European Journal of Statistics and Probability
Vol.9, No.1, pp, 1-10, 2021
Print ISSN: 2055-0154(Print),
Online ISSN 2055-0162(Online)

APPLICATION OF DYNAMIC LINEAR MODEL ON DATA CORRESPONDING TO CHRONIC ASTHMA DISEASE

Mohamed M. Shakandli
Faculty of Science, Statistical Department, University of Benghazi, Benghazi Libya

Nuri H. Salem Badi
Faculty of Science, Statistical Department, University of Benghazi, Benghazi Libya

Emails: *Mohamed.Shakandli@uob.edu.ly* [(1)], *nuri.badi@uob.edu.ly* [(2)]

ABSTRACT: *The linear Gaussian state space model, also known as dynamic linear model in the Bayesian literature, has become one of the standard parametric modeling forms with parameters changing over time in the time series analysis. It provides a unified and flexible framework for describing, modeling and forecasting a wide array of time series and other types of longitudinal data. There are several studies which have been concerned with describing the seasonal pattern of admission to hospital for children with asthma, and have also explained the relationship between unexpected medical contacts and the end of the summer holidays. In this paper we are interested to use asthma chronic disease data for constructing a dynamic linear model and investigate the behavior of this model also, we are interested to make one step ahead forecasting.*

KEYWORDS: Time series, State space model, Dynamic linear model, Local level model, Seasonal dynamic model.

INTRODUCTION

In the last century, one of the most important statistical models emerged which was concerned with modeling and forecasting with regards to time series data. This is called the dynamic linear model, as denoted by (DLM). The term dynamic is defined as a change which occurs in such a process due to changes in time. The dynamic linear model was initially designed in research related to aerospace science, and subsequently developed its applications to include several areas such as: engineering, economics, finance and medicine, this topic discussed extensively by Durbin and Koopman (2001) and Shumway and Stoffer (2006). The state space model originated in engineering. However, it has been applied to disciplines such as, statistics, medicine and economics. The state space model contains two equations: measurement equation, $y_t = F_t\beta_t + v_t$ which describes the relationship between observed variable (y_t) and unobserved variable (β_t), and evolution equation $\beta_t = G_t\beta_{t-1} + w_t$ that describes the dynamic of the unobserved state that evolves during time according to a Markov process of order one. v_t and w_t are sequence of independent variables with zero mean and variances V_t, W_t respectively. Note that the state vector β_t may be containing trend and seasonal components. The prior distribution of state is supposed to be a Gaussian with mean vector m_0 and covariance matrix C_0.

The methodology of Gaussian state space model has been widely discussed by Harvey (1989), West and Harrison (1997) and Durbin and Koopman (2001). Under an assumption of Gaussian errors; maximum likelihood estimation (MLE) (usually performed by using Expectation Maximization (EM) algorithm Dempster et al. (1977)) is the most popular approach to infer unknown hyper parameters such as V_t , W_t , C_0, F_t , G_t , which discussed widely by Shumway and Stoffer (1982). In addition, Kalman filtering, (Kalman (1960)) which is defined as a set of recursion equations is a common estimation approach for the states β_t.

In the literature of the time series, the local level model also called a steady model, is defined as a special case of a linear Gaussian state space model, when both the design vector F_t and the evolution matrix G_t equal to one.

State space mode

Dynamic linear models are also named as State space models whose parameters varying with time. These models are described dynamic system by two parallel linked equations. The first one is called an observational equation which describes the linear relationship between a sequence of time series data $\{Y_t\}$ and states $\{\beta_t\}$ at time *t*. The second equation is system equation which describes an evolution in the states between time t and previous time $t-1$. This relationship between both states is linear. The structure of the system equation which is based on Markova in property is where the vectors of states change linearly over time. In addition, a probabilistic approach is used for describing both the available data and the variation of parameters. The class of state space models is deemed to be helpful in regression and time series analysis. They are able to supply a flexible and unified framework to describe and model a wide range of time series and other types of longitudinal data in different disciplines. According to Migon et al. (2005) the DLM can be seen as a generalization of the regression models which enable changes in the parameters values throughout time by the introduction of an equation governing the temporal evolution of regression coefficients. The form of the state space model is defined by the following set of equations:

$$Y_t = F_t'\beta_t + v_t \qquad v_t \sim N(0, V_t) \tag{1}$$

$$\beta_t = G_t\beta_{t-1} + \omega_t \qquad w_t \sim N_p(0, W_t) \tag{2}$$

The initial information D_0 provides to the expert the prior beliefs about state β_0 that has a normal distribution with m_0 mean and variance C_0:

$$\beta_t \sim N_p(m_0, C_0) \tag{3}$$

- Y_t is a m x 1 vector containing a sequence of observations.
- β_t is a $p \times 1$ vector consisting of unknown parameters (states).
- F_t is a $p \times 1$ design matrix consisting of regression variables.

