Diploma thesis

Notes

Andreas Egger

[Evaluation of forecasting methods 2](#_Toc403113049)

[Time series analysis 2](#_Toc403113050)

[Trend analysis 2](#_Toc403113051)

[Forecasting models 4](#_Toc403113052)

[Simple Exponential Smoothing 4](#_Toc403113053)

[ARIMA models 7](#_Toc403113054)

[Dataset evaluation 9](#_Toc403113055)

[Dataset with hourly time intervals 9](#_Toc403113056)

[Dataset with 5-minute time intervals 12](#_Toc403113057)

[Model Verification 13](#_Toc403113058)

[Accuracy 13](#_Toc403113059)

[Residual characteristics 13](#_Toc403113060)

[Accuracy measures 18](#_Toc403113061)

# Evaluation of forecasting methods

## Time series analysis

Time series may exhibit trend, seasonal and irregular components. In the following we will investigate methods that apply to the given time series and extract promising components.

### Trend analysis

The original, unedited time series of energy prices in belgium in 2010 is depicted in Figure 1.

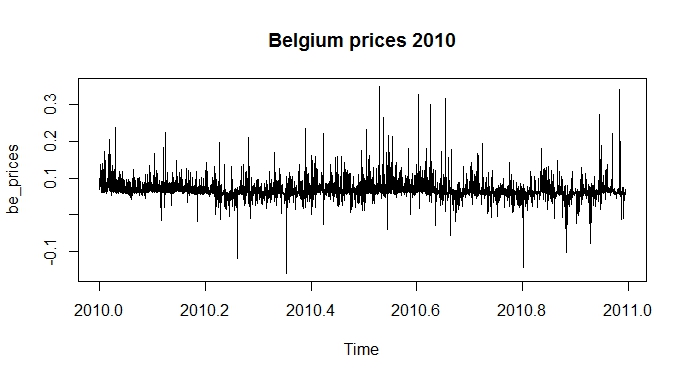


Figure 1: Belgian energy prices over 1 year

This time series shows a strong irregular component and small trends throughout the observed time range, there might also be some seasonal behavior which has to be investigated separately to reveal its characteristics.

For a trend analysis we apply the Simple Moving Average (SMA) method to average out random fluctuations and reveal trends in the data.

The SMA method is configured with a moving window of a specific size to smooth out any disturbing variations in the data. Since the total number of observations exceeds 8000 (number of hours over a year) the window sizes were configured to n=10, n=100 and n=600. The results are shown below.

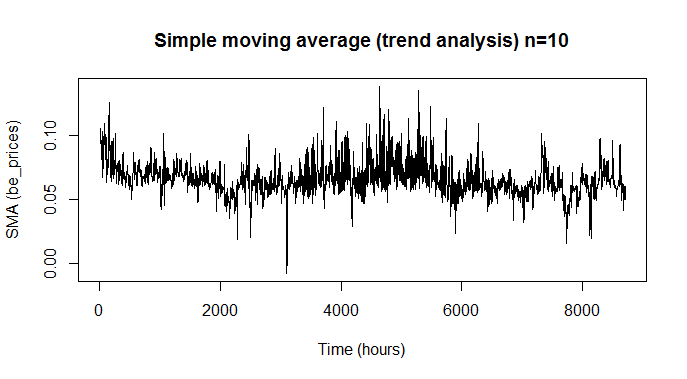


Figure 2: Simple moving average, sliding window size = 10

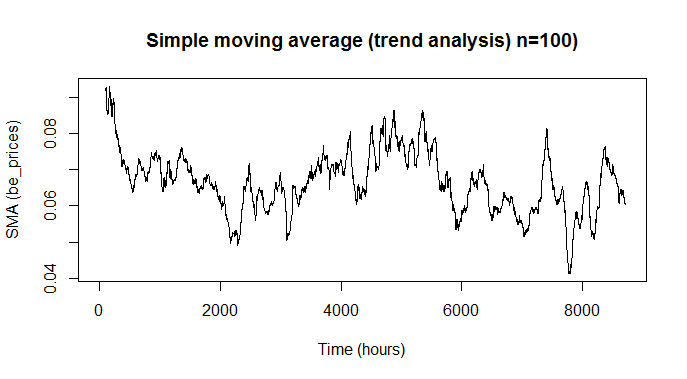


Figure 3: Simple moving average, sliding window size = 100

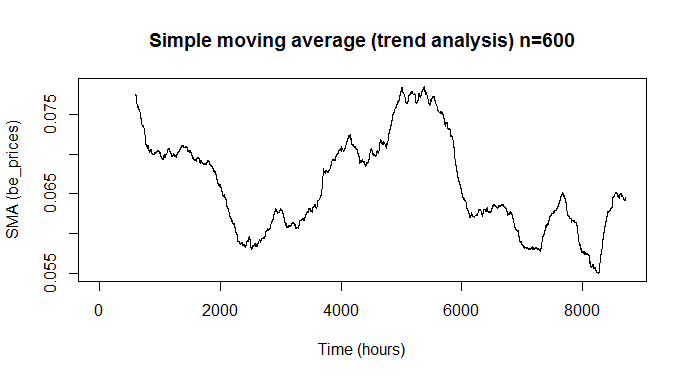


Figure 4: Simple moving average, sliding window size = 600

From the results we can deduce that the dataset exhibits a trend which may help in generating an appropriate forecasting model.

