ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

information
systems

Business Intelligence

– SEMINAR WINTER SEMESTER 2013/2014 –

Autoregressive and Moving Average Processes

– SEMINAR PAPER –

Submitted by:
Nicole Nadine Ludwig

Student-ID:

Advisor:
Prof. Dr. Dirk Neumann

Contents

1 Introduction

Most economic data is measured over time. The GDP, stock market prices or unemployment statistics are just a few examples. A tool to understand the processes that generate such data is time series analysis, which is concerned with analysing the dependencies between the different observations. It is especially interesting to understand the data generating process to make better predictions of future values. This is for example important to companies that operate on the stock market and want to improve their revenues. For this analysis, certain stochastic models have been developed for discrete data.
This paper aims to introduce to basic stochastic models of time series analysis and applies these models to data from the West Texas Intermediate crude oil price and intraday data from the European Energy Exchange Market. Section 2 first takes a look at general properties of time series. Afterwards models to describe the stochastic process are discussed (section 3). Section 4 is concerned with typical deterministic components such as trend and seasonality and possibilities to handle such patterns. Section 5 applies the previously developed tools on the given data with the help of the statistic software "R".

2 Properties of Time Series Processes

The primary aim of time series analysis is probably forecasting. But to do so, one needs to identify the stochastic process which underlies the time series. This section will introduce to basic properties of time series which will be needed later to identify and specify the time series model. To start with the concept of a time series, consider the following model

$$x_t = s_t + m_t + \varepsilon_t,$$

where x_t is the current value of the time series. It depends on deterministic components, such as the seasonal component s_t and the trend component m_t. Furthermore, it also depends on the stochastic part of the time series ε_t. Section 2 and section 3 outline possible paths to model this stochastic part.

2.1 Stationarity

To be able to characterize the stochastic component of a time series, it is necessary to introduce to the concept of stationarity. Although most economic time series have a trend and are thus not stationary, they are based on stationary time series or rather can be transformed so that their residuals are stationary. A time series is considered *strictly stationary* if the stochastic process does not change by shifting the time. That is

$$\mathrm{P}\{x_{t_1} \leq c_1, \ldots, x_{t_k} \leq c_k\} = \mathrm{P}\left\{x_{t_{1+h}} \leq c_1, \cdots, x_{t_{k+h}} \leq c_k\right\},$$

for all $k = 1, 2, \ldots$, all time points $t_1, t_2, \ldots, t_k$, all numbers $c_1, c_2, \ldots, c_k$, and all time shifts $h = 0, \pm 1, \pm 2, \ldots$ (Shumway and Stoffer 2006, p. 23).

These are very strict assumptions and they are not practicable for most time series. Hence, this paper will regard a time series as stationary if it is *weakly (or second order) stationary*, that is for all t given by

$$\mu = \mathrm{E}(x_t), \tag{1}$$

$$\mathrm{Var}(x_t) = \mathrm{E}[(x_t - \mathrm{E}(x_t))^2] = \sigma^2 < \infty, \tag{2}$$

$$\gamma_\tau = \mathrm{E}[(x_t - \mu)(x_{t-\tau} - \mu)] = \mathrm{E}[(x_s - \mu)(x_{s-\tau} - \mu)] \quad \forall t \neq s. \tag{3}$$

Where equation (1) means that the mean value function is constant, equation (2) that the variance is finite and equation (3) that the different values of the time series are only dependent on the interval between t and s but not on the actual point in time. The difference between a random sample and a time series is that the realizations of the time series are not independent from each other. Thus, to characterize this dependency, one can use the autocovariance and the autocorrelation. The autocovariance function γ_τ gives information about the linear dependence of two different observations of the time series. From γ_τ, one can derive the more important *autocorrelation function* (ACF)

$$\rho_\tau = \frac{\mathrm{Cov}(x_t, x_{t-\tau})}{\sigma_{x_t}\sigma_{x_{t-\tau}}} = \frac{\gamma_\tau}{\gamma_0} \quad \in [-1, +1],$$

which also shows this dependency. One can further derive the partial autocorrelation function, which indicates how much additional explanatory contribution π_τ gives when π_1 to $\pi_{\tau-1}$ are also given

$$x_t = \alpha_0 + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \underbrace{\alpha_\tau x_{t-\tau}}_{\pi_\tau} + w_t.$$

2.2 Augmented Dickey-Fuller Test

Judging the stationarity condition by appearance of the time series can be helpful, but it is more efficient to rely on a statistic test. One important test was introduced by Dickey and Fuller (1979). In this test, the null hypothesis is the time series has a unit root and the alternative hypothesis is the time series is stationary. Because the Dickey-Fuller (DF) test relies on the assumption of non autocorrelated errors, it has been expanded to the augmented Dickey-Fuller (ADF) test, which includes the lagged changes Δx_{t-i} into the regression. The idea is to add as many lags k of Δx_{t-i} until w_t is not autocorrelated anymore. The test then is applied to

$$\Delta x_t = \alpha + m_t + \phi x_{t-1} + \sum_{i=1}^{k} \delta_i \Delta x_{t-i} + w_t,$$

where α is a constant, m_t a trend component and w_t a white noise process. The ADF test is basically the same as the standard DF test for $k = 0$. As the test statistic is not normally distributed the test decision is based on the critical values tabulated by Dickey and Fuller. The difficulty of the ADF test is to decide how many lags one should include into the regression. As the test statistic will be incorrect for too few lags and loosing power for too many lags (Wooldridge 2013).

