Predicting stocks prices with time series techniques

Yevheniia Mitriakhina

May 2019

Abstract

Finding patterns in stock prices data is vital for effective trading so we focus on techniques for short-term price prediction. The choice of the asset analyzed is justified by the assumption of presence of significant factor of seasonality in the stock price of company specialized on foods. The Hershey Company(HSY) commonly called Hershey's, is an American company which specializes on manufacturing chocolate, and a variety of sweet products.

Explanatory analysis

The data includes 2833 observations starting from 02.01.2008 downloaded from yahoo.finance. We transform it into proper time series format with frequency of 251, which is mean number of trading days per year. Firstly, we perform simple moving average to find if there is a trend and if seasonality is present.

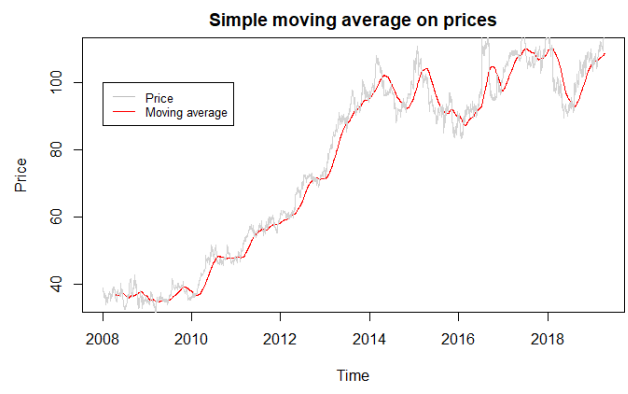


Figure 1: Simple moving average with n=75 (approximately a quarter in trading days)

From the graph we can observe that there is a monotonous increasing trend and probably an annual seasonal pattern. The prices are obviously non-stationary, but we perform Augmented Dickey–Fuller test to see if we can reject the hypothesis that the unit root is present in the sample against the alternative: data is stationary.

```
        Augmented Dickey-Fuller Test

data:  data.prices
Dickey-Fuller = -2.3082, Lag order = 13, p-value = 0.4478
alternative hypothesis: stationary
```

Figure 2: Results of ADF-test on raw data

The p-value of the test is 0.45, so we cannot reject the hypothesis of non-stationarity present in data with 95% confidence. Followingly, we decompose time series into 3 components: trend - the overall upward or downward movement of the data points; seasonal - any monthly/yearly pattern of the data points; remainder - random part of the data.

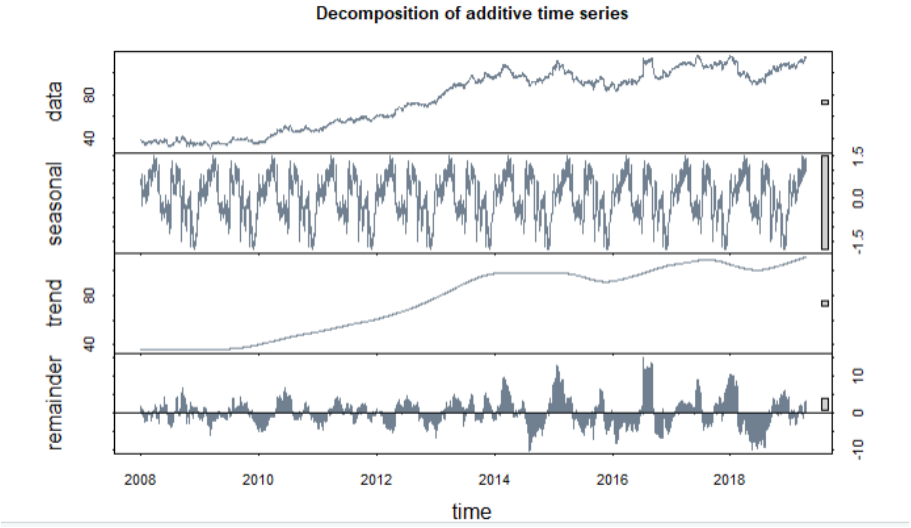


Figure 3: Decomposition with stl() function.