## Forecasting models

### Simple Exponential Smoothing

Time series with a constant level and roughly constant variations over time can be described by an additive model and thus can be applied to the Simple Exponential Smoothing (SES) model.

Simple Exponential Smoothing is based on the HoltWinters method but without considering trend and seasonal components.

We constrain the belgium energy prizes to the first 200 observations in order to get some visibly usable results (Figure 5) .

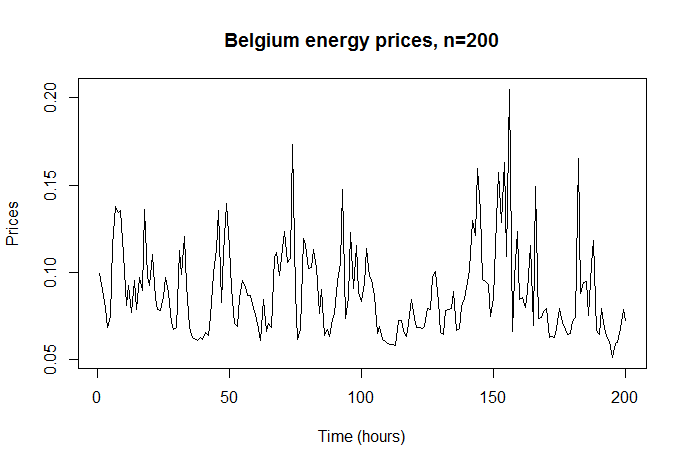


Figure 5: First 200 hours of energy prices in belgium, starting January 2010

Figure 6 depicts the results of applying an SES model to the belgian time series.

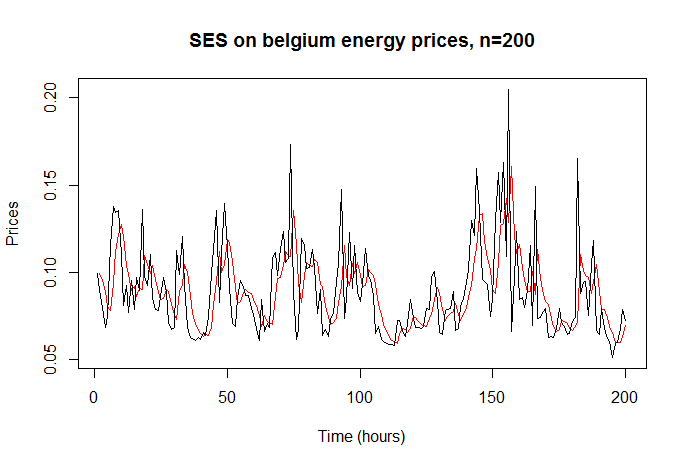


Figure 6: SES applied to belgian energy prices

In Figure 6 the black line shows the original time series whereas the red line denotes the predicted time series based on SES. It can be observed that the SES series lags behind the original series by some degree while additionally some random fluctuations are averaged out.

In Figure 7 an SES forecast beyond the given data is applied to this series. The blue line shows the actual forecast of energy prices whereas the dark and light blue shaded areas denote the 80% and 95% confidence intervals, respectively.

Remarkably the forecast just consists of a straight line, having the same value as the original series’ last observation. This is due to the model being trained without considering trend and seasonality components in the series, so the dataset is actually expected to have constant mean and variation.

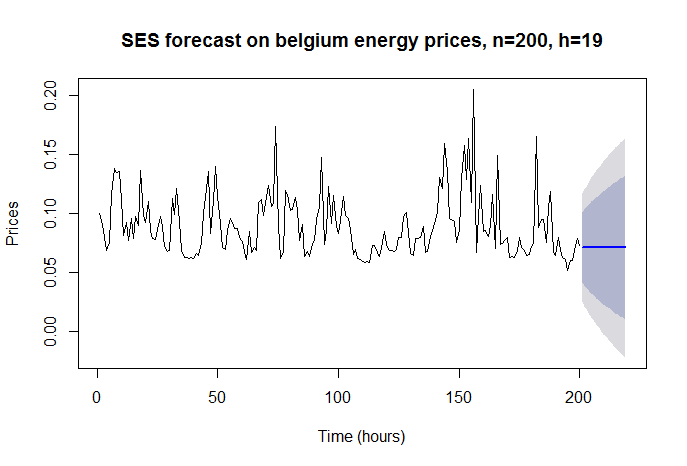


Figure 7: Forecasting 19 hours of belgian energy prices

However, since this dataset does neither have a constant mean nor constant variation, it would first have to be transformed into a series appropriate for this model. A common approach to achieve this is to differentiate the series one or more times, which has to be transformed back after the forecasting is done.