2.3 Non-Stationarity

If the null hypothesis of the ADF test in the previous section cannot be rejected, we have a non stationary time series. In fact most economic time series are not level stationary because they have a trend. Hence, it is also important to take a look at non-stationary time series. The simplest form of a non-stationary process is a random walk

$$x_t = x_{t-1} + w_t,$$

where $w_t \ \ (\sim N(0, \sigma^2))$ is white noise. This process has a stationary mean $\mathrm{E}(x_t \mid x_0) = x_0$ but its variance does not converge and the process is thus not stationary. This non-stationarity can be solved by calculating first differences

$$x_t - x_{t-1} = \Delta x_t = w_t.$$

The resulting time series is now a white noise process and is called integrated of order one $\mathrm{I}(1)$. In general a time series is called integrated of order d, i. e. $\mathrm{I}(d)$ for short, if it is stationary (and thus $\mathrm{I}(0)$), after building the d-th difference.

3 Models for Stationary Time Series

There is a wide range of models that can describe a time series. The basic ones will be considered here. First the autoregressive model (AR, see section 3.1), second the moving-average model (MA, see section 3.2) and finally the mixed models (ARMA/ ARIMA, see section 3.3). They all require the data-generating process to be linear and to have a constant variance (homoscedasticity).

3.1 AR Processes

A very useful model for stationary time series is the *autoregressive* model, where the stationary process is described only in terms of its previous values and a white noise term, i. e.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t,$$

or in terms of the backward shift operator B by

$$\begin{aligned}(1 - \phi_1 \mathrm{B} - \phi_2 B^2 - \ldots - \phi_p \mathrm{B}^p) x_t &= w_t, \\ \phi(\mathrm{B}) x_t &= w_t,\end{aligned}$$

where ϕ_τ is the AR parameter which indicates how strongly the past value $x_{t-\tau}$ influences the present value x_t and w_t represents a white-noise process with expected value $\mathrm{E}(w_t) = 0$ and variance $\mathrm{Var}(w_t) = \sigma^2$ (Box, Jenkins, and Reinsel 2008). The white-noise includes everything unsystematic that is not explained by the past values (e. g. a random shock). The order of the process is determined by p, because only the p past values influence today's value. Generally, the process is called $\mathrm{AR}(p)$, if it is of order p.
As the AR process is basically a regression on the processes own past values the parameters can be estimated with an ordinary least squares estimator. There are two special cases of the AR

parameter. The first is the case that the process has a unit root, and the second is the case where $\phi = 0$ and the pure white noise process is left. The latter case can be tested with a Portmanteau-Test, e. g. the test by Ljung and Box (1978) (the test will be discussed further in section 3.4). Whether the process has a unit root can be tested with e. g. the augmented Dickey-Fuller test (see section 2.2). The process has no unit root if all roots $\phi(\mathrm{B}) = 0$ lie outside the unit circle and is then stationary.
The second moments of the process are giving valuable information over the type and order of the model. For an AR model, the autocorrelation function is infinite and tails off. It is coming closer to zero after the first p lags, either in smooth exponentials or sine waves. Whereas the partial autocorrelation function is finite and cuts of at lag $p+1$, because only the past p values give information about the process. Hence, taking a closer look at the PACF can help to determine the order p of the model.

3.2 MA Processes

Another way to describe the stochastic process underlying a time series can be the *moving average* process. Here, the current value is described in terms of a weighted average over the q past error terms, but the weights do not necessarily sum themselves up to one. The model is defined as

$$x_t = w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2} - \ldots - \theta_q w_{t-q},$$

or in terms of the backward shift operator B by

$$\begin{aligned} x_t &= (1 - \theta_1 \mathrm{B} - \theta_2 \mathrm{B}^2 - \ldots - \theta_q \mathrm{B}^q) w_t, \\ x_t &= \theta(\mathrm{B}) w_t, \end{aligned}$$

where θ_τ is the MA parameter and w_t is Gaussian white noise (Box, Jenkins, and Reinsel 2008). The order of the process is determined by q, because only the past q error terms are determining the current value. Generally a process is called MA(q), if it is of order q. Identically to the AR process, the MA process is only white noise if $\theta_\tau = 0$. The MA parameters can be estimated with a maximum likelihood estimation. Note that the MA process is always stationary.
Considering the ACF and PACF function, the MA process behaves symmetrically to the AR process. Here, the ACF cuts off after lag q and the PACF tails off, while coming closer to zero after lag q.

