RESEARCH ARTICLE

MODELING MONTHLY RAINFALL TIME SERIES USING ETS STATE SPACE AND SARIMA MODELS

[1]Fadhilah Yusof and [2]Ibrahim Lawal Kane

[1,2]Department of Mathematical Sciences, Faculty of Science Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

ARTICLE INFO	ABSTRACT
	The study of rainfall time series of two selected weather stations in Malaysia using various statistical methods enabled to analyse the temporal behaviour of rainfall in the study areas. Time-series analysis is an important tool in modelling and forecasting rainfall. SARIMA $(1, 1, 2)(1, 1, 1)_{12}$, SARIMA$(4, 0, 2)(1, 0, 1)_{12}$ with constant and ETS state space models based on exponential smoothing were built. All the models proved to be adequate. Therefore, could give information that can help decision makers establish strategies for proper planning of agriculture, drainage system and other water resource applications in Malacca and Kuantan.

INTRODUCTION

There are different categories of predictive modeling; the causal method and the time series method. The causal method assumes that the predicted variable is determined by independent explanatory variables, that is regression. The causal method has been applied to predict for example, building price (Runeson 1988), construction cost (Koehn and Navvabi 1989) etc. While the time series methods determines future trend based on past values and corresponding errors. Since a time series method only require the historical data, it is widely used to develop predictive models. A time series is simply a set of observations measured at successive points in time or over successive periods of time. Forecasts essentially provide future values of the time series on a specific variable. Time series forecasting methods are based on analysis of historical data. It makes the assumption that past patterns in data can be used to forecast future data points. The popular ways of time series modelling include an Autoregressive Moving Average; Simple and Weighted Moving Averages; Single, Double and Triple Exponential Smoothing. Among them, the triple exponential smoothing is more sensitive and effective in this situation, because it has three control constant i.e. irregular, trend and seasonal influence. The value of three constants can control and manage influence of time segmentation through the specific time duration.

Simple exponential smoothing was originally developed from heuristics for generating point predictions from historical time series data. Although the Exponential Smoothing methods has been around since the 1950s. There has not been a well-developed modelling framework incorporating stochastic models, like lihood calculation, prediction intervals and

procedures for model selection. Today exponential smoothing refers to a suite of possible forecasting functions (Hyndman et al, 2005). In practice, the forecasting function is generally selected based on the practitioner's assessment of the structure of the relevant data, for example, whether local trends or seasonal factors are important components for describing the data. In addition to being a function of historical data, the forecasting functions rely upon smoothing parameters that must be selected by the practitioner. Both the choice of forecasting function and the degree of smoothing within that forecasting function are generally determined either through experience, or through ad-hoc forecasting criteria such as the minimum mean squared error criterion.

The theory of exponential smoothing was built on heuristics that had a good track record for generating reliable point predictions of demand in business applications. Heuristics also exist for determining the uncertainty surrounding predictions. Of course, accurate measures of uncertainty can be very important. Rainfall estimates are an important component of water resources applications, example, in designing drainage system and irrigation, an accurate estimate of rainfall is needed. There are also concerns with producing valid estimates using appropriate methods. In order to develop a comprehensive solution to the forecasting problem, including addressing the issue of uncertainty in predictions, a statistical model must be employed. One of the most successful automatic forecasting methods in practice is based on exponential smoothing models. There are a variety of such models, each having the property that forecasts are weighted averages of past observations with recent observations given relatively more weight than older observations. The name "exponential smoothing" reflects the fact that the weights decrease exponentially as the observations get older (Hyndman, 2002). The main aim of this paper is to explore and