- G_t is a $p \times p$ matrix which describes the evolution in the states.
- w_t is the term of evolution error. It describes the stochastic changes in the unknown parameters; and follows a *p-variate* normal distribution with zero mean and variance W_t.
- v_t is a term of observational error. It represents a measurement and sampling error corrupting the observation of y_t , assumed normally distributed to be distributed with zero mean and variance V_t .

Additionally, the error terms v_t and w_t at any time point t are represent mutually independent white noise. The above equations, which are described in the Gaussian state space model were identified Harrison and Stevens (1971) in the Bayesian approach. The quadruple quantities $\{F_t\ ,\ G_t\ , V_t\ ,\ W_t\}$ are used for completely determining the dynamic linear model. The algorithm of the recursive Kalman filter is utilised for analytically inferring the unknown states when the value of these quantities is known. West and Harrison (1997) presented greater details for this procedure. The analytical approach within dynamic linear models is not available when the hyper-parameters $\{V_t\ ,\ W_t\}$ are not known. This issue can be solved by using several proposals for performing approximate inference in dynamic linear models using Bayesian methods based on simulation approaches, such as Markov chain Monte Carlo (*MCMC*) methods and sequential Monte Carlo (*SMC*) methodologies, also defined as particle filter, Gordon et al. (1993). The dynamic linear models can be employed to model univariate or multivariate time series. In addition, the *ARIMA* process which was defined within the linear time models Box and Jenkins (1976) can be represented as special case of the state space model when assuming the quadruple quantities $\{F_t\ ,\ G_t\ , V_t\ ,\ W_t\}$ do not change with time. The state space model is called a Hidden Markov model when the variables of state are discrete.

ANALYSIS OF ASTHMA CHRONIC DISEASE DATA

Data description

Since we deal with time series for count data, description the data in this paper is useful for understanding the features within it. In this part, we will present a brief description of the data used in this study. This data used in the study by Julious et al. (2011), considered to school-aged children without asthma were employed in the study as controls based on the assumption that viral infection is the essential factor in increasing the number of unscheduled medical contacts after returning to school. Many studies have illustrated that the number of medical visits for children with asthma reach a peak when returning to school after the summer holiday Storr and Lenney (1989), Grech et al. (2004), and Julious et al. (2007). The data were daily medical contacts for school aged children with asthma in England, during a period of seven years from 1999 to 2005. This data was collected and placed in the database of the General Practice Research (GPRD). Because of the difficulty of knowing the behavior of the daily time series and difficulty to observe its dynamic changes by using a graphic description of the data, we resorted to deal with the total weekly

medical visits instead of daily data over seven years. As a result, a dimension of the used time series in this report has been reduced from 2557 daily observations to 365 observations were represent the total weekly cases.

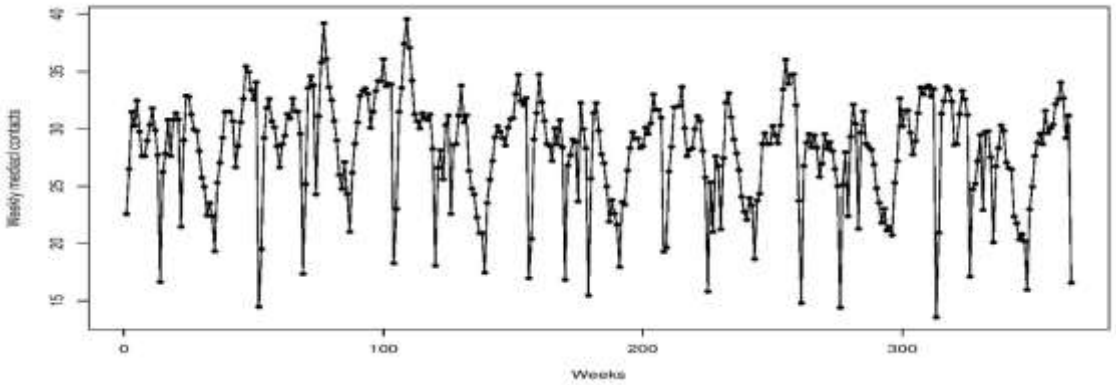

Figure 1: Data description of weekly medical contacts for children with asthma in England