3.3 ARMA and ARIMA Processes

For more complex processes that cannot be described with an AR or MA model, Box and Pierce (1970) introduced the *mixed autoregressive moving average* (ARMA) model, where the stationary process is modeled with an AR part and a MA part. They also expanded the ARMA model to the *autoregressive-integrated moving average* (ARIMA) model, in case the time series is not stationary, but can be transformed into a stationary process through differencing. The general definition is

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q},$$

or as defined by Box and Pierce (1970)

$$\phi(\mathrm{B})\nabla^d x_t = \theta(\mathrm{B})w_t,$$

with B as the backward shift operator, such that $\mathrm{B}x_t = x_{t-1}$, $\phi(\mathrm{B}) = 1 - \phi_1\mathrm{B} - \ldots - \phi_p\mathrm{B}^p$ and $\theta(\mathrm{B}) = 1 - \theta_1\mathrm{B} - \ldots - \theta_q\mathrm{B}^q$. The ARIMA model is of order (p,d,q), where p indicates the order of the AR part of the time series and q the order of the MA part. The d indicates the order of integration, where the time series is integrated of order $\mathrm{I}(d)$ if it is a stable and invertible ARMA process and thus $\mathrm{I}(0)$ after differencing d times. As the MA process is always stationary the ARIMA process is stationary, if its AR part is stationary. For the ARIMA process the ACF and PACF both tail off in exponentials and/or sine waves.
A summary over the three different ways to model the stochastic component of a time series and their properties is given in Table 1.

	Autoregressive Processes	**Moving Average Processes**	**Mixed ARMA Processes**
Model in terms of previous x_t's	$\phi(\mathrm{B})x_t = w_t$	$\theta^{-1}(\mathrm{B})x_t = w_t$	$\theta^{-1}(\mathrm{B})\phi(\mathrm{B})x_t = w_t$
Model in terms of previous w's	$x_t = \phi^{-1}(\mathrm{B})w_t$	$x_t = \theta(\mathrm{B})w_t$	$x_t = \phi^{-1}(\mathrm{B})\theta(\mathrm{B})w_t$
Stationarity condition	Roots of $\phi(\mathrm{B}) = 0$ lie outside the unit circle	Always stationary	Roots of $\phi(\mathrm{B}) = 0$ lie outside the unit circle
Invertibility condition	Always invertible	Roots of $\theta(\mathrm{B}) = 0$ lie outside the unit circle	Roots of $\theta(\mathrm{B}) = 0$ lie outside the unit circle
Autocorrelation function	Infinite (damped exponentials and/or damped sine waves)	Finite	Infinite (damped exponentials and/or damped sine waves after first $q - p$ lags)
	Tails off	Cuts off after lag q	Tails off
Partial autocorrelation function	Finite	Infinite (damped exponentials and/or damped sine waves)	Infinite (damped exponentials and/or damped sine waves after first $q - p$ lags)
	Cuts off after lag p	Tails off	Tails off

Table 1: Summary of Properties (cf. Box, Jenkins, and Reinsel 2008, p. 87)

3.4 Model Selection

After having introduced the different stochastic models, one has to decide which one fits the time series data. To select the best fitting model two different approaches are considered here. One is widely known as the Box-Jenkins-Method, while the other is based on different information criteria. The problem one has to face is to include necessary parameters but not to over-fit the model, as a complex or overfitted model will deteriorate the quality of forecasts.

Portmanteau Test

The Box-Jenkins approach to checking the fit of the model is to estimate the parameters and afterwards prove if the residuals are independently and randomly distributed (Makridakis and Hibon 1997). To test this, Ljung and Box (1978) developed the so-called Portmanteau test

$$\mathrm{Q} = T(T+2)\sum_{\tau=1}^{k}\frac{r^2(\tau)}{(T-\tau)} \sim \chi^2,$$

where T is the number of observations, k is the number of correlation coefficients and r is the estimated autocorrelation function. The null hypothesis of white noise ($\mathrm{Q}=0$) can be rejected if the χ^2 test-statistic with k degrees of freedom is higher than the chosen significance level (Enders 2010). If one cannot reject the null hypothesis, then the residuals are white noise and thus only the purely random component is left of the process. One can use the chosen model for further evaluations and forecasting. If the null hypothesis can be rejected one should find another model to describe the process.