---

*Corresponding author: Fadhilahy@utm.my, ibrahimlk131@gmail.com

model the temporal characteristics of rainfall time series; and compare the performance of Seasonal ARIMA models and State space models with applications to two monthly rainfall series in peninsular Malaysia. The versatile and fully automatic ETS framework requires neither stationary nor strict linearity to produce contemporaneous time series forecasts for variable time horizons. Consequently, it is extensively employed in, e.g., econometrics and inventory control where automatic forecasts are required for a large number of diverse time series. This forecasting framework, whose performance was recently and favourably compared to those of several forecasting techniques across thousands of time series (Makridakis and Hibon, 2000), adapts to underlying alterations in series dynamics and automatically revises forecasts as new observations. We adapt two approaches in modelling and forecasting rainfall in this paper. The first is the state space model based on exponential smoothing methods and the second is Box Jenkins model building technique. The rest of the paper is organized as follows: Section Two gives the methodology employed in this paper; Section Three gives discusses the results of the finding and Section Four gives the Conclusion and Recommendations. The algorithms and modelling frameworks for automatic univariate time series forecasting which we employed in this work are implemented in the forecast package in R software version 2.14.1.

MATERIAL AND METHODS

Seasonal ARIMA

A time series is said to be seasonal if there exists a tendency for the series to exhibit a periodic behaviour after certain time interval. The usual ARIMA models cannot really cope with seasonal behaviour, it only model time series with trends. Seasonal ARIMA models are formed by including an additional seasonal terms in the ARIMA models and are defined by seven parameters. The seasonal ARIMA denoted by ARIMA(p,d,q)(P,D,Q)s is given as:

$$(1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p)(1-\beta B^s-\beta_2 B^{2s}-...-\beta_P B^{Ps})(1-B)^d(1-B^s)^D y_t =$$

$$c+(1-\psi_1 B-\psi_2 B^2-...-\psi_q B^q)(1-\theta B^s-\theta_2 B^{2s}-...-\theta_Q B^{Qs})\varepsilon_t$$

Where

$(1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p)$ = Autoregressive part of order p (AR(p))

$(1-\beta B^s-\beta_2 B^{2s}-...-\beta_P B^{Ps})$ = Seasonal Autoregressive part of order P (AR$_s$(P))

$(1-\psi_1 B-\psi_2 B^2-...-\psi_q B^q)$ = Moving average part of order q (MA(q))

$(1-\theta B^s-\theta_2 B^{2s}-...-\theta_Q B^{Qs})$ = Seasonal Moving average part of order Q (MAs (Q))

$(1-B)^d$ = differencing of order d (I (d))

$(1-B^s)^D$ Seasonal differencing of order D (I$_s$ (D))

s is the period of the seasonal pattern appearing

The idea behind the seasonal ARIMA is to look at what are the best explanatory variables to model a seasonal pattern. Details of such model can be found in Box and Jenkins (1976) and Madsen (2008).

The decomposition of a time series

A time series can be thought of as consisting of four different components

- Trend (T): In general, natural, human, economic and other processes produce gradual trends. A trend is a long-term component that represents a growth or a decline of a time series over an extended period of time.
- Seasonal component (S): This term of seasonality is used for time series defined at time intervals which are fractions of a year. It is a pattern of change that repeats itself from year to year.
- Cyclical component (C): Changes in time series sometimes show a wavelike fluctuation around a trend, which shows the possible existence of periodicity with longer intervals.
- Irregular component (ε): This is a part of a time series represented by residuals, after the above-mentioned components have been removed.
- The time series that contained the above components is usually described as

$$y_t = T_t + S_t + C_t + \varepsilon_t$$

Ljung-Box Test

Ljung and Box (1978) proposed a Q-Test called Ljung–Box test which is commonly used in autoregressive integrated moving average (ARIMA) modelling. It is applied to the residuals of a fitted ARIMA model, not the original series, and in such applications the hypothesis actually being tested is that the residuals from the ARIMA model have no autocorrelation, or it performs a lack-of-fit hypothesis test for model misspecification, which is based on the Q-statistic given as:

$$Q = N(N+2)\sum_{j=1}^{L}\frac{\hat{\rho}_j^2}{(N-j)}$$

Where N = sample size, L = number of autocorrelation lags included in the statistic, and $\hat{\rho}_j^2$ is the squared sample autocorrelation at lag j. Under the null hypothesis of no serial correlation, the Q-test statistic is asymptotically Chi-Square distributed.

