Unravel Before You Travel: A Decision System for Air-ticket Purchase

Parneet Kaur, Prateek Prasanna, Ravneet Arora
Rutgers University

Abstract

Given past 2 years Airline Price Data, we try to predict whether to 'Buy' or 'Not Buy' an airline ticket in some given future time. Given data is modeled using four different approaches. Performance of each of the models is evaluated and compared based on the number of correct decisions made.

1. Introduction

With so many airlines in business these days, getting a good price offer is a profitable and important task for the passengers. Machine learning and statistical models can be of great help in predicting trends in airline prices and thus can assist the passengers in planning their trips and saving lot of money. These models can provide inputs on many decisions like when to buy or not buy a ticket, the future trend in prices (which can follow current series of prices) and the future price of a ticket. In this project we try to fit some different models to the available airline price data and forecast the future trend to help customers in determining whether to 'Buy' or 'Not Buy' a ticket. Forecasted outputs of these models are compared by treating some part of data as training data and rest as test data. Past 2 years airline price data is being extracted from traveling sites like *farecompare.com*. The data represents the lowest price of a ticket recorded in a past time for traveling in some future bucket time. This data is available for 52 cities and between each pair of city. In this project, Auto Regressive Integrated Moving Average (ARIMA), Double Exponential Moving Average (DEMA), Dynamic Linear Model (DLM) and Discrete Time Markov Model (DTM) are used to either predict future prices or analyze the trend. This information is then processed to provide customer the Buy or Not Buy decision. Each model is observed to possess its own property and thus fits best with different kinds of data. The performance of the models is compared by counting the number of correct decisions made by each model.

2. Methods Used

2.1. Autoregressive Integrated Moving Average

An ARIMA model predicts a value in a response time series as a linear combination of its own past values, past errors (also called shocks or innovations), and current and past values of other time series. The ARIMA approach was first popularized by Box and Jenkins [2], and ARIMA models are often referred to as Box-Jenkins models. Using lags and shifts in historical data it uncovers patterns and predicts the future.

A p-th order autoregressive model has the general form

$$Y_t^{[1]} = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \varepsilon_t \quad (1)$$

where
Y_t is the response (dependent) variable at time t.
Y_{t-i} are response variables at time $t-i$.
ϕ_i are the coefficients to be estimated.
ε_t is the error term at time t

The autoregressive model is an all-pole infinite impulse response filter.
A q-th order autoregressive model has the general form

$$Y_t^{[2]} = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - .... - \theta_q \varepsilon_{t-q} \quad (2)$$

Y_t is the response variable at time t.
μ is the constant mean of the process
θ_i are the coefficients to be estimated.
ε_t is the error term at time t
ε_{t-i} are the errors in the previous time periods that are incorporated in the response Y_t

The moving average filter is a finite impulse response filter.
ARIMA is the generalized form of ARMA model which has the form.

$$Y_t = Y_t^{[1]} + Y_t^{[2]} \quad (3)$$

The error terms are usually independent identically-distributed random variables (i.i.d.) sampled from a normal distribution with zero mean. The model is applied to stationary series. There is an initial differencing step corresponding to the "integrated" part of the model which can be applied to remove the non-stationarity. ARIMA models are the most general class of models for forecasting a time series. A correct ARIMA model requires identification of the right number of lags [4] and the coefficients that need to be used. A non-stationary time series is identified by how fast the Autocorrelation Function(ACF) plot decays to zero. If it stays positive out to a large number of lags, it is non-stationary and differencing is required. Process of differencing is continued till the lag-1 ACF goes negative, taking care that the standard deviation of the corresponding differenced series doesn't rise in between. Once the difference order d is identified, the Partial Autocorrelation Function (PACF) and ACF of the d-order differenced series is analysed. The lag at which PACF shows a sharp cut-off is p. Similarly, lag at which ACF shows a sharp cut-off is q. At times ,we observe that there might betwo or more values p and q i.e, the plots show sharp cut-off at multiple points. When such a case arises, we need to check the (Akaike Information Criterion) AIC score of the corresponding model. AIC is a measure of the amount of information lost.

$$AIC = 2k - 2ln(L) \tag{4}$$

where *k* is the number of parameters and *L* is the maximized value of the likelihood function of the estimated model.