Information Criteria

The second method is based on information criteria. They give an answer to the question whether the chosen model fits the data but they penalize additional parameters so that the most complex model is not always preferred. Two often used information criteria are the Akaike information criterion (AIC) and the Bayesian (or Schwartz) information criterion (BIC). It is the aim to minimize the following equations

$$\begin{aligned}\mathrm{AIC} &= -2\ln L + \frac{k}{T} = T\ln(\mathrm{RSS}) + 2n,\\ \mathrm{BIC} &= -2\ln L + \frac{k}{\ln T} = T\ln(\mathrm{RSS}) + n\ln T,\end{aligned}$$

where L is the value of the maximized log likelihood function, RSS is the residual sum of squares, k is the number of parameters used in the model and T is the number of observations (Enders 2010). To select the best model one should determine a maximum order and than compare all possible models with a lower order according to their AIC and BIC. The model with the lowest AIC and/or BIC should be chosen for further investigation. Generally, the BIC penalizes additional parameters more severely.

4 Trend and Seasonality

Recapitulating the definition of a basic time series in section 2, we have introduced methods to characterize the stochastic part. But most time series do not only consist of a purely stochastic part but do also have a trend and seasonality component. To model such time series one has to eliminate the trend and seasonality before fitting any of the models considered in section 3 to the residuals of the adjusted time series. In this section, a few methods to adjust the time series will be discussed.

4.1 Trend

Trend is the most common form of non stationarity, especially in economic time series. As seen in the random walk example, integrating the data can help to make the process stationary. But differencing is not the right solution to eliminate trend if the trend is not stochastic but deterministic. Differencing a time series with a deterministic trend component can lead to a non-invertible MA(1) process and assume correlations where there are none. Thus, if a time series is stationary around a trend, the proper way to handle the trend problem is to use dynamic regression to extract the trend component from the stochastic process.
For a time series with a stochastic trend one can use the first or higher differences of the data and fit an ARIMA model as defined in section 3.3. If the data has a trend it should not be stationary, so the ADF test should not reject the null hypothesis. But as the ADF test only tests for unit roots Kwiatkowski et al. (1992) developed the KPSS test to discover a trend component. In this test the null hypothesis of the data being either stationary around a linear trend or around a level is tested (Kwiatkowski et al. 1992).

4.2 Seasonality

As most time series data is taken in small intervals over many years it is likely to contain seasonal factors. Of the various possibilities to handle seasonality two approaches are considered here. The first uses multiple regression with dummy variables or harmonic functions, the second approach fits a seasonal ARIMA Model to the data. Although these approaches are now considered separately it can also be a solution to combine them.

Regression

If the seasonal effects in the time series are constant and additive they can be extracted using a regression model with dummy variables for each season. For a time series that contains monthly seasonality $(s_1, \ldots, s_{12})$ the regression could be considered as

$$x_t = \beta_1 s_1 + \beta_2 s_2 + \ldots + \beta_{12} s_{12} + \varepsilon_t,$$

where the time series consists of the additive seasonal components s_t, a stochastic noise term ε_t, that may be autocorrelated, and the regressors $\beta_1, \ldots, \beta_{12}$. If the seasonality is assumed to be multiplicative it can be useful to take the logarithm of the data to achieve additive components.

Some seasonal effects are better considered not as linear parameters but as smoothly varying over time or seasons. Therefore sine and cosine functions can be estimated to build a model with smooth seasonality. Cowpertwait and Metcalfe (2009) define a harmonic model as

$$x_t = m_t + \sum_{i=1}^{[s/2]} \left[s_i \sin(2\pi i \frac{t}{s}) + c_i \cos(2\pi i \frac{t}{s}) \right] + \varepsilon_t,$$

where s are the number of seasons, $[s/2]$ are the possible cycles, i the frequency and s_i and c_i are unknown parameters.

Seasonal ARIMA Model (SARIMA)

If one considers the seasonality in the time series to be stochastic, one can fit an ARIMA model with an extra multiplicative seasonal component. Box, Jenkins, and Reinsel (2008) divide the multiplicative model into "*between periods*"

$$\Phi_P(\mathrm{B}^s)\nabla_s{}^D x_t = \Theta_Q(\mathrm{B}^s) w_t,$$

and "*within periods*"

$$\phi_p(\mathrm{B})\nabla^d x_t = \theta_q(\mathrm{B}) w_t.$$

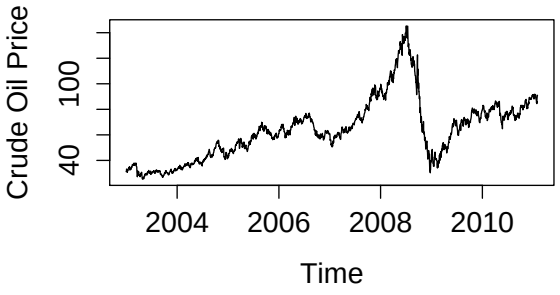
The general multiplicative case for a SARIMA$(p,d,q)\times(P,D,Q)_s$ model is then defined as

$$\phi_p(\mathrm{B})\Phi_P(\mathrm{B}^s)\nabla^d\nabla_s{}^D x_t = \theta_q(\mathrm{B})\Theta_Q(\mathrm{B}^s) w_t,$$