The original weekly data together with time series components are plotted in Figure1. In order to look at the dynamic in the data, we detect from the graph that the data set fluctuates around mean with a clear evidence of evolution over time. However, due to the Christmas day, the last week in each year exhibits more uncertainty around the mean. Another finding is that the weekly medical contacts in each year during seven years have a similar pattern of seasonality, but the level differs. Since we are dealing with the weekly data, we consider the seasonal cycle to be equal to 52 weeks. This means that the pattern of cases in any given week in the first year is expected to be repeating in the corresponding weeks in other years, subject to other sources of variation, such us local level. In order to specify whether there is clearly seasonal variability within the components of the time series, the decomposition of additive time series with frequency 52 is used for as shown in Figure 2.

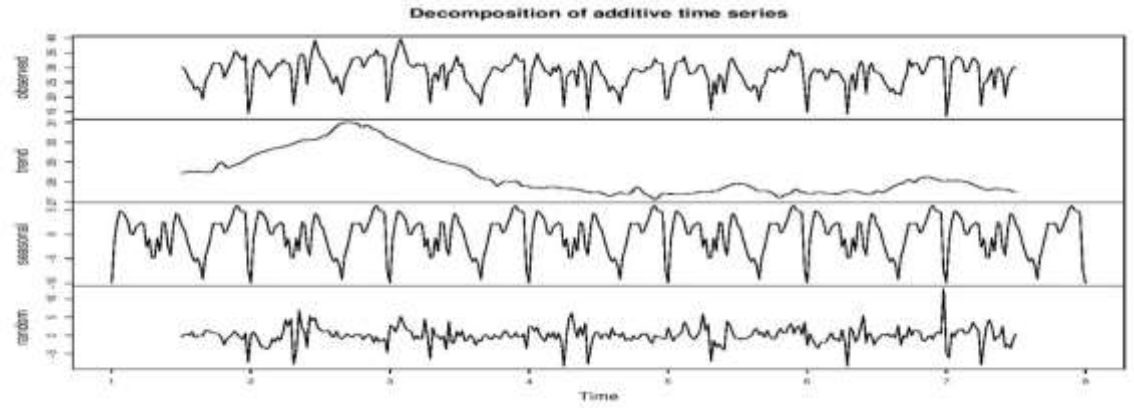

Figure 2: The decomposition of additive time series.

As can be observed from this figure, the pattern of the data repeats the same frequency after 52 weeks. Therefore, we expect to be an annual seasonality in our data.

Results and discussion of the local level model

As mentioned before the local level model is a class of a Gaussian state space form when the design vector and the evolution matrix equal one. In this section we are interested in constructing a dynamic linear model for our data set in order to describe changes in the rate of the medical contacts over time and then apply the recursive kalman filter for forecasting. In order to achieve this goal, the local level model and the seasonal dynamic linear model are proposed models to fit this data. The local level model is the simplest model in univariate time series. The observation y_t is a function of state variable with a single parameter and a stochastic error. The quadruple $\{1, 1, V_t, W_t\}$ are used in the definition of this model. The local level model can be written in the following way:

$$y_t = \beta_t + v_t \ , \qquad v_t \sim N(0, V_t)$$
$$\beta_t = \beta_{t-1} + \omega_t \ , \qquad \omega_t \sim N(0, W_t)$$

For the implementation of the recursive Kalman filter for updating a state from time t − 1 to t, and obtaining one-step forecasting of the observations, we assumed the prior distribution of states to follow a normal distribution with mean zero and variance 1000.The large value of the state variance indicates significant uncertainty in the initial belief of the weekly medical visits for infected children. In addition, the maximum likelihood method has been used to estimate the variance of observation and the transition covariance matrix. The log likelihood is 710.6584. Figure 3, shows the original data (solid points), the one week ahead forecasts based on the local-level model and the recursive Kalman filter (dotted lines with ticks), and the 95% forecast intervals (dashed lines). As we note the vast majority of the original data lies within the predictions interval except for some points over the series. Although, some of the forecast values are quite close to the real data, there are some observations far from the mean of prediction and closer to the interval.