State Space Models

In order to improve predictability, ES models will be fitted to each dataset, then the fitted models are used to make the out-of-sample forecast. The whole datasets from 1968:1 to 2003:12 were used as an example to illustrate the model fitting process. The ES models, which are encapsulated within a state-space framework hereafter referred to as ETS for error (E), trend (T), and seasonal components (S). The E component is either additive (A) or multiplicative (M), T and S components may be A, M, or inexistent (N); last, T may also be dampened additively (A$_d$) or multiplicatively (M$_d$). For example, the ETS method MMN has E(M), T(M), and S(N) structures. There are 30 possible ES combinations within the forecasting framework, comprising linear and non-linear ones (Hyndman and Khandakar, 2008). Exponential smoothing methods were originally classified by Pegels' (1969) taxonomy. This was later extended by Gardner (1985), modified by Hyndman et al. (2002), and extended again by Taylor (2003), giving a total of fifteen methods seen in table 1. However, only the 15 ES methods with multiplicative error structures (heteroskedastic)

were considered for time series analysis (Medina et al, 2008). The models also yield more realistic 95% prediction interval values. Furthermore, a reduction in the number of ES methods evaluated also diminishes the expensive computational time. Summary of the models are given in table 1.

Table 1 Fifteen exponential smoothing methods

Trend component	Seasonal component		
	N None	A Additive	M Multiplicative
N (None)	NN	NA	NM
A (Additive)	AN	AA	AM
A_d (Additive Damped)	A_dN	A_dA	A_dM
M (Multiplicative)	MN	MA	MM
M_d (Multiplicative damped)	M_dN	M_dA	M_dM

Additive error model: ETS (A, A_d, N)

Let $\mu_t = \hat{y}_t = l_{t-1} + b_{t-1}$ denote the one-step forecast of y_t assuming that the values of all parameters are known. Also, let $\varepsilon_t = y_t - \mu_t$ denote the one-step forecast error at time t. Hyndman and Khandakar (2008) found that

$$y_t = l_{t-1} + \phi b_{t-1} + \varepsilon_t \qquad (1)$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha \varepsilon_t \qquad (2)$$

$$b_t = \phi b_{t-1} + \beta^* (l_t - l_{t-1} - \phi b_{t-1}) = \phi b_{t-1} + \alpha \beta^* \varepsilon_t \qquad (3)$$

The last expression can simplify by setting $\beta = \alpha \beta^*$. The three equations above constitute a state space model underlying the damped Holt's method. Note that it is an innovations state space model (Anderson and Moore 1979; Aoki 1987) because the same error term appears in each equation. Written in standard state space notation by defining the state vector as $x_t = (l_t, b_t)'$, and expressing (1) – (3) as

$$y_t = \begin{bmatrix} 1 & \phi \end{bmatrix} x_{t-1} + \varepsilon_t \qquad (4)$$

$$x_t = \begin{bmatrix} 1 & \phi \\ 0 & \beta \end{bmatrix} x_{t-1} + \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \varepsilon_t \qquad (5)$$

The model is fully specified once the distribution of the error term ε_t is stated. Usually the assumption is that these are independent and identically distributed, following a normal distribution with mean 0 and variance σ^2, written as $\varepsilon_t \square NID(0, \sigma^2)$. The model ETS (A,A,N) has additive errors, additive trend and no seasonality, in other words, this is Holt's linear method with additive errors.
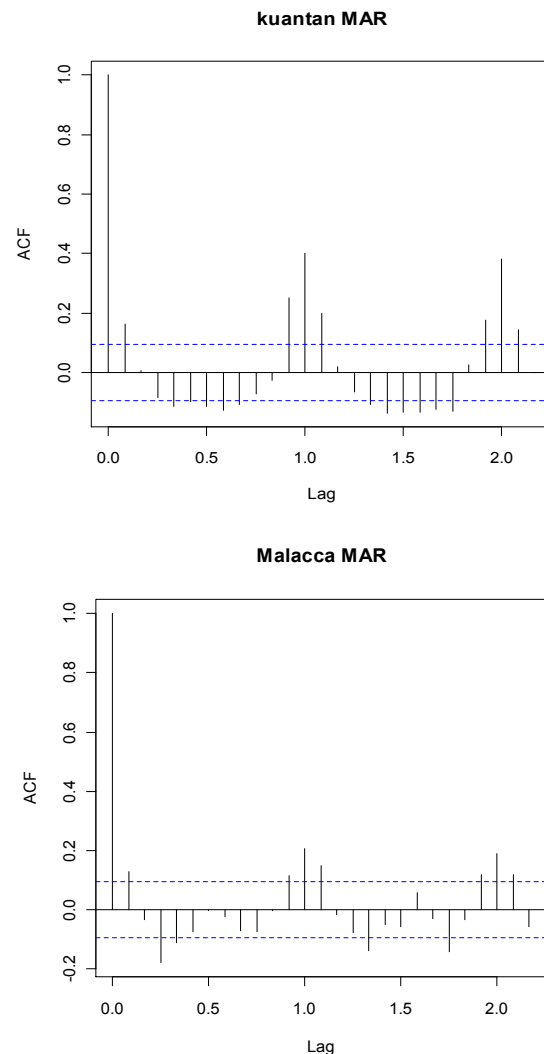
Data and data analysis