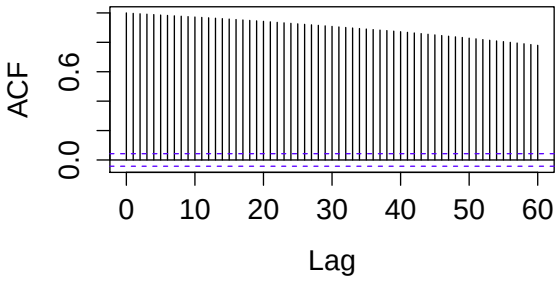
where Φ and Θ are the AR and MA coefficient of the seasonal pattern, respectively. And B^s is the seasonal back shift operator. The order of the ARIMA-part is p,d,q, and the order of the seasonal pattern is P,D,Q. The seasonal period is denoted by s. The variable D refers to the seasonal difference, e. g. if the time series consists of monthly data over several years, the seasonal difference would be $\Delta s_t = s_t - s_{t-12}$ (Box, Jenkins, and Reinsel 2008). The model can only be fit to the data if the process is stationary or stationarity in both parts can be achieved through differencing. That means one has to check for unit roots in the "within" part and also for seasonal unit roots in the "between" part.

5 Analysing Time Series in R

The previous chapters have shown the mathematical backgrounds. The remainder of the paper will test these concepts first on the West Texas Intermediate (WTI) crude oil price and, afterwards, on electricity prices from intraday data of the European Energy Exchange (EEX) market. This analysis will be done using the statistic software "R". The commands in the program are written as `command`. Most functions used are from the package "stats" developed by the R Core Team (2013).



(a) Crude oil price in $ per barrel.



(b) Autocorrelation function across lags with 95 % confidence interval (blue dotted line).

Figure 1: The WTI crude oil price plot in business days per week and corresponding ACF in period 2002-2011.

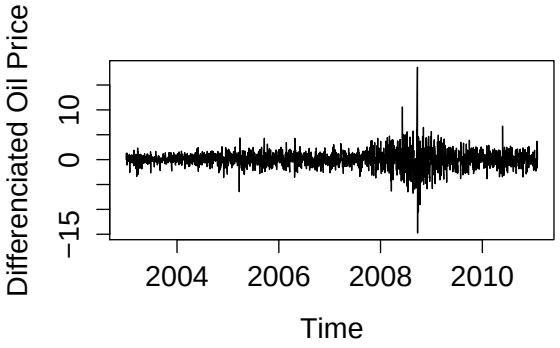
5.1 WTI Crude Oil Price

The WTI Oil Price is given from Dec 31, 2002 until Feb 01, 2011, for all five business days per week. In "R", one can create a time series object with the function `ts`. A first look at the plot of the time series in Figure 1a gives valuable information concerning trending behaviour. In this case the graph plotted with `plot.ts` shows a possible upward trend. The autocorrelation function, which plots the sample correlation against the lags (`acf(oilts, lag.max = 100)`), Figure 1b shows a slow decay, which also indicates non-stationarity. Finally, the ADF test (section 2.2) cannot reject the null hypothesis of a unit root with a p-value of 0.5587 and the KPSS test rejects the null hypothesis of trend stationarity on a significance level lower than 1 %. The last two tests are part of the "R"-package `tseries` (see Trapletti and Hornik 2013)

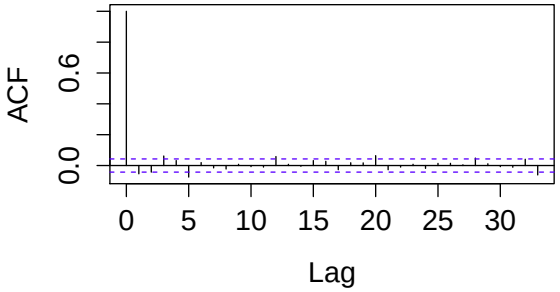
```
1 adf.test(oilts)
2 kpss.test(oilts, null="Trend")
```

Considering the previous stationarity tests, it seems appropriate to take a look at the first difference of the oil price data in Figure 2a. The corresponding function is `diff(oilts)`. If the process would follow a random walk, the resulting sample autocorrelation function of the differentiated data should not be autocorrelated anymore. Figure 2b shows that this is almost the case. But other models might be more accurate to describe the process. Table 2a shows selected models for the differentiated I(1) oil price data. If one takes a look at the information criteria and the Ljung-Box p-value (as discussed in section 3.4), the AIC is lowest for the AR(5) process, the BIC is lowest for the MA(1) process and the Ljung-Box p-value is highest again for the AR(5) process.

