_j are $k \times k$ matrices.

Using the back-shift operator b, the VAR (p) model can be written as

$$(I - \Phi_1 b - \ldots - \Phi_p b^p)\, Y_t = \pi + e_t,$$

Where I is the $k \times k$ identity matrix. In a compact form as follows

$$\Phi(b)\, Y_t = \pi + e_t,$$

Where $\Phi(b) = I - \Phi_1 b - \ldots \Phi_p b^p$ is a matrix polynomial.

Stability in VAR (p) process is one important characteristic. It generates stationary time series, where the means, variances and covariance are time invariant. Reverse characteristic polynomial can be evaluated to check its stability.

$$\det(I - \Phi_1 b - \ldots - \Phi_p b^p) \neq 0,\ b \leq 1.$$

If $b=1$,in the solution, then some variables are integrated

If Y_t is weakly stationary, then we have

$$\mu = E(Y_t) = (I - \Phi_1 b - \ldots - \Phi_p b^p)^{-1} \pi = [\Phi(1)]$$

Provided that the determinant exists since determinant of $[\Phi(1)]$ is different from zero. Expressing VAR (p) in deviation form from its mean, we define $\mu = E(Y_t)$. Then in deviation form is given by;

$$Y_t - \mu = \Phi_1 (Y_{t-1} - \mu) + \Phi_2 (Y_{t-2} - \mu) + \cdots + \Phi_p (Y_{t-p} - \mu)$$

$\tilde{Y}_t = Y_t - \mu$, then the VAR(p) model becomes

$$\tilde{Y}_t = \Phi_1 \tilde{Y}_{t-1} + \ldots + \Phi_p \tilde{Y}_{t-p} + e_t \quad (2)$$

Given, $\hat{e}_t^{(i)} = Y_t - \hat{\Phi}_i^{(i)} Y_{t-1} - \cdots - \hat{\Phi}_i^{(i)} Y_{t-i}$ as the residual of the model, the residual covariance matrix is defined as;

$$\hat{S}(p) = \frac{1}{T} \sum_{t=1}^{T} \hat{e}_t (\hat{e}_t),$$

The AIC of a VAR (i) model under the normality assumption is defined

$$AIC(p) = ln\left(|\hat{S}(p)|\right) + \frac{2n^2 p}{T}$$

We will select the VAR of order p such that the calculated value is minimum, that is $\text{AIC}(p) = \text{minimum } 1 \leq p \geq \text{AIC}(p)$, where p is an integer for a given vector time series. Other multivariate information criterion measures include;

$$BIC(p) = ln\left(|\hat{S}(p)|\right) + \frac{lnT}{T} pn^2$$

$$HQC(p) = ln\left(|\hat{S}(p)|\right) + \frac{2\ln(\ln(T))}{T} pn^2$$

$$FPE(p) = \left(\frac{T + p^*}{T - p^*}\right)^n |\hat{S}(p)|$$

Where p^* is the total number of parameters in each equation, and the VAR model of order p lags such that the criterion information is minimum, is selected.

2.1.2. Estimation of Constants

After obtaining the order of the model, p, of the vector series, we now derives the estimators of the constants.

Consider the consecutive VAR models:

$$Y_t = \pi + \Phi_1 y_{t-1} + e_t$$

$$Y_t = \pi + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + e_t$$

$$\ldots = \ldots$$

$$Y_t = \pi + \Phi_1 y_{t-1} + \ldots + \Phi_i y_{t-i} + e_t \quad (3)$$

The most common methods of estimating parameters are the maximum likelihood estimator (MLE) and the ordinary least square estimator (yang & yuan 1991). Here, we will apply the ordinary least squares (OLS) method to estimate the parameters of these models and applied equation by equation.

For i^{th} equation (3), let $\hat{\Phi}_i^{(j)}$ be the OLS estimate of Φ_i and $\hat{\Pi}_i^{(j)}$ be the OLS estimate of π. Where (i) is used to denote the estimate of a VAR (i) model. The estimates of the coefficients of the VAR model, $\Phi_1 \tilde{Y}_{t-1} + \ldots + \Phi_p \tilde{Y}_{t-p} + e_t$, when estimated using unrestricted VAR (p) model are considered to be fixed quantities. These estimates of coefficients do not accurately reflect the underlying relationship because some of the estimated coefficients of the VAR model are non-zero purely by chance when estimated by OLS so restrictions may be imposed to reduce the number of parameters being estimated.

2.1.3. Diagnostic Check

Suppose the orders and constants have been chosen for a VAR model underlying the data, then residuals are checked whether they are normally, identically and independently distributed.

(i) Statistical Tests

The portmanteau-test of box and pierce (1970)

It checks whether the estimated residuals $e_t = Y_t - {}_t, \text{t} = 1,2,\cdots,\text{n}$, behave approximately like realizations from a white noise process. It is defined by;

$$Q_h = T \sum_{j=1}^{h} tr\left(\hat{C}_j \hat{C}_o^{-1} \hat{C}_j \hat{C}_o^{-1}\right)$$

Where, $\hat{C}_j = \frac{1}{T} \sum_{t=i+1}^{T} \hat{e}_t \hat{e}_{t-1}$. The test statistic has an

approximate of chi-square with $(k^2h - n^*)$ degrees of freedom where n^* is the number of coefficients excluding deterministic terms of a VAR (p) model.

(ii) Arch Tests

The multivariate arch tests, is used to test for heteroskedasticity. The test statistic is defined as

$$VARCH_{LM}(q) = \frac{1}{2}TK(K+1)R_m^2,$$

Where $R_m^2 = 1 - \frac{2}{k(k+1)}tr(\hat{s}\hat{s}_0^{-1})$ and $\hat{s}$ assigns the covariance matrix of the model. The test statistic is a chi-square distribution with $\frac{qk^2(k+1)^2}{4}$ degrees of freedom and is implemented in r using vars package.