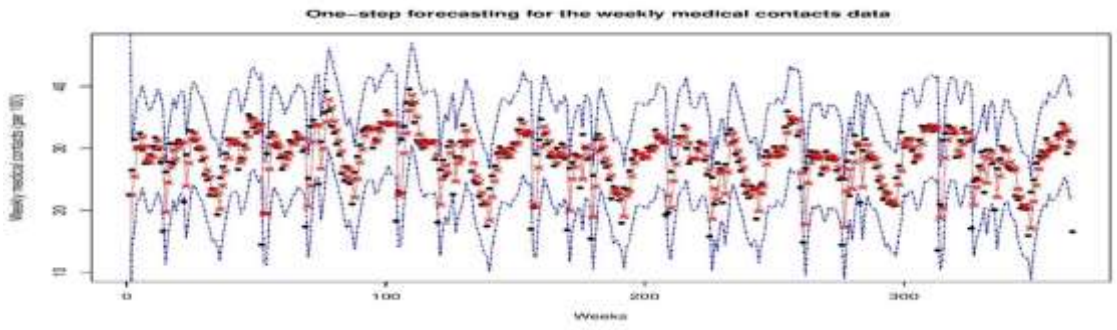

Figure 3: Original data together with one-step forecasts and 95% confidence intervals.

In order to assess the quality and the goodness of fit of the model, we will analyze the residuals that generated by the fitted model. These residuals that obtained from the fitted model will achieve the assumptions of linearity and normality. This can be done by inspected graphically for residuals and by using the measures of the performance of the goodness of fit, such as the mean of squared residuals (MSE) and the mean absolute deviation of the residuals(MAD). If the fit is good, the values of MSE and MAD should be close to zero.

Figure 4, shows the graphic description of the residuals generated from the fitted model. shown are the residuals in panel (a), a histogram of the residuals in panel (b), a normal Q-Q plot in panel (c), and the autocorrelation function in panel (d). From panel (a), we see no obvious patterns, although there are some outliers and this indicates that there is a seasonality within the data. The histogram and the normal QQ in panels (b) and (c) respectively are used to check for normality assumption. From the histogram, we see the distribution of residuals is skewed in, the normal QQ shows that the residuals are quite close to normality except for residuals of the lower tail. The autocorrelation plot in panel (d) which is used to the independence test, displays some peaks at different lags that lie outside the standard bounds $\pm 1.96/\sqrt{n}$; this means the residuals are not independent. The values of the two performance measures MSE and MAD are calculated from the residuals 0.9988 and 0.6883 respectively. Finally, all the above results for the residuals indicate that the level model is not suitable for fitting the data.

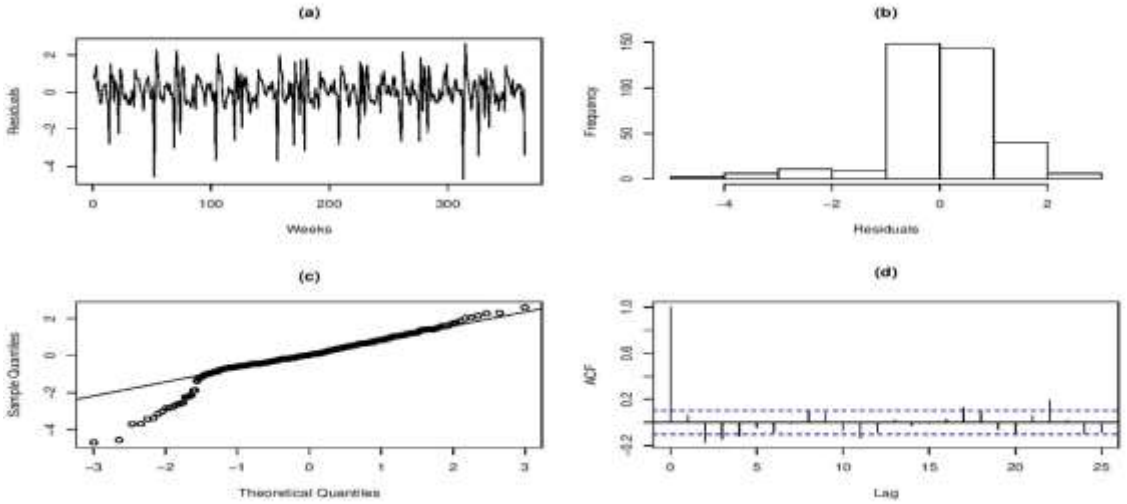

Figure 4: Diagnostics of on the residuals generated from the local level model.