Building an ARIMA model

The time-series data analysed in this section correspond to the monthly observations of the homogeneous rainfall datasets for the time period 1968:1 – 2003:12, for Kuantan and Malacca (Tropical rainforest climate) locations in peninsular Malaysia. Any modelling effort on these datasets will have to be based on an understanding of the variability in the rainfall datasets.

The data was examined to check for the most appropriate class of ARIMA processes through selecting the order of the consecutive and seasonal differencing required making the series stationary, as well as specifying the order of the regular and seasonal autoregressive and moving-average polynomials necessary to adequately represent the time series model. The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) are the most important elements of time-series analysis and forecasting. The ACF measures the amount of linear dependence between observations in a time series that are separated by a lag k. The PACF plot helps to determine how many autoregressive terms are necessary to reveal one or more of the following characteristics: time lags where high correlations appear, seasonality of the series, trend either in the mean level or in the variance of the series. The time plots in Fig. 2 upper panel shows that these time series could probably be described using an additive model, since the random fluctuations in the data are roughly constant in size over time. The autocorrelation function (Fig. 1) is a signature of periodicity in the time series. Figure 1 Autocorrelation function of monthly average rainfall data

Figure 1 Autocorrelation function of monthly average rainfall data with a periodicity behavior

kuantan MAR

Malacca MAR

Decomposing Time Series

Decomposing a time series means separating it into its constituent components, which are: usually trend component, irregular component, and if it is a seasonal time series, a seasonal component. Figure 2 Time series decomposition plots

Figure 2 Time series decomposition plots separating trend, seasonal and error components from the observed rainfall series