(iii) Normality Check

The Jarque-Bera normality tests for multivariate series are implemented and applied to the residuals of a VAR (p) as separate tests for multivariate kurtosis and Skewness. The test statistics is defined as

$JM_{mv} = s_3^2 + s_4^2$ Where $s_3^2$and $s_4^2$are computed as follows;

$$s_3^2 = TC_1C_1/6$$

$$s_3^2 = T(C_2 - 3_k)(C_2 - 3_k)/24$$

Where C_1 and C_2 are the third and fourth non-central moment vectors of the standardized residuals. The JM_{mv} is a chi- square distribution with 2k degrees of freedom. The multivariate Skewness and kurtosis test are chi- square distribution with 2k degrees of freedom.

(iv) Unit Root Tests

(a) Dickey-Fuller Test & Augument Dickey-Fuller Test

DF tests the null hypothesis of the unit root against the alternative that there is no unit root in the process. It gives a formal test for checking stationarity of the model. For example, given a differenced equation of an AR (1) as,

$$\Delta X_t = (\alpha - 1)X_{t-1} + e_t$$

Then if the X_t is a random walk, then$X_t = 0$, but if X_t is stationary, Then coefficient is negative. Using the standard t-test statistic, it will be formed as

$$\hat{t}_n = \frac{1 - \hat{\alpha}}{\sqrt{\hat{\sigma}^2(\sum_{t=2}^{n} X_{t-1}^2)^{-1}}}$$

$\hat{\alpha}$ and $\hat{\sigma}$are the least squares estimators for α and $\hat{\sigma}^2$

For large n, the statistic converges to functional distribution of wiener process

$$\hat{t}_n \rightarrow \frac{W^2(1) - 1}{2\left\{\int_0^1 W^2(\mu)d\mu\right\}^{1/2}}$$

Where w is a standard wiener process.

ADF statistic is an augmented version of the df test. It is used to test for a more complicated and larger set of time series models. The test used in ADF statistic is a negative number and the more negative, the stronger the rejection of the null hypothesis. It can be applied to each variable to check for existence of co integration.

(b) Durbin Watson

This is a test statistic used to detect the presence of autocorrelation in the residuals from a regression analysis. It tests for the null hypothesis that the residuals are serially uncorrelated against the alternative hypothesis that residuals follow a stationary first order auto regression.

Given that e_t is the residual associated with the observation at time t, and then the Durbin Watson test statistic is given by;

$$d = \frac{\sum_{t=2}^{T}(e_{t-1})^2}{\sum_{t=1}^{T} e_t^2},$$

Where t is the number of observations and $d \cong 2(1 - r)$. R is the sample autocorrelation of the residuals.

If d=2, there is no autocorrelation. Then if d<2, there is evidence of positive correlation. For d>2,

Then error terms are negatively correlated which imply an underestimation of level of statistical significance. But if d<1, causes an alarm.

(c) The Qk (m) Statistic

Qk(m) is used to check the adequacy of the fitted model. It can be applied to the residual series to check the assumption that there are no serial or cross-correlations in the residuals. If the model is fit, the Qk (m) statistic of the residuals is asymptotically a chi-squared distribution with k^2m-g degrees of freedom, where g is the number of estimated parameters in the AR coefficient matrices.

(d) The T -Statistic

t-statistic is used to test the statistical significance of parameters. It is carried out to check if the model is over specified. It is also important to assess whether the stationarity and invertibility conditions are satisfied. If we factorize characteristic polynomials and one of their roots is close to unity, it may be an indication of lack of stationarity and or invertibility.

An inspection of the covariance matrix of the estimated parameters allows us to detect the possible presence of high correlation between the estimates of some parameters which can be a manifestation of the presence of a common factor in the model (Box and Jenkins; 1976).

A vector error correction model (VECM) can be used to incase of presence of a common factor (co integration) in the model. It describes the nature of any non stationarity among different component series. Applying it improves longer term forecasting over an unconstrained model.

2.1.4. Prediction

Prediction is the estimation of values whether future, current or past with respect to the given data. Future values of the original process can be predicted based on the model which fits the given data. It starts with certain assumptions and the estimates projected into the coming future either weeks, months and even years using techniques such as box-Jenkins

models, exponential smoothing, regression analysis, moving averages and projection. Any error in the assumptions results in a similar or magnified error in forecasting. Therefore, the sensitivity techniques for analysis are used to assign a range of values to uncertain factors or variables.

Once the model has been identified and passed through diagnostic check, we can use it for it for forecasting. For time $Y_1, Y_1, \cdots, Y_T$, the basic problem is then to estimate future values $\hat{Y}_{T+h}$

Of the h-steps ahead forecast made at time t. We need to forecast $\hat{Y}_{T+k}$

In such a way that the mean squared error (MSE) of the prediction is minimum.

$$MSE\left(\hat{Y}_{T+k}\right) = E\left(Y_{T+h} - \hat{Y}_{T+k}\right)^2$$

For a VAR (p) model, the 1-step ahead forecast at the time origin h

Is

$$Y_h(1) = \Pi + \sum_{i=1}^{p} \Phi_i \, Y_{h+1-i}$$

The associated forecast error is $a_h = e_h + 1$

The covariance matrix of the forecast error is Σ. If Y_t is weakly stationary, then l-step ahead forecast $Y_1(l)$converges to its mean vector, μ, as the forecast horizon increases.

(i) Advantages of VAR Modeling

1. VARs are more flexible than univariate AR models by allowing the value of a variable to depend on more than just its own lags or combinations of white noise terms

2. The researcher does not need to specify which variables are endogenous or exogenous with VAR model. All variables are endogenous.

3. Forecasting with VARs are often better compared to 'traditional structural' models.

(ii) Disadvantages of VARs

1. VARs use little theoretical information about the relationships between the variables to guide the specification of the model.

2. Often not clear how the VAR estimates of coefficient should be interpreted.

3. There are so many parameters to be estimated.

2.2. SARIMA Model

Seasonal autoregressive integrated moving average (SARIMA) processes are designed to model time series with trends, seasonal patterns and short time correlation. They have developed from the standard model of box and Jenkins (1970) and incorporate both seasonal autoregressive and moving average factors into the modeling process.

It is a class of ARIMA models suitable for data exhibiting seasonal variations. Suppose that a time series y_t has a polynomial trend of degree d. Then we can eliminate this trend by considering the process $(\Delta^d y_t)$, obtained by d times differencing.