Results and discussion of the seasonal dynamic linear model

A seasonal time series is a time series exhibiting some a seasonal or cyclic pattern. The seasonal components are described by a harmonic component which is defined as a sinusoidal function. The statistical expertise represents a way to choose the harmonic components; moreover, the fundamental harmonic which is defined as the first harmonic

will be necessary for describing the seasonality in the data series. The Fourier series is used to represent the seasonal component within the state space approach. For more details, see Harvey (1989) and West and Harrison (1997). The seasonal dynamic linear model is the second proposed model for this data. As mentioned before the period of the data is 52. To describe the seasonality in the data, we used five harmonic components, because it is difficult to deal with the full seasonal form DLM which need 52 harmonic components for implementing. The observation equation and the transition equation for the model can be written as follows:

$$y_t = \acute{F}_t \beta_t + v_t \quad , \qquad \beta_t = G\,\beta_{t-1} + \omega_t \ ,$$

where the design vector F and the evolution matrix G are defined as:

$F = E_2 = (1,1,0,1,0,1,0,1,0,1,0,1)' = (1, E'_2, E'_2, E'_2, E'_2, E'_2, 1)'$

$$G = J(\omega) = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & J(\pi/26) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & J(5\pi/26) & 0 \\ 0 & 0 & \cdots & 0 & -1 \end{pmatrix}$$

The single harmonic component model can be written as DLM

$$F = E_2 = (1,0)' \quad , \quad G = J(\omega) = \begin{pmatrix} \cos\omega & \sin\omega \\ -sin\omega & cos\omega \end{pmatrix}$$

Where $\omega = \frac{2\pi r}{c}$, $\omega \in (0\ ,\ \pi)$ is defined as a frequency of the time series and rth , r = 1, ..., h is called as the number of harmonic in the model. However, the dynamic linear model with ω = π which is defined as the Nyquist frequency can be written as F = 1, G = −1.

The harmonic component is defined in the evolution matrix by J(φ) which is responsible for the seasonal variation, however, the last element in the transition matrix is -1 which correspond to the Nyquist frequency π. The observation variance and the evolution covariance matrix are the same as before in the first model, however, the distribution of the prior state is a normal $\beta_0 \sim N\,((0,0,0,0,0,0,0,0,0,0,0,0)', 1000I_{12})$ where I is the identity matrix. The log likelihood is 710.6584.

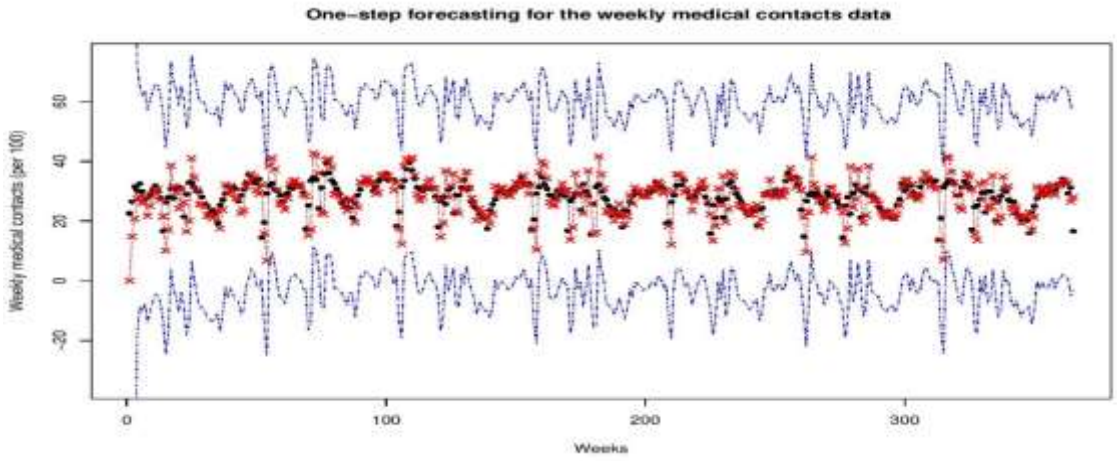

Figure 5: Original data together with one-step forecasts and 95% confidence intervals.