In "R" the corresponding function to fit an AR model is `ar(x, method=c("yw", "mle", "ols","burg", order.max)`. One can chose between different methods of parameter estimation (Yule-Walker, maximum-likelihood estimation, ordinary-least-squares or burg-algorithm) and determine the maximum order of the model, if no maximum order is selected the function chooses the order according to the AIC. The output is then the selected process of order p and the estimated AR parameters.



(a) First difference of the crude oil price.



(b) ACF of the differentiated oil price across lags with a 95 % confidence interval (blue dotted line).

Figure 2: Differentiated crude oil price time series plot and corresponding ACF.

Model	AIC	BIC	Ljung-Box
Random Walk	8111.71	8123.02	0.0136
AR(1)	8105.62	8116.93	0.9092
AR(5)	8089.96	8123.89	0.9746
MA(1)	8105.09	8116.39	0.9256
MA(5)	8090.16	8124.09	0.9575
ARMA(1,1)	8106.29	8123.25	0.8929
ARMA(2,2)	8093.93	8122.20	0.4896

(a) Different models for the differentiated oil price.

	Orig. Data	Random Walk	AR(5)	MA(1)
1	71.93	74.69	74.841	74.711
2	73.97	74.69	74.943	74.711
3	74.99	74.69	74.885	74.711
4	74.52	74.69	74.748	74.711
5	74.52	74.69	74.802	74.711

(b) Comparison of the original data of the first days of the test set to the forecast of the best fitted models of the training set.

Table 2: Results to the analysis of the oil price data.

```
ar(x = diff(oilts), method = "mle")

Coefficients:
      1       2       3       4       5
 0.9526  0.0085  0.0824  −0.0094  −0.0361

Order selected 5  sigma^2 estimated as  2.707
```

The command SelectModel(x, ARModel = "AR", Criterion=c("AIC", "BIC"), Best=x), where the best x models are shown, that have been chosen either according to the AIC or BIC (McLeod and Zhang 2008), is also helpful. To fit more complex models to the data the function arima(x, order=c(p,d,q)) can be used. The output also includes the AIC and BIC of the model. One can also find a built-in function to select the order of an ARIMA model automatically auto.arima(x, stationary = TRUE,FALSE, ic="bic","aic", stepwise=TRUE, trace=FALSE, test="adf") (see Stoffer 2012). Table 2a has been created by using the arima function, fitting all possible models and afterwards compare their AIC and BIC. To choose the best model one might also consider how much of the original time series the different fitted models can actually explain . In Figure 3 the AR(5) and MA(1) fitted values are plotted in comparison to the original differentiated data. As can be seen here, neither model can take the high volatility of the oil price into account. The prediction is therefore not as accurate as

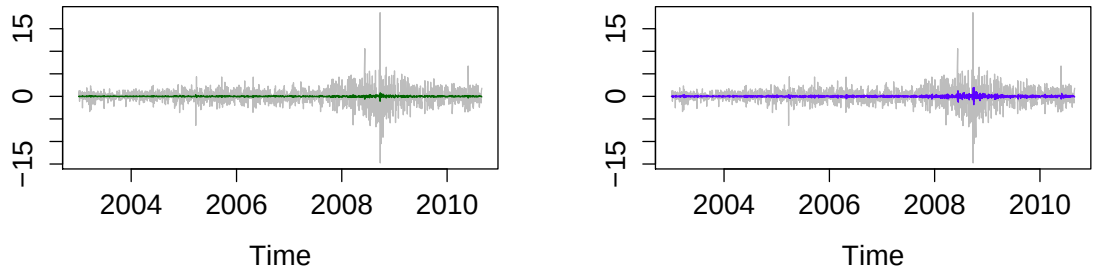
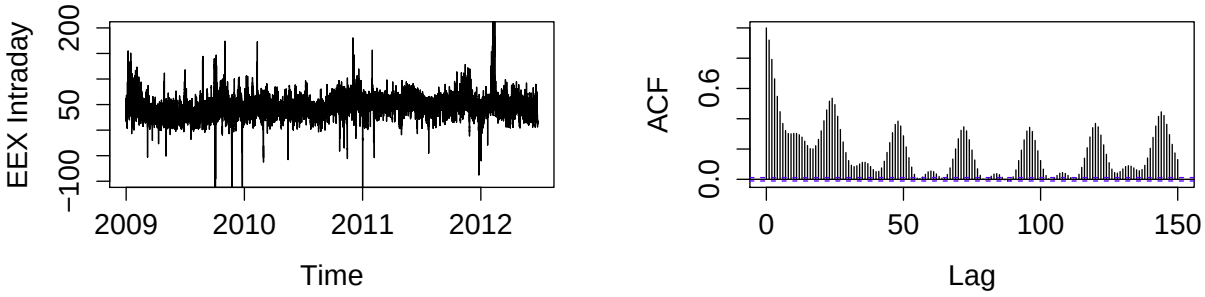


Figure 3: Fitted models of MA(1) (left) and AR(5) (right) in comparison to the original differentiated crude oil price (grey).