If the filtered process $(x_t = \Delta^d y_t)$, is an ARMA (p, q) process satisfying the stationarity condition, the original process (y_t) is said to be an autoregressive integrated moving average of order p,d,q, denoted by ARIMA(p,d,q). In this case constants $a_1, a_2, \ldots a_p, b_0 = 0, b_1, \ldots, b_q \epsilon \mathbb{R}$ exist such that $x_t = \Delta^d y_t = \sum_{u=1}^{p} a_u \Delta^d y_{t-u} + \sum_{w=0}^{q} b_w \varepsilon_{t-w}, t \epsilon \mathbb{Z}$ is a white noise.

Let y_t, t=0, 1, 2 … be a non-stationary time series possibly containing seasonality.

Then y_t depends on past values such as $y_{(t-1)s}, y_{(t-2)s}$, as well as $y_{(t-1)}, y_{(t-2)}, \ldots, x_t = \Delta^d y_t$, Where Δ is a differencing operator, d is the order of non-seasonal differencing is ARMA (p, q) Process.

$$A_p(z)x_t = B_q(Z)\varepsilon_t$$

Then y_t, is ARIMA (p, d, q) model but if $x_t = \Delta^d \Delta^D y_t$

Where d is the order of seasonal differencing, the model is referred to as seasonal autoregressive integrated moving average SARIMA written as ARIMA (p,d,q)(P,D,Q)

Is given by

$$A_p(z)A_P(Z^s)x_t = B_q(Z)B_Q(Z^s)\varepsilon_t$$

Where $A_p(z), A_P(Z^s), B_q(Z)$ and $B_Q(Z^s)$, are polynomials of order p, P, q, Q respectively, z is a backshift operator, s is the seasonal period of the series. A_p is an AR process of order p, $A_p$is an AR process with order of seasonal component, B_q is an ma process of order q and B_Q is an MA process with order of seasonal component.

Fitting SARIMA models to the meagre data using a semi-automated approach based on a combination of the box-Jenkins method with small-sample, bias-corrected Akaike information criteria (AIC) model selection (Rothschild et al., 1996; Brockwell and Davis, 2002). This approach involved three major steps:

2.2.1. Model Identification

Choosing the parameters p and q, the order q of a moving average MA (q) -process can be estimated by means of the empirical autocorrelation function r (k)

The order p of an AR (p) -process can be estimated in an analogous way using the partial autocorrelation function. MA (p, q) -process we take the pair (p, q), which minimizes some function i. e Akaike's information criterion, Hannan information criterion with bias correction and the Bayesian information criterion discussed earlier under VAR.

2.2.2. Estimation of the Model Coefficients

The coefficients $a_1, \cdots, a_p$ and $b_1, \cdots, b_q$ are estimated using maximum likelihood method or otherwise. Most of the estimation methods are already implemented in existing computer software using iterative techniques.

2.2.3. Diagnostic Check

The fit of the ARMA (p, q) model with the estimated coefficients is checked. These involve scrutinizing the estimated residuals and ensure that they behave approximately like the realizations from a white noise process.

2.2.4. Forecasting

The forecast of future values of the original process is based on the model which is adequate and fits the given data.

2.3. The Concept of Re-Shaped Time Series

Given n -observations in univariate time series data and assuming the series is seasonal with period s.

For illustration, we will use (s=4), where each observation has 4 variables i.e x_1, x_2, x_3 and $x_4$which in our case will be seasons (s=4 variables). Arranging the data in a serpentine manner for all the n –observations $x_1, x_2, \ldots, x_n$, we will treat each row as a vector of $Y_i's$, where i=1,2,…,t. The vector time series is of order s×t.

For a seasonal univariate time series data, $x_1, x_2, \ldots, x_n$ we can reshape it as shown in table 1 below;

Table 1. *Concept of re-shaped time series*

	Y_1	Y_2	Y_3	Y_4
	↓	↓	↓	↓
	Q_1	Q_2	Q_3	Q_4
Y_1	x_1	x_2	x_3	x_4
Y_2	x_5	x_6	x_7	x_8
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
Y_T	x_{n-3}	x_{n-2}	x_{n-1}	x_n
$\hat{Y}_{T+1}$ real value	$\hat{x}_{n+1}$	$\hat{x}_{n+2}$	$\hat{x}_{n+3}$	$\hat{x}_{n+4}$

For $x_1 = Q_{1t}, x_1 = Q_{1t}, \ldots, x_1 = Q_{1t}$, then the VAR form of the re-shaped time series is written as;

$$Y_t = \pi + Q_1 y_{t-1} + \ldots + Q_{t-p} y_{t-p} + e_t, \; t=1,2,3,\ldots t$$

In matrix form it is represented as follows;

$$\begin{bmatrix} Q_{1t} \\ Q_{2t} \\ \vdots \\ Q_{nt} \end{bmatrix} = \begin{bmatrix} \Pi 1 \\ \Pi 2 \\ \vdots \\ \Pi n \end{bmatrix} + \begin{bmatrix} \Phi_{11}^1 & \Phi_{12}^1 & \ldots & \Phi_{1n}^1 \\ \Phi_{21}^1 & \Phi_{21}^1 & \ldots & \Phi_{2n}^1 \\ \vdots & \vdots & \ldots & \vdots \\ \Phi_{n1}^1 & \Phi_{n2}^1 & \ldots & \Phi_{nn}^1 \end{bmatrix} \begin{bmatrix} Q_{1t-1} \\ Q_{2t-1} \\ \vdots \\ Q_{nt-1} \end{bmatrix} + \begin{bmatrix} \Phi_{11}^2 & \Phi_{12}^2 & \ldots & \Phi_{1n}^2 \\ \Phi_{21}^2 & \Phi_{21}^2 & \ldots & \Phi_{2n}^2 \\ \vdots & \vdots & \ldots & \vdots \\ \Phi_{n1}^2 & \Phi_{n2}^2 & \ldots & \Phi_{nn}^2 \end{bmatrix} \begin{bmatrix} Q_{1t-1} \\ Q_{2t-1} \\ \vdots \\ Q_{nt-1} \end{bmatrix} + \ldots + \begin{bmatrix} \Phi_{11}^p & \Phi_{12}^p & \ldots & \Phi_{1n}^p \\ \Phi_{21}^2 & \Phi_{21}^p & \ldots & \Phi_{2n}^p \\ \vdots & \vdots & \ldots & \vdots \\ \Phi_{n1}^p & \Phi_{n2}^p & \ldots & \Phi_{nn}^p \end{bmatrix} \begin{bmatrix} Q_{1t-p} \\ Q_{2t-p} \\ \vdots \\ Q_{nt-p} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ \vdots \\ e_{nt} \end{bmatrix}$$