Figure 5, shows the one-step predictions (dotted lines with ticks) with 95% predictive intervals (dashed lines) against the original date (solid points). We can see that all data lies within the predictive interval. Moreover, it can have been observed that there are some observations of the data series that are closed to their forecast mean, others somewhat further away.

The analysis of residuals obtained by the seasonal dynamic linear model with period 52 is used in order to assess the quality and the goodness of fit of the model. The graphic display and the performance measures were utilized to do it.

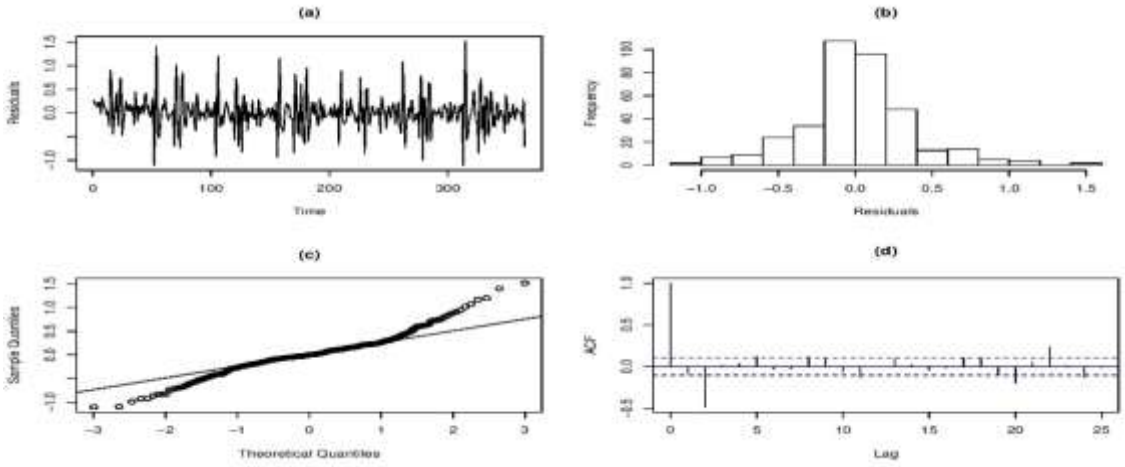

Figure 6: Diagnostics of on the residuals generated from the seasonal dynamic linear model.

Figure 6 presents the graphic description of the residuals generated from the seasonal DLM where the time plot of the residuals, a histogram of the residuals, a normal QQ plot, and the autocorrelation function in are shown panels (a), (b), (c), and(d) respectively.

From panel (a), we see no obvious pattern in the residuals. The histogram and the normal QQ in panels (b) and (c) respectively are used to check for normality assumption. From the histogram, we observe that the distribution of residuals may be approximately normal, also the normal QQ shows that the residuals are quite close to normality except for extreme values in the tails. For the independence test, the autocorrelation function in panel (d) shows some peaks at different lags especially at the second lag which means that the residuals are correlated. The values of the performance measures MSE and MAD are calculated as 0.1324 and 0.2548 respectively.

In order to make a comparison between the two proposed models in terms of quality of forecasting as well as getting a clear vision for the variations within the data series, we looked at the first 106 observations. Figure 7 displays the original data (solid line with circles) together with one-week forecasting by using the local level model (ticks) and the seasonal dynamic linear model (solid points). It can be seen that there are more forecasts close to observations by using the seasonal dynamic linear model than the local level model. Furthermore, the seasonal dynamic linear model has smaller values of the performance measures compared to the local level model.