Decomposition of additive time series

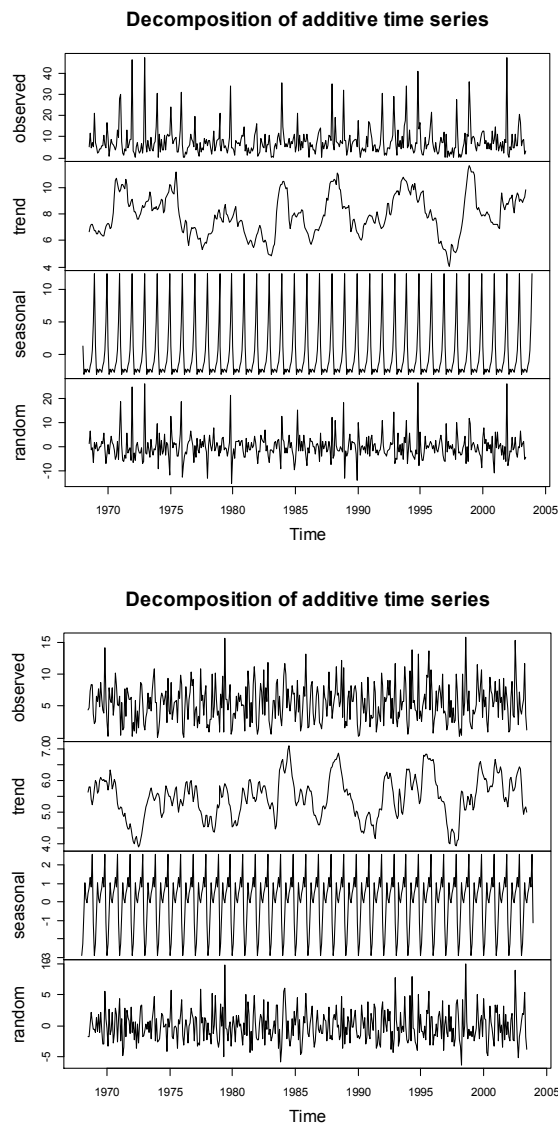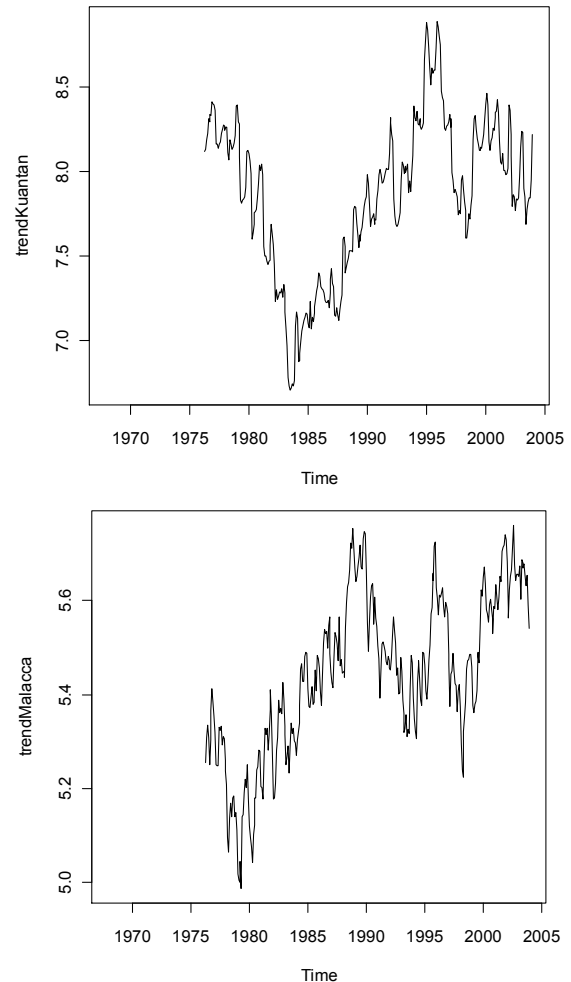Decomposition of additive time series

Figure 2 shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component (bottom). The plots show that the estimated trend component shows no evidence of moving trend. But, by smoothing using a simple moving average of order 100, plots of the smoothed time series gives a clearer picture of the trend component for the rainfall series, and observed that the Kuantan rainfall to have decreased from 1976 to 1984, then increased until 1997. Malacca rainfall trend seems to be increasing since 1979 as seen in Fig.3. Figure 3 plots of the smoothed time series data showing trend movement

Figure 3: plots of the smoothed time series data showimg trend movement

The basic approach to fit a suitable Seasonal ARIMA model is to transform a nonstationary monthly rainfall time series to stationary time series, in this, the time series were decomposed, the seasonal effects were calculated and subtracted from the series rendering the series free of seasonality. The first difference is taken and it is clear that the first difference of both series has zero mean. The orders p and q of the ARIMA models were identified and estimated for both series following Box and Jenkins methodology. Table 2 displays the results of fitted Seasonal ARIMA models. The models were selected based on information criteria.