2.4. Performance Measures

S-step ahead forecast, $\hat{x}_s$, will be equivalent to one-step ahead forecast, $\hat{Y}_{T+1}$, for vector.in order to get the best forecasting, the parameters of the models are adjusted to minimize the forecasting error. By defining a forecasting error (loss function), we search for the best parameters in both models that minimizes this function.

To compare performance of the VAR and SARIMA models, we compute a set of indicators for the quality of time series forecast methods. These include;

2.4.1. Mean Squared Error (MSE)

For the forecast in both predictions should be minimum

$$MSE(\hat{Y}_t) = E(\hat{Y}_t - Y_t)^2$$

The results of both forecasts will be compared with the real values.

2.4.2. Root Mean Squared Error (RMSE)

Is the square root of the average of all squared errors, according to Wang and Lim (2005). It ignores any over and under- estimation

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}$$

2.4.3. Mean Percentage Error (MPE)

MPE is a percentage error measurement which allows comparison of under and over-estimation. It takes into account whether a forecasting method is biased.

$$MPE = \frac{1}{T}\sum_{t=1}^{T}\frac{(Y_t - \hat{Y}_t)}{Y_t} \times 100$$

2.4.4. Mean Absolute Percent Error (MAPE)

It is a percentage error measurement, which allows comparison of under and over-estimation. It is particularly useful when the units of measurement are relatively large.

$$MPE = \frac{1}{T}\sum_{t=1}^{T}\frac{|Y_t - \hat{Y}_t|}{Y_t} \times 100$$

2.4.5. Theil's U Statistic

Theil's U-statistics see Theil (1958) is used as a measure of forecasting error that is minimized. It is a relative measurement based on comparison of the predicted change with the observed change. The value of u lies between 0 and 1. If u equals to 0, there is a perfect fit, whereas u equals to 1 implies that forecasting of data is very poor.

3. Results

3.1. Introduction

This chapter presents results using secondary data in which the concept of SARIMA model has been known. The data set was used to test the concept of VAR and examine its appropriateness. This was analyzed by use of statistical software for model fitting and testing. Akaike information criterion (AIC) is used to select the best fitting model. Bayesian information criterion (BIC) also was used.

3.2. Seasonal ARIMA Model

3.2.1. Testing the Stationarity

Macro dataset used by stock and Watson in their introduction to econometrics where all unemployment rates for 16 years and above data is used for analysis. We denote the quarterly data on macro dataset as unemploymentrates. Time

series plot shows evidence of cyclic and random components. The spikes display evidence of strong seasonality. The time series plot is shown in figure 1 below.

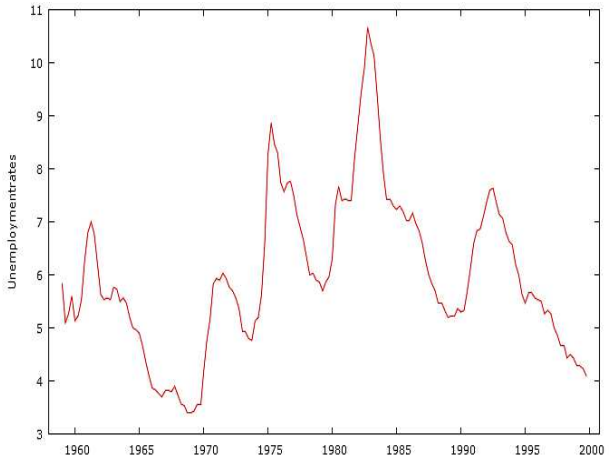

Figure 1. *Time series plot.*

Autocorrelation function and partial correlation function plot of unemploymentrates shown below is evident of presence of a unit root. This is shown by the PACF value close to 1 confirming non-stationarity. The ACF on the other hand shows the autocorrelation between the variables and the lags of itself. It is evident that in all the lags, there is statistically significance of autocorrelation at 1%. ACF and PACF plot is shown in figure 2.

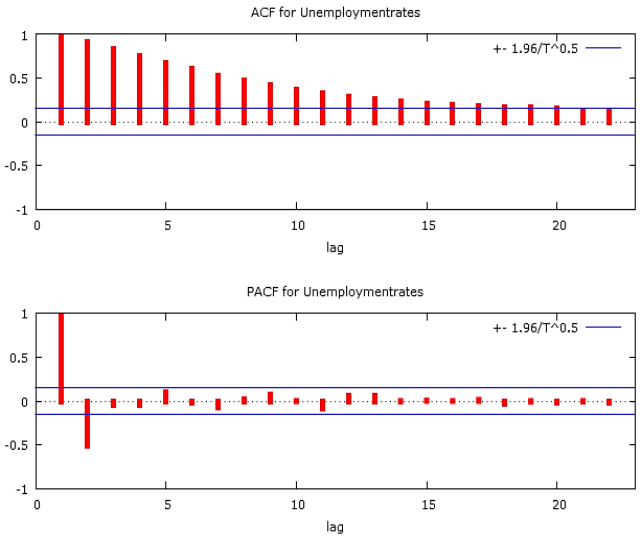

Figure 2. *ACF and PACF of unemployment rates*

We also test for presence of unit root test using ADF test. The ADF test results show that p-value is greater than 0.01 hence we fail to reject the null hypothesis of non stationarity of the process and conclude that the data is non-stationary. We take the logarithm transformation of the data in order to reduce problems of heteroscedasticity. Taking the first difference of the data to ensure stationarity assumptions of ARIMA model is satisfied, the data shows stationarity with its plot. Testing for the presence of unit roots using ADF test, the p-value is less than 0.01. We therefore reject the null hypothesis of non-stationarity and conclude that the data is stationary at critical value of 1%. This is shown in table 2.