Thus, lower the AIC score, better is the model. So to carry out a n-step ahead prediction, we select that model which has the lowest corresponding AIC score. Using the parameters obtained, we proceed with the prediction. For this purpose, we divide our data into *training phase* and *validation phase*. The training set consists of 120-day data and the validation can be either for 15 or 30 days.

2.2. Double Exponential Moving Average

Moving averages are used to smooth out the noise in the data and find the underlying trend in the data. They have been used in stock market to make the decisions so as to maximize users profit. Traditional approaches include Simple Moving Average (SMA) and Exponential Moving Averages(EMA). Double Exponential Moving Average (DEMA) was introduced by Mulloy in [6] for trend analysis. it is also known as End Point Moving Average as described in [5]. It is based on least square fit. For a given set of points, an equation of a straight line can be obtained using least least squares technique such that the sum of the squares of all the distances between the set of points and the straight line is minimum. The relationship between the current price and the last value of the least square fit can be useful in determining if the current price value is higher, lower or same as the best fit line. We can compute the best-fit line each day from previous 'N' days and plot the most recent value obtained each day as a trend line. This trend-line is called end point moving average or double exponential moving average.

EMA uses 'forgetting-factor' λ to have mean of the weighted sum of the samples as the predicted value.

$$a_n = \frac{y_n + \lambda y_{n-1} + \lambda^2 y_{n-2} + ... + \lambda^n y_0}{1 + \lambda + \lambda^2 + ... + \lambda^n} \tag{5}$$

If a window of past N days is considered, equivalent forgetting factor $\lambda = \frac{N-1}{N+1}$. Also, gain factor is defined as $\alpha = 1 - \lambda$. Then, EMA is given by:

$$a_n^{[1]} = \lambda a_{n-1}^{[1]} + \alpha y_n \tag{6}$$

DEMA is obtained by further smoothening the output of EMA. This is done by applying EMA once again to the output obtained in equation 6.

$$a_n^{[2]} = \lambda a_{n-1}^{[2]} + \alpha a_n^{[1]} \tag{7}$$

DEMA indicator is then given by:

$$a_n = 2a_n^{[2]} - a_n^{[1]} \tag{8}$$

Once $DEMA$ trend-line is obtained, following rules are applied to make a decision 'Buy' or 'Not Buy' on a day 'x':
1. If 'N' days $DEMA$ moves down and current price is less than $DEMA$, then 'Buy'.
2. If 'N' days DEMA moves up and (current price$(x-1) <$ $DEMA(x-1)$ and price$(x) > DEMA(x)$), then 'Buy'.
3. If 'N' days DEMA neither moves nor down and current price is less than DEMA, then 'Buy'.
4. If none of the above three conditions are met, decision 'Not Buy' is made.

2.3. Dynamic Linear Models

2.3.1 State Space Models (SSM)

State Space Models treats the time series as the system whose output is affected by various kind of components such as random noise, trend and seasonal components [7]. They have special structure that allows application of various recursive algorithms for the problems such as estimation and forecasting. State Space Models can be applied to both univariate and multivariate time series and can easily handle non-stationary data. In SSM the states of the time-series are assumed to hold Markovian property. It means that a particular hidden state S_t depends only on the state S_{t-1} and is independent of all the earlier states. This can be written as:

$$p(S_t \mid S_{1:t-1}) = p(S_t \mid S_{t-1}) \tag{9}$$

And also a particular observational value Y_t depends on the corresponding state only and thus all the observational values are independent of each other.

$$p(Y_t \mid S_{0:t-1}, Y_{1:t-1}) = p(Y_t \mid S_t) \tag{10}$$

State can be some hidden quantity which mostly remains constant. For example, in airline price data, state can be taken as the expected average price of airline ticket and observational value is the actual price which is visible after adding components such as random noise, long-term trend etc.

2.3.2 Dynamic Linear Models

DLM is an important class of State Space Models. A DLM is specified by equations:

$$Y_t = F_t S_t + v_t, \quad v_t \sim \mathcal{N}_q(0, V_t)$$

$$S_t = G_t S_{t-1} + w_t, \quad w_t \sim \mathcal{N}_p(0, W_t)$$

and prior

$$S_0 \sim \mathcal{N}_p(m_0, C_0)$$

where, p is the dimension of a state, q is the dimension of observational value, G_t and F_t are matrices (of size pxp and qxp, respectively) and v_t and w_t are sequences of Gaussian distributions with mean 0 and variance matrices V_t and W_t. For a particular DLM usually F and G are fixed for a particular model which means they are independent of time. Also, one can initialize the m_0 and C_0 for a particular set of data. Thus, the parameters which need to be estimated are V and W.