Table 2: Results of the fitted SARIMA models

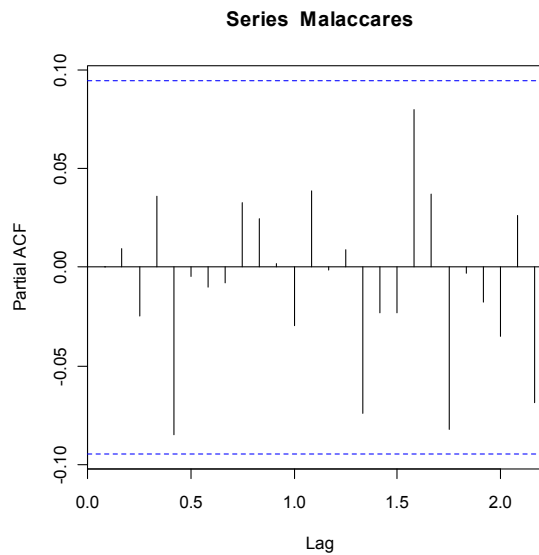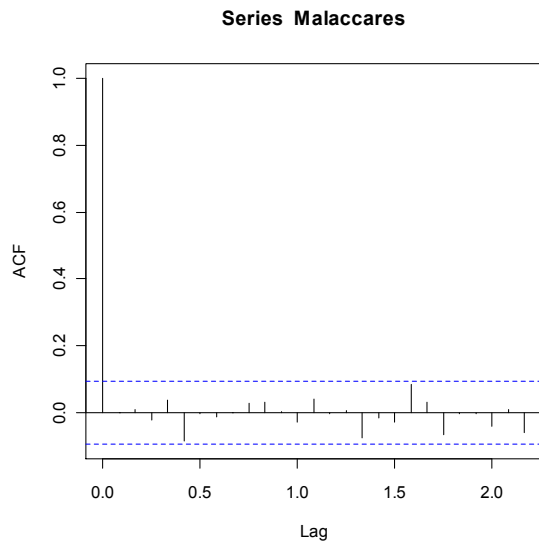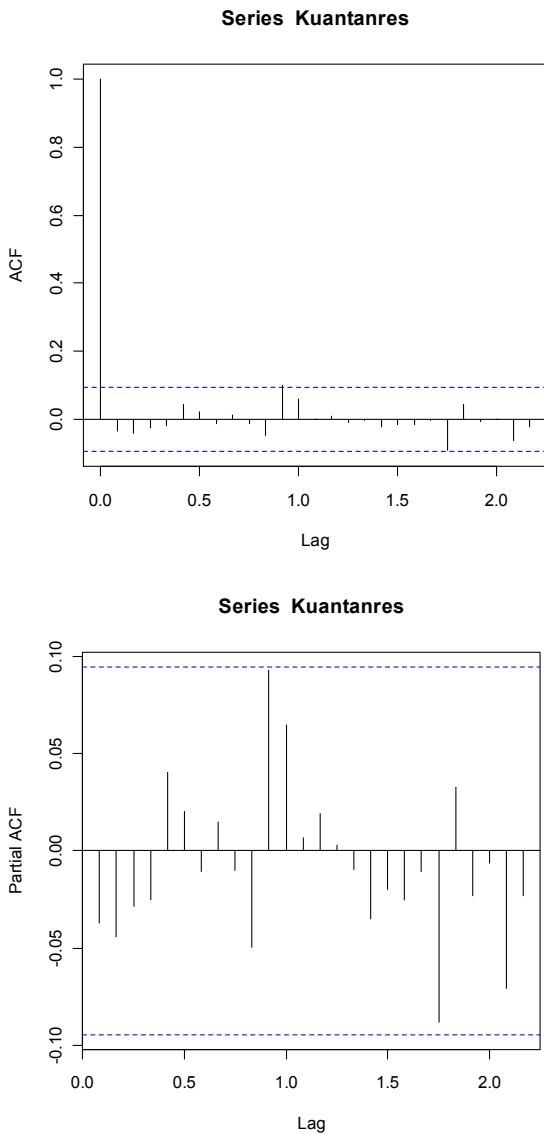|	Malacca		Kuantan	
	SARIMA(4,0,2)(1,0,1)[12]		SARIMA(1,1,2)(1,1,1)[12]	
Parameter	Estimate	Std.error	Estimate	Std.error
AR1	0.0668	0.1254	-0.7481	0.1130
AR2	-0.3044	0.1010	-	-
AR3	-0.0780	0.0485	-	-
AR4	-0.0909	0.0480	-	-
MA1	-0.0684	0.1314	-0.2391	0.1155
MA2	0.3011	0.1017	-0.6812	0.1250
SAR1	0.9818	0.0099	-0.0498	0.0512
SMA1	-0.9376	0.0228	-0.8810	0.0302