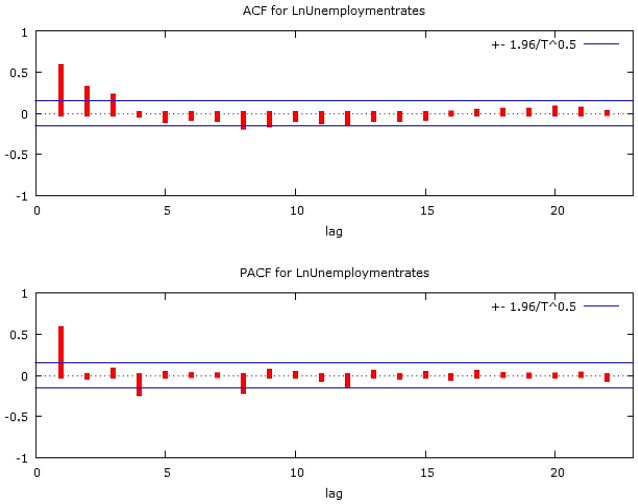

Figure 3. *ACF and PACF of log of unemployment rates*

Table 2. *ADF test for both unemploymentrates and its first difference.*

Variable	ADF test statistic	P-value	Critical value
Unemploymentrates	-1.58286	0.4914	0.01
Δlnunemploymentrates	-4.31345	0.0004159	0.01

We can also test the stationarity using KPSS test with the null hypothesis of stationarity. The t-statistic is less than the critical value at 1%. We therefore fail to reject the null hypothesis of stationarity at 1%. Therefore, we conclude that the data is stationary. This is shown in table 3 below.

Table 3. *KPSS of differenced lnunemploymentrates.*

	Test statistic	P-value
KPSS test statistic	0.0878173	0.738

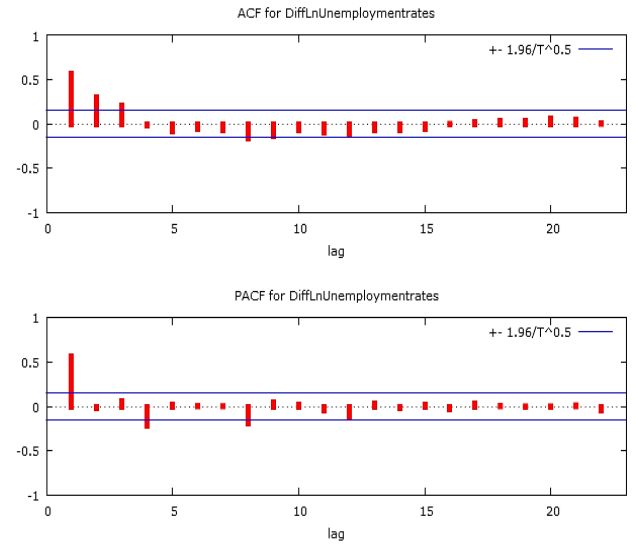

Figure 4. *ACF and PACF of differenced lnunemployment rates.*

Both plots of differenced data show that there is a little

autocorrelation in ACF. The PACF shows that the there is no spike close to one. This confirms the stationarity of the data. The plots are shown in figure 4.

3.2.2. Selection of the Order

The order of the model is checked to identify the appropriate SARIMA model. The model which minimizes the information criterion is chosen. In our data, the ARIMA

$(0, 1, 1\times0, 1, 0)_4$ model is found to be the best model since its information criterions are minimum.

3.2.3. Estimation of coefficient Parameters

From the fitted model, we can obtain its coefficients. The coefficients of ARIMA $(0, 1, 1\times0, 1, 0)_4$ model based on the information criterion is shown below with its coefficients being statistically significant the AIC value is minimum and standard error variance is also small.

Model 30: ARIMA, using observations 1960:2–1999:4 (t = 159)

Dependent variable: $(1 - L)(1 - L^s)$ lnunemploymentrates standard errors based on hessian coefficient std. Error z p-value

Const 0.000773 0.00927 0.0833 0.9336

Θ 0.928806 0.0809547 1.4732 0.0000

Mean dependent var 0.000312

s.d. Dependent var 0.074352

Mean of innovations 0.000050

s.d. Of innovations 0.060778

Log-likelihood 218.6787

Akaike criterion −431.3574

Schwarz criterion −422.1507

Hannan–quinn −427.6187

Real Imaginary Modulus Frequency

MA

root 1 −1.0767 0.0000 1.0767 0.5000

The results are shown in table 4 below.

Table 4. *Estimates of parameters ARIMA (0, 1, 1×0, 1, 0)$_4$.*

	Coefficient	Std. Error	Z	P-value
constant	-1.000	0.0226214	-44.21	0.000
Thetha-1	-0.576546	0.0840321	-6.861	6.84e-012
S.d dependent var.	0.038212	Aic	-629.6445	
Log-likelihood	318.8223	Hannan-quinn	-624.9712	

From the fitted model we can see clearly that the p-values of the coefficients are statistically significant. The equation for the fitted model is expressed as;

$$(1-Z)(1-Z^4)Xt = (1 + 0.928806B)\varepsilon_t$$

3.2.4. Diagnostic Check

To check if the fitted model is significant, one of the ways of checking this is by plotting the frequency distribution of the residuals. Figure 5 below, shows the frequency distribution of the residuals of the fitted model ARIMA $(0, 1, 1\times0, 1, 0)_4$. This shows that the residuals are normally distributed since the p-value is greater than 0.01.