We can see a clear annual seasonal pattern, and an incresing trend, which is more steep in the years 2008-2014 and more flat afterwards. The remainder part, which is our 'error' has greater magnitude over time, so we would need to take logarithm to stabilize the variance of errors. Given these points, we need to take logarithm, detrend and difference the data in order to achieve stationarity and use ARIMA model.

Data preparation

After taking logarithm the values are on the scale from 3.4 to 4.2. We would fit the linear trend to this series with *tslm* which is largely a wrapper to lm, but it allows to account for trend and seasonality components. We don't use the trend values obtained in previous section, because it appears to be overfitted to the data and is impossible to extrapolate to future values.

The formula looks as follows:

$$log(prices) = \beta_0 + \beta_1 trend + \epsilon$$

After running regression we obtain:

```
Call:
tslm(formula = log.prices ~ trend)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31849 -0.09897 -0.01017  0.10545  0.36461

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.606e+00  5.775e-03   624.4   <2e-16 ***
trend       4.605e-04  3.530e-06   130.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1536 on 2831 degrees of freedom
Multiple R-squared:  0.8574,    Adjusted R-squared:  0.8573
F-statistic: 1.702e+04 on 1 and 2831 DF,  p-value: < 2.2e-16
```

Figure 4: Parameters for linear trend

We have a good fit, R-squared is 0.85 which means that 85% of variation in our data is well explained by this trend. The slope is obviously positive, because the trend is increasing. On the plot it looks as follows:

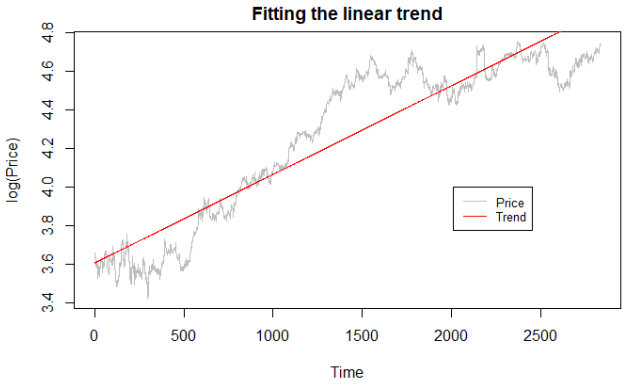


Figure 5: Linear trend

Now we would subtract from each observation value which is explained by the trend and model the residuals.

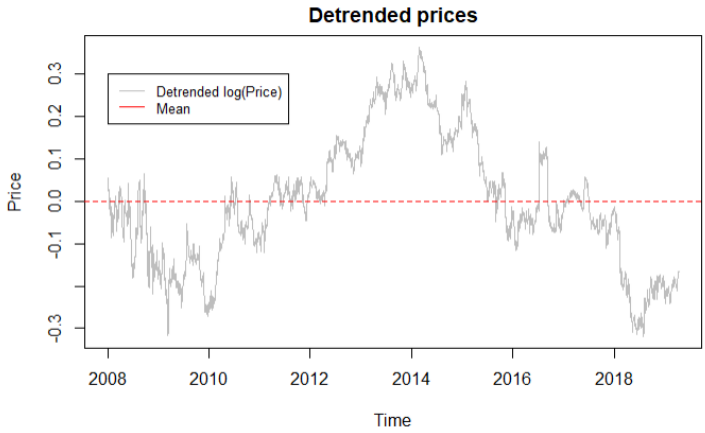


Figure 6: Detrended prices

Finally, we take the first difference of the price and see if the data is stationary now. The differenced data has following distribution:

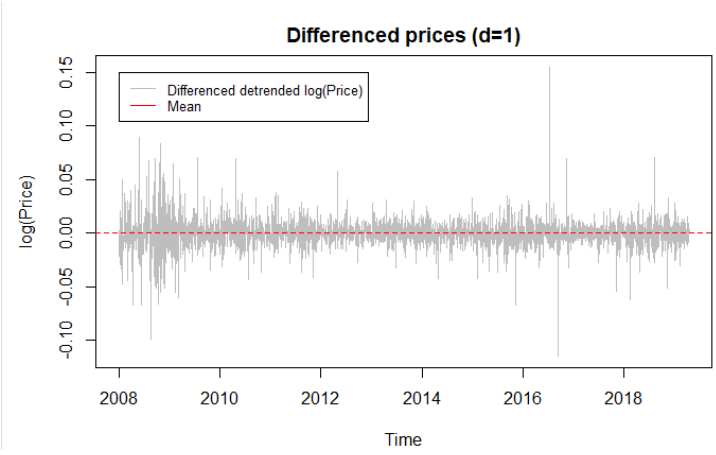


Figure 7: Differenced prices

And the test values:

```
        Augmented Dickey-Fuller Test

data:  differenced.prices
Dickey-Fuller = -14.672, Lag order = 14, p-value = 0.01
alternative hypothesis: stationary
```

Figure 8: Results of ADF-test for differences data

Now we can reject the hypothesis that the data is non-stationary with 95% confidence.

Specifying the model

we will use ARIMA model for forecasting introduced by Box-Jenkins. Our data already corresponds to the assumptions of the model, namely stationarity. In ARIMA model, the future value of a variable is a linear combination of past values and past errors, expressed as follows:

$$y_t = \phi_0 + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_1 \epsilon_{t-1}$$

Where y_t is the actual value, ϵ_t is random error; ϕ_i and θ_i are the coefficients and p and q are parameters referred to as auto-regressive and moving average order. To correctly specify the model we need to determine p and q. We examine ACF and PACF functions. ACF stands for auto-correlation function and it shows how correlated points are with each other, based on how many time steps they are separated by. Partial auto-correlation function, sometimes called conditional auto-correlation shows auto-correlation with its own lagged values, regressed on all the previous lags. From the plots one can observe that data correlates strongly on the most recent values and both the function converge when $n \to \infty$. Lags behind 250-th are already not statistically significant.

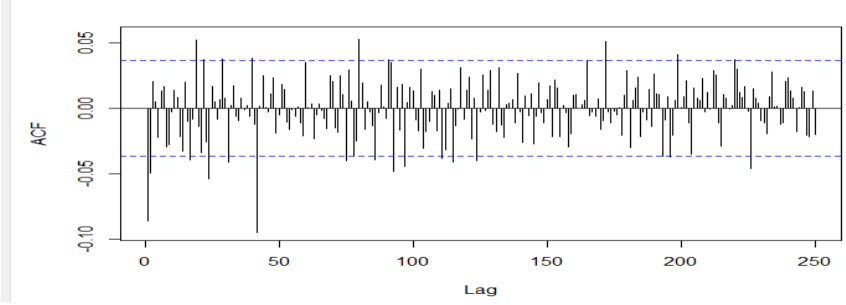


Figure 9: ACF

PACF also allows to observe the seasonal pattern (we have negative spikes approximately every 50 trading days).

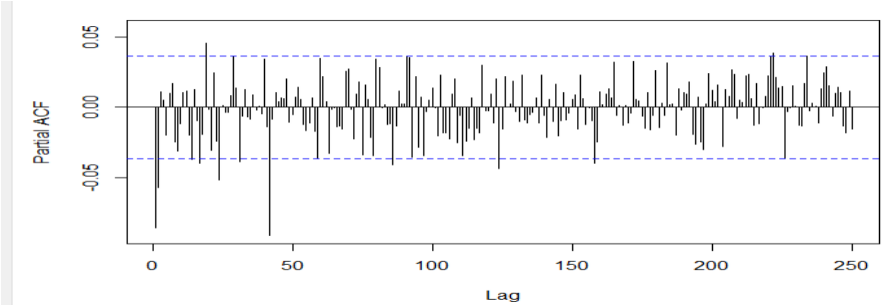


Figure 10: PACF

We will choose parameters based on Akaike information criteria and standart error of regression. The results of the few trials are shown in the table below.