In diagnostic checking, the residuals from the fitted SARIMA models were examined against adequacy. This is usually done by correlation analysis through the residual ACF/PACF

functions and the goodness-of-fit test by means of Ljung-Box test. If the residuals are correlated, then the model should be refined. Otherwise, the residuals are white noise and the model is adequate to represent the time series. Residuals from the SARIMA models were tested using Ljung-Box test as shown in Table 3. From the table, the hypothesis that the autocorrelations of the residuals are zero cannot be rejected for both models. This means that the residuals are uncorrelated, indicating that models are adequate, that is fits the series well. The autocorrelation function in (Fig. 4) also confirmed the Ljung-Box test results.

Table 3 Ljung-Box Chi-Square test for residuals from SARIMA models

Lag	SARIMA(4,0,2)(1,0,1)[12]		SARIMA(1,1,2)(1,1,1)[12]	
	Chi-Square	P-value	Chi-Square	P-value
12	10.055	0.611	5.2413	0.9494
24	15.518	0.905	15.7818	0.8958
36	25.957	0.892	40.0501	0.2949
48	34.458	0.929	51.3538	0.3437

Figure 4 Autocorrelation and Patial Autocorrelation Functions of residuals from fitted SARIMA models

Series Kuantanres

Series Kuantanres

Series Malaccares

Series Malaccares

Exponential smoothing methods are useful for making forecasts, and require no assumptions about the correlations between successive values of the time series. However, to make prediction intervals for forecasts made using exponential smoothing methods, the prediction intervals requires that the forecast errors should be uncorrelated and normally distributed with mean zero and constant variance. The exponential smoothing state space models that adequately fit the data in can be described as the ETS (A, A_d, N). The parameters used in generating these models are shown in Table 4. These parameters were chosen to generate data that look reasonably realistic. Clearly, the algorithm has a very high success rate at determining whether the errors should be additive or multiplicative.

Table 4 ETS (A, A_d , N) models parameters

Location	α	β	ϕ	δ	1	b
Malacca	-0.0359	0.0129	0.6497	2.6148	3.9424	2.6148
Kuantan	-0.0955	0.0085	0.9115	5.4617	2.173	0.6668

The sample correlogram in Fig. 5 shows that most sample autocorrelation coefficients of the residuals from the fitted ETS state space models are within the limits upon which the models are based, therefore the residuals are white noise suggesting that the models are adequate following (Ramasubramanian, 2007). To confirm the evidence of no autocorrelations, the Ljung-Box test results in table 5 shows that there is evidence of no autocorrelations in the in-sample forecast errors, and the distribution of forecast errors. This suggests that the simple exponential smoothing method provides an adequate predictive model for Kuantan and Malacca rainfall, which probably cannot be improved upon.

Figure 5 Autocorrelation and Patial Autocorrelation Functions of residuals from fitted State space models
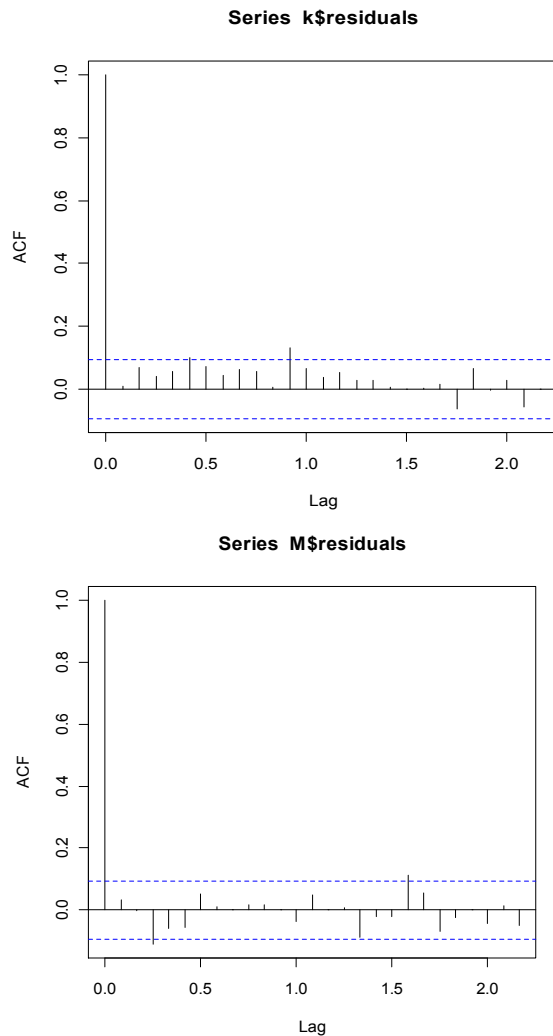
Series k$residuals