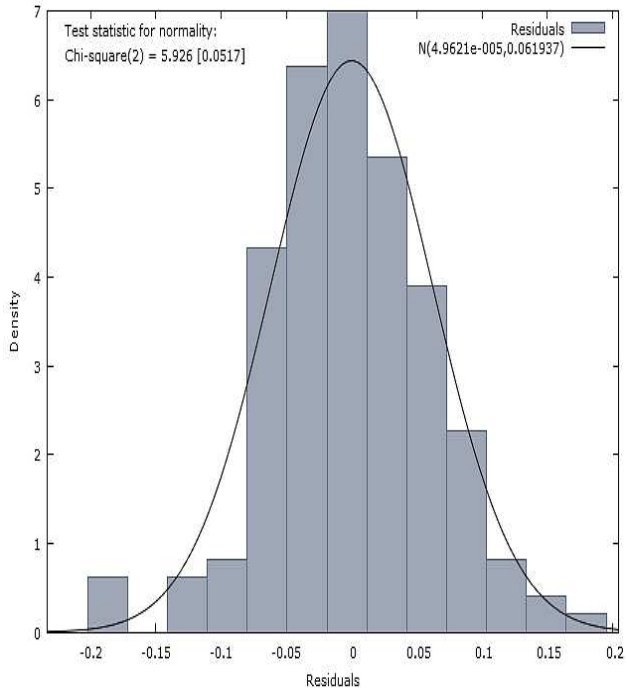

Figure 5. *Frequency distribution of the residuals.*

We can also check the normality of the residuals using the QQ-plot. From the plot, there are no outliers and all points lie on the line making us to conclude that the residuals of the fitted model are normally distributed.

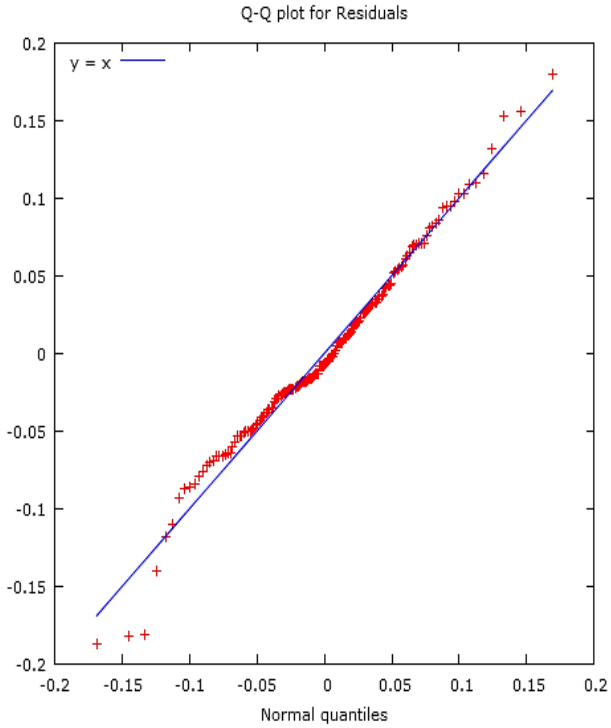

Figure 6. *normal QQ-plot of residual.*

3.2.5. Forecasting

Forecasting for ARIMA $(0, 1, 1\times0, 1, 0)_4$ model is shown in the table below.

Table 5. *predictions results of ARIMA (0, 1, 1×0, 1, 0)$_4$ for 1998.*

Variable	Actual value	Prediction value	Error
1998:1	4.6667	4.3456	0.3211
1998:2	4.4333	4.7373	0.304
1998:3	4.5000	4.0605	0.4395
1998:4	4.43333	4.7509	0.3176

3.3. VAR Modeling

The seasonal univariate time series data is reshaped to form vector form of four variables since the data is quarterly. Data between 1959 and 1998 are used in-sample estimation and data between 1998 and 1999 are used for the out-of-sample forecasting purposes. Figures below shows the time series plots of the four variables during the sample period. The variables are denoted Q1, Q2, Q3, and Q4. The figure below displays all the plots.

3.3.1. Testing Stationarity

Autocorrelation function and partial correlation function plot of the variables shown below in figure 7, 8, 9 and 10 is evident of presence of a unit root. This is shown by the PACF value close to 1 confirming non-stationarity. The ACF on the other hand shows the autocorrelation between the variables and the lags of itself.