Parameters	*Standard error*	*Akaike information criteria*
(1, 0, 0)	0.0001824	-16349.99
(1, 0, 1)	0.000182	-16354.38
(2, 0, 0)	**0.0001818**	**-16357.28**
(0, 0, 1)	0.0001823	-16352.43
(2, 1, 0)	0.0002492	-15460.31
(2, 1, 2)	0.0001819	-16340.53
(2, 1, 1)	0.0001819	-16342.28
(0, 0, 2)	0.0001819	-16356.4

The best fit is ARIMA(2, 0, 0). So, we run regression with this parameters.

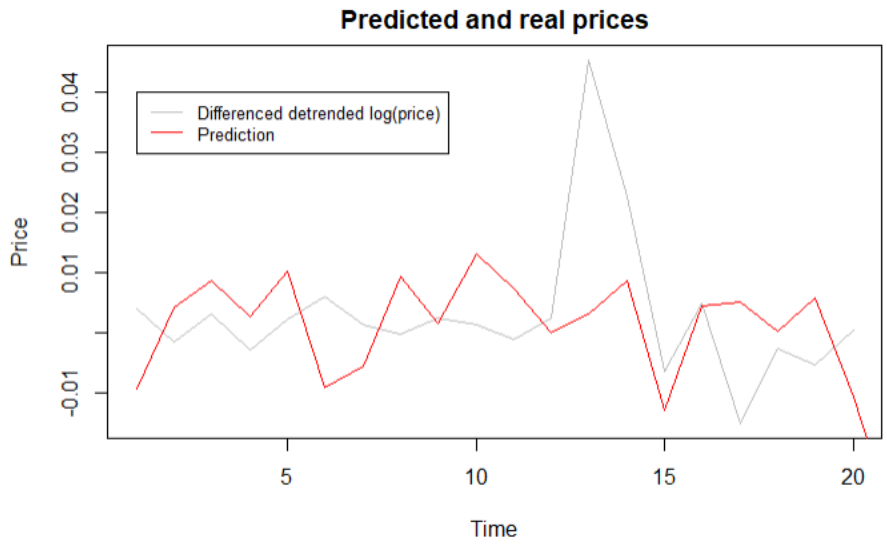
```
Call:
arima(x = differenced.prices, order = c(2, 0, 0))

Coefficients:
          ar1      ar2  intercept
      -0.0910  -0.0572     -1e-04
s.e.   0.0188   0.0188      2e-04

sigma^2 estimated as 0.0001818:  log likelihood = 8182.64,  aic = -16357.28
```

Predicting prices

We set time horizon to 30 days, rather short-term prediction and expose real data from that period to the same transformations we used for the main dataset - take logarithm, detrend and take first difference. The forecast is shown as the red line on the graph below. We can observe that ARIMA was satisfactory at forecasting: the lines are close enough and mostly moving at same direction. Except for the leap at 13th day real prices belong to the 80% confidence interval of the point estimator.



Summary

The experimental results obtained with best ARIMA model demonstrate the potential of ARIMA models to predict stock prices satisfactory on short-term basis if data is prepared and is appropriate according to the assumptions of the model.

References

1. Time Series and Forecasting
https://www.statmethods.net/advstats/timeseries.html
2. Stationarity and differencing
https://otexts.com/fpp2/stationarity.html
3. How To Identify Patterns in Time Series Data: Time Series Analysis
http://www.statsoft.com/Textbook/Time-Series-Analysis