2.3.3 Forecasting Using DLM

DLM has special property which allows the use of recursive procedures for estimation of the forecasts. For DLM, recursive procedure is applied to one-step ahead prediction to get the k-step ahead predictions. k-step ahead prediction for state S_t is denoted by $S_{t+k} \mid y_{1:t}$ and is defined as normal distribution $\mathcal{N}(b_t(k), R_t(k))$ and, k-step ahead prediction for observational value Y_t is denoted by $Y_{t+k} \mid y_{1:t}$ is defined by normal distribution $\mathcal{N}(l_t(k), Q_t(k))$ The recursive procedure to calculate k-step prediction using 1-step ahead prediction is as follows as explained in [3]:
Let $b_t(0) = m_t$ and $R_{t_0} = C_t$, then distribution for S_{t+k} is

$$\mathcal{N}(b_t(k), R_t(k))$$

where, $b_t(k) = Gb_t(k-1)$,
$R_t(k) = GR_t(k-1)G^T + W$
Distribution for Y_{t+k} is:

$$\mathcal{N}(l_t(k), Q_t(k))$$

where, $l_t(k) = Fb_t(k)$,
$Q_t(k) = FR_t(k)F^T + V$

2.3.4 Application of DLM on Airline Price Data

Due to additive structure of DLM several component models can be summed together. Each model is associated with a particular characteristic of the data such as periodicity, random noise and linear trend. For the airline price data 2 models namely, Local Level Model and Linear Growth model have been applied.
1. Local Level Model (LLM) is a very simple model for univariate time-series. It is also known as random walk plus noise model. This model captures random noise which is inherent in the data. Equation for LLM is:

$$Y_t = u_t + v_t, \quad v_t \sim \mathcal{N}(0, V)$$

$$u_t = u_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, W)$$

W and V are of single dimension in this particular model. Also $F = G = 1$ and $S_t = u_t$. This model is applicable to the time-series for which there isn't any particular pattern or trend visible. This model seems very appropriate for the airline price time-series data as lot of random patterns are present in the plot of the data.
2. Linear Growth Model (LGM) is used to capture some existing trend present in the time-series. This model is defined by the equations:

$$Y_t = S_t + v_t, \quad v_t \sim \mathcal{N}(0, V)$$

$$u_t = u_{t-1} + H_{t-1} + w_{t,1}, \quad w_{t,1} \sim \mathcal{N}(0, \sigma_1^2)$$

$$H_t = H_{t-1} + w_{t,2}, \quad w_{t,2} \sim \mathcal{N}(0, \sigma_2^2)$$

$$S_t = \begin{bmatrix} u_t \\ H_t \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$W = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, \qquad F = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

The term H stands for the slope of the curve. Linear Growth Model plays a very important part in determining 'Buy' or 'Not Buy' decisions as it forecasts the trend which follows the currently available data.

2.4. Discrete Time Markov Chains

The airline data can be modelled as a Discrete Time Markov Chain(DTMC) problem. In order to do this . we first identify states in the given data. This can be done by dividing the entire price range into slabs of $20 each. This would result in *n* number of states 1,2,....n. Once this is done, we calculate the frequency of inter-state transition and collect the data as a transition frequency matrix . Then we construct the transition probability matrix by dividing each transition frequency by the row sum of the frequencies. Here we assume that all transitions are possible. Once we have the transition probability matrix for the Markov

Chain[8], we can compute its steady state distribution by solving the equation

$$\pi' = \pi^{'} P \tag{11}$$

where π is the limiting state probability vector

$$\pi = lim_{n\to\infty}\mathbf{p}(n) \tag{12}$$

The one-step transition matrix can be represented by P which is a square stochastic matrix with non-negative elements and unit row sums. It is of the form

$$P = \begin{bmatrix} P_{00} & P_{01} & . & . & P_{0K} \\ P_{10} & P_{11} & & . & P_{1K} \\ . & . & . & . & . \\ . & . & . & . & . \\ P_{K0} & P_{K1} & . & . & P_{KK} \end{bmatrix}$$