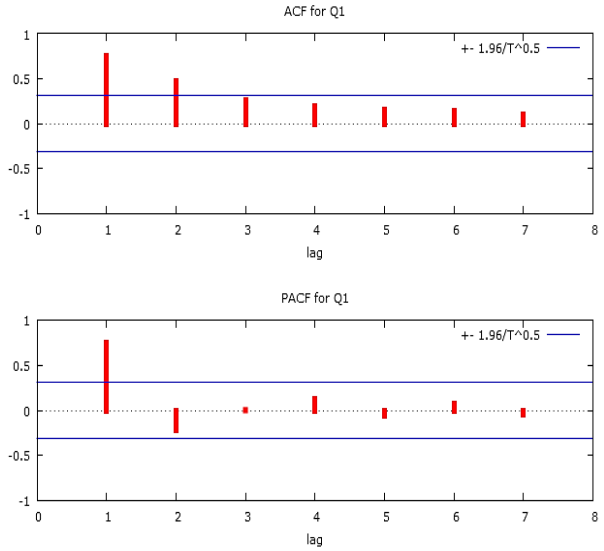

Figure 7. *correlogram of Q1.*

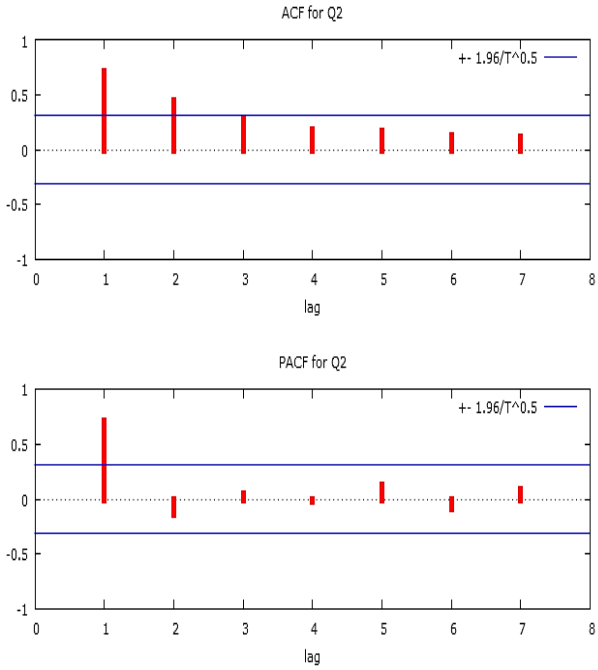

Figure 8. *correlogram of Q2.*

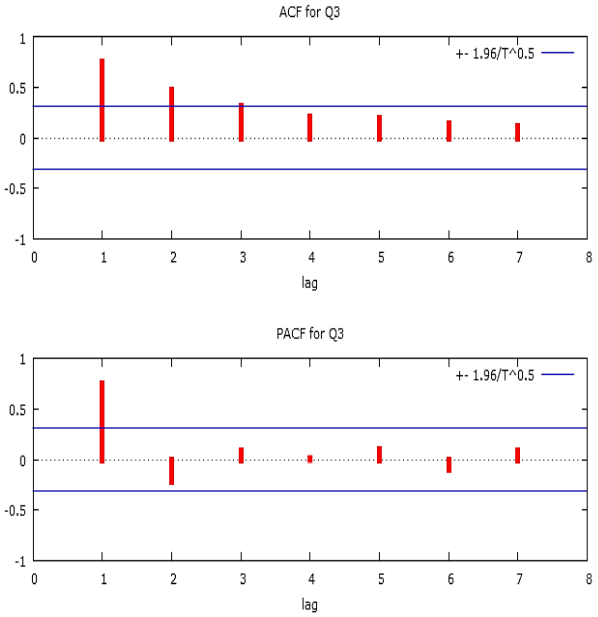

Figure 9. *correlogram of Q3.*

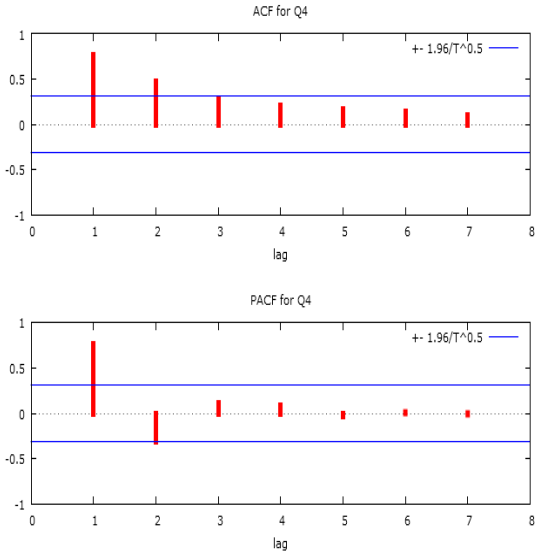

Figure 10. *correlogram of Q4.*

We can also check the stationarity by testing the unit roots. One of the tests used is the augmented dickey-fuller test applied to the series in order to test for unit-roots. Table 6 below shows the ADF results of both the variables and their first differences of the series.

From the table above, the ADF test results indicate that all variables are non-stationary by not rejecting the null hypothesis of unit-root for all the variables at 1% critical value, but they are all stationary after first differencing. We can therefore conclude that all the time series are integrated of at least order one; therefore, we use differenced series in our analysis.

Table 6. *ADF unit root test results.*

Variable	Deterministic terms	Test value	P-value	Conclusion
Q1	Constant, trend	-2.2501	0.4496	Non-stationary
Differenced q1	Constant	-5.57543	3.236e-005	Stationary
Q2	Constant, trend	-2.34281	0.402	Non-stationary
Differenced q2	Constant	-5.98549	1.264e-005	Stationary
Q3	Constant, trend	-2.57543	0.5616	Non-stationary
Differenced q3	Constant	-5.35931	7.847e-005	Stationary
Q4	Constant, trend	-2.75285	0.2152	Non-stationary
Differenced q4	Constant	-4.90132	3.236e-005	Stationary

3.3.2. Model Identification

For VAR model to be identified, the data should be stationary. We then determine the true lag order for the model. Selecting a higher order lag length than the true lag lengths increases the mean square forecast errors of the VAR, and selecting a lower order lag length than the true lag lengths usually causes auto correlated errors. This was pointed out by Lutkepohl. Hence, accuracy of forecasts from VAR models highly depends on selecting the true lag lengths. There are several statistical information criterions for selecting a lag length. This include: Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-quinn information criterion (HQC). The table below shows the VAR model lag order selection criteria.

Table 7. *VAR(p) model lag order selection criteria.*

Lag	Aic	Bic	Hqc
1	1.841568	2.730338*	2.148371
2	1.493935*	3.093722	2.046181*
3	1.650475	3.961277	2.448163

The AIC and HQC is minimized at p=2. But BIC is minimized at p=1. We take BIC since it is appropriate for data with sample less than 120 quarters. Therefore, we have 20 parameters to estimate.

Table 8. *The coefficients of estimated VAR model.*

	Δq1	Δq2	Δq3	Δq4
Constant	0.0176441	0.0135887	-0.00790380	-0.0443565
Δq1(-1)	0.0230364	0.232444	0.153521	-0.135078
Δq2(-1)	-0.259410	-0.501622	0.361898	0.853679
Δq3(-1)	-0.556027	-0.888681	-1.88762	-1.90887
Δq4(-1)	1.53304	1.74647	1.77344	1.27800

The VAR(1) can be expressed as follows:

$\Delta Q1_t$=0.0176441+0.0230364*$\Delta Q1_{t-1}$-0.259410*$\Delta Q2_{t-2}$-0.556027*$\Delta Q3_{t-3}$+1.53304*$\Delta Q4_{t-4}$

$\Delta Q2_t$=0.0135887+0.0232444*$\Delta Q1_{t-1}$-0.501622*$\Delta Q2t$-2-0.888681*$\Delta Q3t$-3+1.74647*$\Delta Q4_{t-4}$

$\Delta Q3_t$=-0.00790380+0.153521*$\Delta Q1_{t-1}$+0.361898*$\Delta Q2_{t-2}$-1.88762*$\Delta Q3_{t-3}$+1.77344*$\Delta Q4_{t-4}$

$\Delta Q4_t$=-0.0443565-0.135078*$\Delta Q1_{t-1}$+0.853679*$\Delta Q2_{t-2}$-1.90887*$\Delta Q3_{t-3}$+1.27800*$\Delta Q4_{t-4}$

3.3.3. Model Diagnostic Checking

We test for the autocorrelation of the model, to check whether the residuals of the fitted model are identically distributed. Table 9 below shows that the p-value of all the four equations are greater than critical value at 1% hence we do not reject the null hypothesis of no autocorrelation, meaning there is consistent estimators.