one might hope. In Table 2b, the forecasting performance for the three best models, the random walk (RW), AR(5) and MA(1), are summarized. For that a subset of the original data was used as "training" set. The three models were then fitted to the training data and used to forecast the next five oil prices. The first column of the table shows the values of the original data, the following columns show the predicted values from the different models. In "R", the forecasting can be done with the function forecast if the object one wants to forecast is of the class arima. This function is included in the "forecast"-package by Hyndman (2013). One can also use predict from the "stats"-package, which can also forecast other objects.
One can see that all three models perform rather poorly. As the AR(5) and MA(1) model fail to include the high volatility one could consider the random walk as most fitting for the data. The random walk hypothesis is also supported widely for stock market prices (see also Fama 1970, Lee 1992). Although the residuals of the differentiated oil price process are not normally distributed, it seems appropriate to take the random walk into account and calculate a high signal in noise that cannot be evaluated with a classical AR model.

5.2 EEX Intraday Data

The intraday Data from the European Energy Exchange (EEX) is given as hourly weighted average from Jan 01, 2009 01:00am to June 26, 2012 11:00pm. The plot of the time series Figure 4a shows that the process seems to be stationary. The ADF test confirms this by rejecting the null hypothesis of a unit root with a p-value smaller than one percent. The KPSS test also rejects the null hypothesis of trend stationarity on a $1\,\%$ significance level. Differencing the data should, thus, not be needed. Looking at the ACF function of the process, it becomes clear that the process has various seasonal patterns (Figure 4b).
In section 4, a few methods to eliminate the seasonality have been shown. In this example, it seems most appropriate to fit a harmonic model as in section 4.2 to the time series. Cowpertwait and Metcalfe (2009, p. 101 ff) describe a way to define possible sine and cosine functions in "R" via matrices. For the EEX Data a weekly cycle seems to be fitting. Because the data is given hourly, one season is here 168 hours and thus a cycle is $168/2 = 84$ hours. This is implemented in "R" as follows



(a) Plot of the EEC intraday price.

(b) ACF of the EEX intraday price across lags with a 95 % confidence interval (blue dotted line).

Figure 4: EEX intraday data from 2009 to 2012 in hours and corresponding ACF.

```
SIN <- COS <- matrix(nr=length(time(WeighAvPrice)), nc=84)
for (i in 1:84) {
  COS[,i] <- cos(2*pi*i*time(WeighAvPrice)/168)
  SIN[,i] <- sin(2*pi*i*time(WeighAvPrice)/168)
}
```

Now, one can regress the different harmonics on the weighted average price of the data. This was done with all 84 sine and cosine functions to find out which harmonics are significant. Afterwards all insignificant harmonics with a t-statistic less than a critical value of $1,96$ were taken out, so that 17 harmonics were left. With the help of the AIC criterion various regression models have been checked including different dummy variables and different harmonics. The code below shows the best regression model that was used for further investigation.

```
season <- lm(WeighAvPrice ~ 0 + time(WeighAvPrice) + I(time(WeighAvPrice)^2)
             + I(time(WeighAvPrice)^3) + I(time(WeighAvPrice)^4)
             + COS1[,1] + COS1[,2] + SIN1[,1] + SIN1[,2]
             + COS[,1] + COS[,2] + COS[,3] + COS[,4]  + COS[,5]
             + COS[,6] + COS[,9] + COS[,10]
             + SIN[,1] + SIN[,2] + SIN[,3] + SIN[,4]  + SIN[,6]
             + SIN[,8] + SIN[,9] + SIN[,15] + SIN[,16]
             + Hour + Season + Year + Date)
```

The constant 0 has been included, so that no dummy category is omitted, and the trend component is fitted as a polynomial of order four. The four last regressors are dummy variables for hour, season, year and date respectively. The SIN and COS functions are harmonics for the weekly cycles, whereas SIN1 and COS1 are harmonics for the daily cycles. The residuals of this regression should now include only the stochastic components of the data.

Looking at the ACF and PACF of the deseasonalized data, one can still find reoccurring peaks. This is a sign for an additional stochastic seasonal component, which can be characterized by a SARIMA model (section 4.2). In Table 3 several models are summarized according to their AIC, BIC and Ljung-Box test p-value. The best model to describe the intraday data seems to be the seasonal ARMA$(12,0)(1,0,1)_{24}$ model. It has the lowest AIC, BIC and the highest p-value and the residuals of the model show hardly any autocorrelation. The fitted model of the seasonal AR

Model	AIC	BIC	Ljung-Box
$AR(13)$	201994	202119	0.8715
$AR(14)$	201974	202108	0.9695
$ARMA(2,3)$	202008	202058	0.9632
$ARMA(4,1)$	202013	202063	0.9931
$ARMA(12,0)(1,0,1)_{24}$	199560	199686	0.9957
$ARMA(14,0)(1,0,0)_{24}$	204733	204866	0.9722
$ARMA(14,0)(1,0,1)_{24}$	200325	200467	0.9957
$ARMA(13,0)(1,0,1)_{24}$	200354	200487	0.8778
$ARMA(8,1)(1,0,1)_{24}$	200381	200481	0.9966

Table 3: Comparison of selected models for the deseasonalized EEX intraday price.