We can find the n-step transition matrix as

$$\mathbf{P}(n) = \mathbf{P}^n \tag{13}$$

The transition probability matrix is essentially a sparse matrix, since we do not observe transition between a large number of states. When such a matrix, with unknown entries as zero, is used to calculate the steady state probability distribution, we might arrive at a wrong conclusion and infer that the ticket will not attain a state to which we did not observe any transition. Thus we need to refine our data by estimating the missing frequencies and used the modified transition frequency matrix to calculate the transition probability matrix. One method of estimating missing frequencies from a given set of observation is the Good-Turing Frequency Estimation method [1] .

The steps of calculating the missing frequencies are:

1.Calculate a table of frequency of frequencies versus frequency. If r denotes a frequency, N_r will denote the number of time the frequency is observed.

2. The total number of objects observed is thus

$$N = \sum rN_r \tag{14}$$

3. The total probability of all unobserved objects is estimated as

$$p_0 = \frac{N_1}{N} \tag{15}$$

Where N_1 is the frequency of the frequency 1

4. The modified frequencies will then be

$$r^* = (r+1)\frac{E(N_{r+1})}{E(N_r)} \tag{16}$$

In most cases, smaller values of r will have high frequencies and thus E(N_r) can be replaced by the corresponding value N_r. For larger values of r, we need to approximate E(N_r) by a smoothed value S(N_r). In addition to this, the highest frequency will always have a estimated zero value. We chose to use a linear smoothing function to avoid complexity. The smoothing function is computed by constructing a linear least square estimator of the form

$$N_r = ar + b \tag{17}$$

Finally, this value of N_r is used to calculate r^*.

3. Evaluation and Discussion

If a decision is to be made on day 'x', previous 120 days of data is taken as training set and a model is fit. Then based on the predicted prices or trend observed on that day, decision 'Buy' or 'Not Buy' is made. For evaluating the performance of the models, the actual values of prices available for next 30 days are used to make a decision that would have been beneficial. Both the decisions are then compared.

For **ARIMA**, on day x, a model is fit by selecting the values of p,d and q for the lowest AIC score as explained in section 2.1. The predictions are then made for next 30 days. Then, on day 'x', if the predicted value in next 30 days goes lower than the current price on day 'x', decision 'Not Buy' is made. If the price increases or remains constant in next 30 days, decision 'Buy' is made. If decision is 'Not Buy' then next day again the same procedure is repeated, and this continues till a definitive decision 'Buy' is made. By doing so, in a real time scenario, every new information available is incorporated for making a decision on a particular day. For city pair CMH-DFW, prices from day 550-670 is taken to forecast prices from 670-700. Similarly training data from range 580-700, 610-730 and 640-760 are taken to forecast prices for 700-730, 730-760 and 760-766 range respectively. The predictions are as shown in figure1, the predicted prices do not match the actual prices, but the overall trend is followed. This facilitates taking an appropriate decision.

This approach works considerably well as evident from the results. It is worth noting that though the results are fairly good, at times the order of the model becomes high. In such cases, the complexity of the calculation increases. Ideally, the order of the mode (p and q) should be restricted to a low value. We also see that there isn't a single best-fit for a given data set. Different ARIMA models need to be fit to different parts of the data before making the corresponding n-step ahead prediction.

ARIMA model makes a good prediction if the value of 'n' is small. To analyze this, 15 and 30 day prediction is done as shown in figure2. It is observed that 30-day price prediction fails to follow the trend of prices sometimes,