Series M$residuals

Table 5 Ljung-Box Chi-Square test for residuals from ETS State space models

Location	Chi-Square	P-value
Kuantan	11.44	0.934
Malacca	22.85	0.297

The p-values above 0.05 indicate the acceptance of the null hypothesis of model adequacy at significance level 0.05 (Wang et al. 2005)

Conclusion

Time-series analysis is an important tool in modelling and forecasting rainfall. Two time series analysis techniques were applied in this study, the state space models based on exponential smoothing and the Box-Jenkins modelling method. Seasonal ARIMA models and ETS state space models were developed. Both models were found to be adequate for forecasting. The diagnostic checking confirms the adequacy of the models. Although, this piece of information was not appropriate to predict the exact monthly average rainfall but could give us information that can help the decision makers establish strategies for proper planning of agriculture, drainage system and other water resource applications in Malacca and Kuantan.

REFERENCE

Anderson, B.D.O and Moore, J.B. 1979. Optimal Filtering. Prentice-Hall, Englewood Cliffs.

Aoki, M. 1987. State Space Modeling of Time Series. Springer-Verlag, Berlin.

Box, G.E.P. and Jenkins, G.M. 1976. Time Series Analysis: Forecasting and Control. Holden-Day, Boca Raton, Fla., USA.

Gardner, Jr. E.S. 1985. Exponential Smoothing: The State of the Art. Journal of Forecasting, 4, 1-28

Hyndman, R. J. 2002. Forecasting based on state space models for exponential smoothing

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D. 2005. Prediction intervals for exponential smoothing using two new classes of state space models. Journal of Forecasting 24: 17–37.

Hyndman, R.J. and Khandakar, Y. 2008. Automatic Time Series Forecasting: The forecast Package for R. Journal of statistical software. http://www.jstatsoft.org/.

Koehn, E. and Navvabi, M. H. 1989. "Economics and social factors in construction." *Cost Engineering*, 31(10), 15-18.

Ljung, G.M. and Box, G.E.P. 1978. On a Measure of a Lack of Fit in Time Series Models. *Biometrika* **65** (2): 297–303. doi:10.1093/biomet/65.2.297

Madsen, H. 2008. Time Series Analysis, London: Chapman & Hall/CRC

Makridakis, S. and Hibon, M. 2000. The M3-Competition: results, conclusions and implications. International Journal of Forecasting 16: 451–476.

Medina, D. C., Findley, S. E. and Seydou, D. 2008. State–Space Forecasting of *Schistosoma haematobium* Time-Series in Niono, Mali. www.plosntds.org. 2(8).

Pegels, C.C. 1969. Exponential Forecasting: Some New Variations." Management Science, 15(5), 311-315.

Ramasubramanian V. 2007. www.lasri.res.ln/.../5.../2-time_series_analysis_22-02-07_revised.pdf.

Runeson, K. G. 1988. "Methodology and method for price-level forecasting in the building industry." *Construction Management and Economics*, 6(1), 49-55.

Taylor, J.W. 2003. Exponential smoothing with a damped multiplicative trend. International journal of Forecasting 19: 715–725.

Wang, W. Van Gelder, P.H.A., Vrijling, J. K. and Ma, J. 2005. Testing and modelling autoregressive conditional heteroskedasticity of streamflow processes, *Nonlin. Processes Geophys.*, 12, 55–66, doi:10.5194/npg-12-55-2005.

****