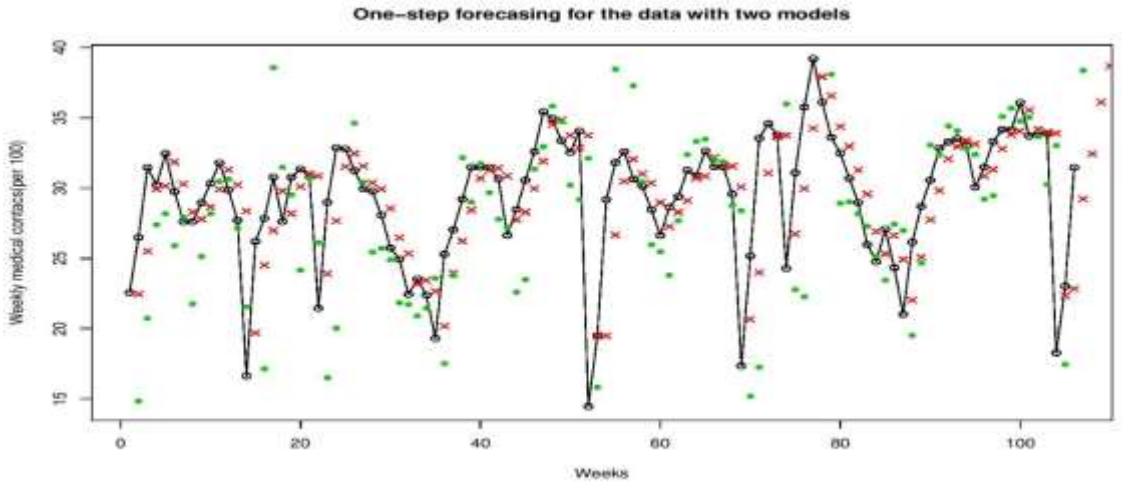

Figure 7: Data against one-week forecasting by using two proposed models.

CONCLUSION

Attention has been directed in this paper to constructing the dynamic linear model for asthma chronic disease data and investigate the behavior of the DLM. The local level model and the seasonal dynamic linear model, with period 52, have been proposed to fit the asthma data. In fact, the results appeared that there are more forecasts close to observations when using the seasonal dynamic linear model and has smaller values of the performance measures. Based on the results of the one-step ahead forecasting, the seasonal dynamic linear model has a good performance than the local level model.

References

Abramowitz, M. and Stegun, I. (1964). *Handbook of mathematical functions*. New-Work: Dover.

Box, G. and Jenkins, G. (1976). *Time Series Analysis*: *Forecasting and Control*. Holden-Day; San Francisco.

Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practise,* Spring- Verlag; New York

Dempster, A., Laird, N. and Rubin, N. (1977). Maximum likelihood estimation from incomplete data via the em algorithm (with discussion). *Roy. Statist. Soc. Ser*., **39**:1-38.

Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods*, Oxford University Press Inc, New York

Gordon, N., Salmond, D. and Smith, A. (1993). Novel approach to nonlinear-nongaussian bayesian state estimation. *IEEE Proceeding*, **140**: 107-113.

Grech, V., Balzan, M. and Distefano, S. (2004). Paediatric wheezy admissions at and around school holiday periods. *Malta Med J,* **16**: 23-26.

Harvey, A. (1989). *Forecasting Structrual Time Series Model and Kalman Filter*.Press, London.

Harrison, P. and Stevens, C. (1971). A bayesian approach to short-term forecasing, *Operations Research Quarterly*, **22**: 341-362.

Julious, S., Campble, M., Bianchi, S. and Murray-Thomas, T. (2011). Seasonality of medical contacts inschool-aged children with asthma: Association with school holiday. *Public Health*, **10.1016**.

Julious, S., Jiwa, M. and Osman, L. (2007). Hospital asthma admissions in school age asthmatics associatedwith the return back to school. *public Health*, **121**: 484-485.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**: 35-45.

Migon, H., Gamerman, D., Lopes, H. and Ferreire, M. (2005). Dynamic models, *In Handbook of Statistics.*

Shumway, R. and Stoffer, D. (1982). An approach to time series smoothing and forecasting using the em alogerithm, *Time series analysis*, **3**: 253-264.

Storr, J. and Lenney, W. (1989). School holidays and admissions with asthma. *Arch Dis Child*, **64**: 103-107.

Shumway, R. and Stoffer, D. (2006). *Time series Analysis and Its Application with R Examples.* Springer.

West, M. and Harrison, P. (1997). *Bayesian Forecasting and Dyanmic Models,* Springer-Verlag, London.

West, M., Harrison, P. and Migon, H. (1985). Dynamic generalized linear models and bayesian forecasting.*American Statistical Association*, **80**, 73-96.