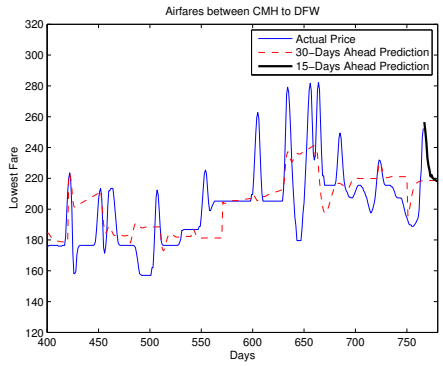

Figure 1. N-step ahead prediction using ARIMA for city pair CMH-DFW

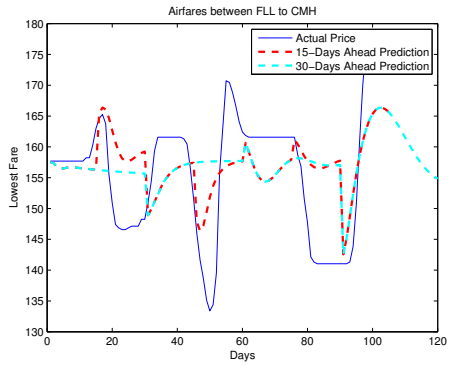

Figure 2. 15 and 30-step ahead prediction using ARIMA for city pair FLL-CMH

whereas 15 day price prediction is able to follow it. So, the decision is more accurate if the ticket purchase is to be made in a small time frame from a day 'x'.

For **DEMA**, $N = 200$ is selected, which means that a moving average on a particular day is computed based on previous 200 days. The decision making is as described in section 2.2. Using this approach we can't tell the customer when to buy a ticket in future. This technique aims at telling if or not the price is reasonably good today.

For **Local Level Model**, training data for 120 days and testing data of 30 days is selected. Thus, forecasted data can be compared with the actual time-series. Graph for city pair BWI-CVG is given at 4. From the graphs it is clear that LLM is able to predict small changes in the data, but it cannot foresee some very abrupt changes appearing in the form of peaks and valleys. Firstly data from day 550-670 is taken to forecast prices from 670-700. Similarly training data from range 580-700, 610-730 and 640-760 are taken to forecast prices for 700-730, 730-760 and 760-766 range respectively.

For **Linear Growth Model**, since a trend depends on

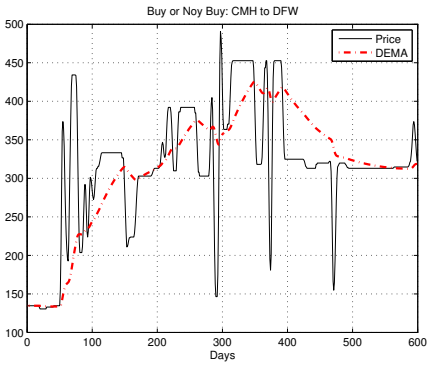

Figure 3. Decision Making based on Double Exponential Moving Average for CMH-DFW

Table 1. Decision making using ARIMA

	Buy	Not Buy
Predicted Buy	26	5
Predicted Buy	12	33

the whole range of data available, firstly data from 1-300 is taken to predict the trend for period 300-400. Similarly training data from 1-400, 1-500 and 1-600 is taken to predict the trend from 400-500, 500-600 and 600-700 respectively. As shown in figure 5, linear growth model is quiet successful in predicting long range trend for the series. But again, the sudden peaks and changes in data presents problem for this model.

For **Discrete Time Markov Model Approach**, the airline data from CMH to DFW was considered. The training data consisted of the first 650 values. The price range was divided into 10 states , each of \$20 . The transition frequency matrix was calculated after which Good-Turing estimation was applied to estimate the missing frequencies. After obtaining the modified transition frequency matrix, each element was divided by the corresponding row sum to obtain the state probability matrix. Suppose we need to calculate the 10-step ahead prediction, the obtained matrix P was raised to the power 10. The resultant matrix was the 10-step transition matrix. The last state occupied in the training data was state 6. When the obtained matrix P was raised to the power 10, the fourth element corresponding to the sixth row had the maximum probability in that row (0.2386). Thus, given that the Markov chain was in state 6 in the beginning, it is most probable to occupy state 4 after 10 transitions.

The decision making results obtained from ARIMA, DEMA and DLM(LLM) are tabulated in tables 1, 2 and 3. The accuracy of the models can be determined from the tables.