Table 9. *Autocorrelation tests for the four equations.*

	Ljung-box q	P-value
Equation 1	1.77125	0.183
Equation 2	0.657489	0.417
Equation 3	5.18819e-005	0.994
Equation 4	0.140521	0.708

We also test for the auto effect correlation using arch. The p-value is found to be greater than 0.01. Meaning we fail to reject the null hypothesis that there is no auto effect. This implies that there is conditional homoscedasticity which allows for a valid inference. This is shown in table 10 below;

Table 10. *ARCH test for the equations.*

	Test statistic	P-value
Equation 1	4.10135	0.0428491
Equation 2	0.0950886	0.757805
Equation 3	0.746557	0.387568
Equation 4	0.0531222	0.817717

The p-value is greater than 0.01. Hence we fail to reject the null hypothesis that the residuals of the VAR model are normally distributed. Hence we can make an inference.

Residual correlation matrix, c (4 x 4)
1.0000 0.84237 0.70092 0.53609
0.84237 1.0000 0.93270 0.69950
0.70092 0.93270 1.0000 0.87207
0.53609 0.69950 0.87207 1.0000
Doornik-hansen test, Chi-square (8) = 9.26726 [0.3203]

Testing all the residuals, the residuals of the VAR model have the p-value of 0.3203 greater than 0.01, shows that all the residuals are normally distributed with the VAR model fitted being valid.

3.3.4. Prediction and Results

After obtaining the valid model, we now check the forecasting performance measures of each variable and compare with the actual value. The forecasting results of VAR-1 model for 1998 are shown in table 11 below.

Table 11. *The Forecasting results of VAR-1 model for1998.*

Variable	Actual value	Prediction value	Error
Δq1	4.6667	4.5916	0.0751
Δq2	4.43333	4.5905	-0.1572
Δq3	4.5000	4.6554	-0.1554
Δq4	4.4333	4.7524	-0.3191

3.3.5. Performance Comparison

Using the performance measures mentioned above, the table below shows the summary comparison performance of the two models.

Table 12. *Forecasting results of VAR and SARIMA model for 1998.*

Variable	Actual value	Prediction (SARIMA)	Prediction (var)	Error (SARIMA)	Error (Var)
Q1	4.6667	4.3456	4.5916	0.3211	0.0751
Q2	4.4333	4.7373	4.5905	-0.304	-0.1572
Q3	4.5000	4.0605	4.6554	0.4395	-0.1554
Q4	4.4333	4.7509	4.7524	-0.3176	-0.3191
MsE				0.1224	0.0391
RmsE				0.3498	0.1977
MpE				2.646%	-3.147%
MapE				7.662%	3.951%
Theil's u				0.3498	0.0216

3.4. Cumulative Results

The purpose for this project is to examine if VAR model is appropriate in forecasting a reshaped univariate seasonal time series. SARIMA model has been used for comparison purposes. Macro dataset used by stock and Watson in their introduction to econometrics where all unemployment rates for 16 years and above data were used. The data was quarterly and was reshaped into four variables and apply VAR for forecasting. This was analyzed in detail and the summary of other data set one was shown. The data contains observations from 1959 to 1999 where observations from 1959 to 1997 were used for model fitting. The year 1998 was used for prediction and comparison purposes.

Taking the detailed sample given above as an example, the performance of the VAR model is found to be better than the SARIMA model with the performance measures being minimum. The mean absolute percentage error for SARIMA is 7.662% and mean absolute percentage error for VAR is 3.951%. The other performance measures also show that VAR is a better model for forecast. Theil's u statistic for VAR is almost close to 0, implying almost a perfect fit.

However, in practice we normally face the challenge of the observations of seasonal univariate time series data that can be reshaped. This becomes a challenge in reshaping the data. Also fitting the reshaped data set into a VAR model is a bit complicated than the SARIMA model since each of the reshaped variables is non-stationary. However, VAR model is still evident to be the best model for all the observations than the SARIMA model.

References

[1] Mei, q., liu, y. & jing, x. (2011).*Forecast the gdp of shanghai based on the multi-factors VAR model*. Journal of Hubei University of Technology, 26(3).

[2] Box, g.e.p., Jenkins, g.ma.1976 .*Time series analysis Forecasting and control, 2nd ed. Holden-Day.San.*

[3] Box, g. E. P. & pierce. D. A. 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical association*, 65(332), 1509–1526

[4] Sims, c.a. 1980. *Macroeconomics and reality*. Econometrica, 48, 1-48

[5] Zivot, e. And J.wang (2006). *Modeling Financial Time series with Splus, 2nd Edition.Springer*, *p.p. 385-386.*

[6] Clarida, r. H. & Friedman, b. M. (1984). The behavior of u.s. Short-term interest rates since October 1979. *The Journal of finance, 39(3), 671-682.*

[7] Mills, T.C. (1999). *The Econometric modeling of financial time series, second edition*. Cambridge university press, Cambridge

[8] Xiao Han Cai (2008). *Time series analysis of air pollution co in California south coast area, with seasonal ARIMA model and VAR model*. University of California, Los Angeles.

[9] Zivot, E. and J.Wang (2006), *Modeling Financial Time Series with SPlus*, Springer.

[10] Pfa, Bernhard (July 2008), 'Var, svar and svec models: Implementation within r package vars',*Journal of Statistical software published by the Statistical Association.*

[11] Aidoo, E (2010), 'Modelling and forecasting inflation rates in ghana: An application of sarima models', *Hogskolan Dalarna.*

[12] Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and* Statistics .

[13] Halim, S. and I.N. Bisono (2008), 'Automatic seasonal autoregressive moving average models and unit root test detection', International Journal of Management Science and Engineering Management.

[14] J.D., Cryer and K.S Chan (2008), *Time Series Analysis with Applications in R.*, Springer Science +Business Media.