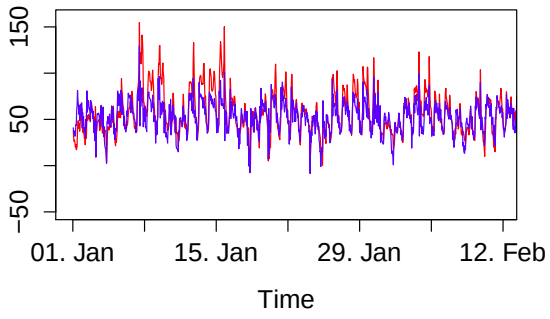


Figure 5: Fitted values of the seasonal AR process and the deterministic seasonality (blue) in comparison to the original intraday price (red) for the first months in 2009.

process plus the deterministic seasonality is plotted in Figure 5. To predict the data, one would now need to add the deterministic seasonal and trend components to the stochastic forecast.

6 Conclusion

The analysis of the crude oil price has shown that a part of the stochastic process can be described in terms of an appropriate AR model. But the chosen models cannot really account for the high volatility of the stock price. Here, one might consider other approaches to model the time series to improve the forecasts. One could for example relax the assumptions taken e. g. heteroscedasticity. As for the EEX intraday data, the resulting model can characterize the stochastic and deterministic process quite well. It has been shown that even more complex data, with different seasonal patterns can be adjusted so that one can fit one of the basic stochastic models. Putting it all in a nutshell, the classic time series models can help to understand and classify the stochastic process underlying discrete time series data. Even those rather simple models can describe the data fittingly, but the prediction quality leaves room for further improvement. In "R", a huge variety of different packages and functions is included to simplify the analysis.

A References

BOX, G. E. P. and D. A. PIERCE (1970). *Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models*. English. In: *Journal of the American Statistical Association*, Vol. 65, No. 332, pp. 1509–1526.

BOX, G. E. P., G. M. JENKINS, and G. C. REINSEL (2008). *Time series analysis : forecasting and control*. 4. ed. Wiley series in probability and statistics. Hoboken NJ: Wiley.

COWPERTWAIT, P. S. P. and A. V. METCALFE (2009). *Introductory time series with R*. eng. Use R! Dordrecht ; Heidelberg [u.a.]: Springer, p. 254.

DICKEY, D. A. and W. A. FULLER (1979). *Distribution of the Estimators for Autoregressive Time Series With a Unit Root*. In: *Journal of the American Statistical Association*, Vol. 74, No. 366, pp. 427–431.

ENDERS, W. (2010). *Applied econometric time series*. eng. 3. ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley, p. 517.

FAMA, E. F. (1970). *Efficient Capital Markets: A Review of Theory and Empirical Work*. English. In: *The Journal of Finance*, Vol. 25, No. 2, pp. 383–417.

HYNDMAN, R. J. (2013). *forecast: Forecasting functions for time series and linear models*. R package version 4.8.

KWIATKOWSKI, D. et al. (1992). *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?* In: *Journal of Econometrics*, Vol. 54, No. 13, pp. 159 –178.

LEE, U. (1992). *Do stock prices follow random walk?:: Some international evidence*. In: *International Review of Economics & Finance*, Vol. 1, No. 4, pp. 315 –327.

LJUNG, G. M. and G. E. P. BOX (1978). *On a measure of lack of fit in time series models*. In: *Biometrika*, Vol. 65, No. 2, pp. 297–303.

MAKRIDAKIS, S. and M. HIBON (1997). *ARMA Models and the Box-Jenkins Methodology*. In: *Journal of Forecasting*, Vol. 16, No. 3, pp. 147–163.

MCLEOD, A. I. and Y. ZHANG (2008). *Improved Subset Autoregression: With R Package*. In: *Journal of Statistical Software*, Vol. 28, No. 2.

R CORE TEAM (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

SHUMWAY, R. H. and D. S. STOFFER (2006). *Time series analysis and its applications : with R examples*. 2. ed. Springer texts in statistics. New York NY: Springer.

STOFFER, D. (2012). *astsa: Applied Statistical Time Series Analysis*. R package version 1.1.

TRAPLETTI, A. and K. HORNIK (2013). *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-32.

WOOLDRIDGE, J. M. (2013). *Introductory econometrics : a modern approach*. 5. ed., internat. ed. [Mason, Ohio u.a.]: South-Western Cengage Learning, p. 878.

B List of Figures

C List of Tables