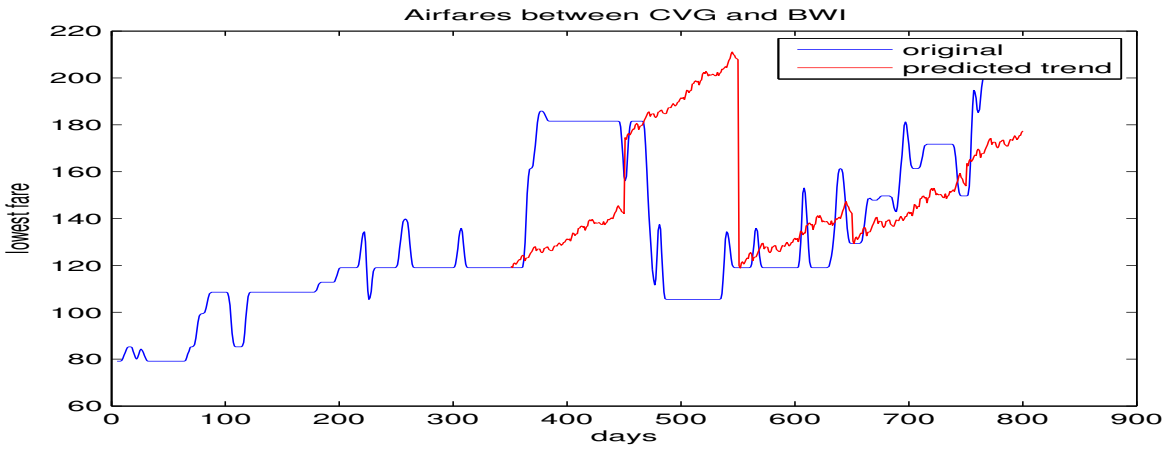

Figure 5. N-step ahead prediction using Linear Growth Model for city pair BWI-CVG

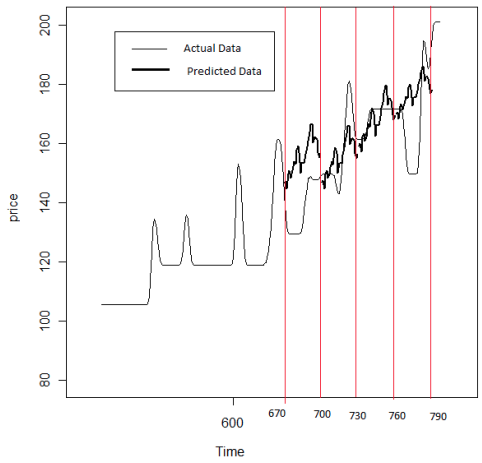

Figure 4. N-step ahead prediction using Local Level Model for city pair BWI-CVG

Table 2. Decision making using DEMA

	Buy	Not Buy
Predicted Buy	43	2
Predicted Buy	13	19

Table 3. Decision making using DLM(LLM)

	Buy	Not Buy
Predicted Buy	16	7
Predicted Buy	5	12

4. Conclusion and Future Work

From the results obtained we observe that ARIMA performs well for short term prediction and it is computationally heavy. However, since ARIMA can't handle stationary data, it is better to use DLM. Accuracy of DLM can be increased by adding more complex components or models. Though the Markov model doesn't provide a concrete decision making criteria, it lets the customer know what is the probability of attaining a particular price in future. Thus, it is more of a probabilistic approach rather than determinists approach. DEMA works well if the decision to buy or not buy has to be made today. However, it does not tell what time in future will the customer get the best price and how good the price will be.

In the proposed models we have taken only the history of airfare prices. As an extension of this project, it will be a good idea to analyze how airfares are affected by other factors such as oil prices, seasons, distance, seasonality and popularity of destination and incorporate their effects.

References

[1] W. A. Gale and G. Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*. 4

[2] G. R. George Box, Gwilym M. Jenkins. *Time Series Analysis: Forecasting Control*. Prentice Hall, 3rd edition, 1994. 1

[3] P. C. Giovanni Petris, Sonia Petrone. *Dynamic Linear Models with R*. Springer, 2009. 3

[4] http://www.duke.edu/ rnau/411arim.htm. Website accessed 01 may 2011. 2

[5] P. E. Lafferty. The end point moving average. *Technical Analysis of Stocks Commodities*, pages 413–417, October 1995. 2

[6] P. G. Mulloy. Smoothing data with faster moving averages. *Stocks Commodities*. 2

[7] G. Petris. dlm: an r package for bayesian analysis of dynamic linear models. 2

[8] D. J. G. Roy D. Yates. *Probability and Stochastic Processes*. John Wiley Sons, 2 edition, 2005. 4