In this case, we differentiate the series one time. The results are shown in Figure 8.

From the results we can estimate that the resulting series does have a constant mean and constant variation on average. After applying the SES model to this time series the predicted values smooth out all random fluctuations and take values at the approximated mean of the series (Figure 9).

The actual forecast using SES based on the differentiated series is again a straight line, but since the series has zero mean and near constant variance this can be taken as reasonable forecast. Also, the prediction intervals stay the same with increasing forecast window size due to the series constant variance and mean. When transforming the data back to the original series the forecasted values have to be transformed as well to obtain the forecasts of the original dataset.

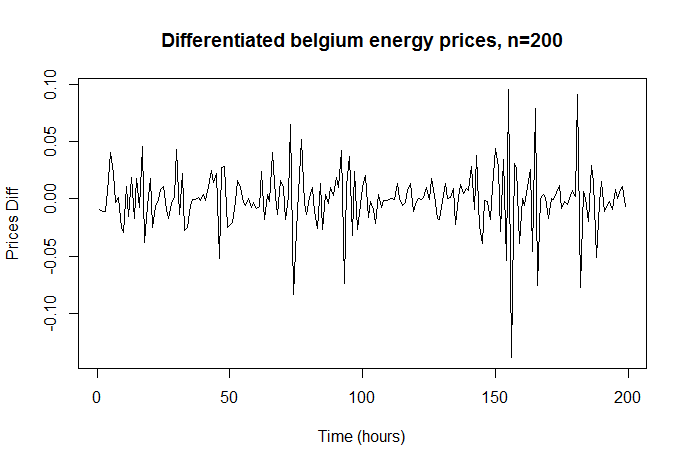


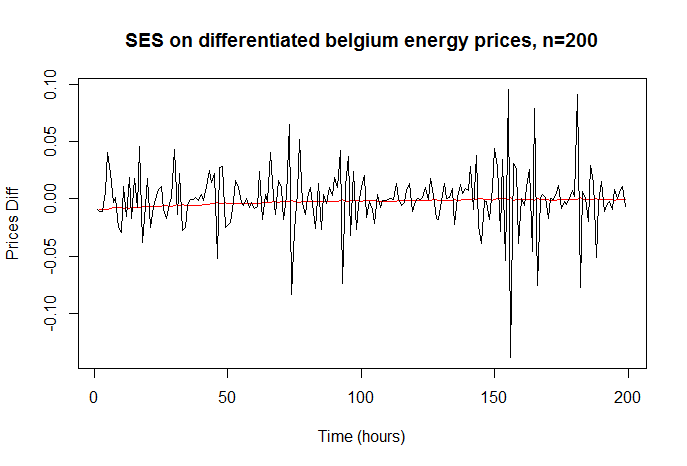
Figure 8: Belgian energy prices, one time differentiated

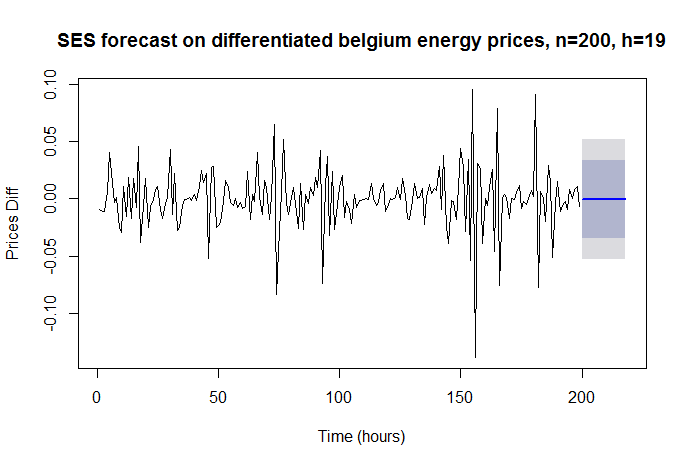
Figure 9: SES model applied to differentiated time series

Figure 10: Forecast on differentiated time series, 19 time intervals into the future

### ARIMA models

ARIMA is an acronym that stands for “autoregressive integrated moving average”. It may consist of three parts, an autoregressive (AR), moving average (MA) and integrated component. It is generally denoted as ARIMA(p,d,q) where parameters p d and q refer to the autoregressive, integrated and moving average components, respectively.

An ARIMA model may be applied to stationary time series only as it is considered to be an additive model. On model creation an irregular component is added which is represented by an artificial time series with zero mean and constant variance. This is done in order to account for random fluctuations in the dataset. Thus, the dataset to be examined must also be stationary, which means that it should exhibit a near constant level and variance.

As ARIMA models may only be applied to stationary datasets a preprocessing step may be necessary prior to the application of any autoregressive (AR) or moving average (MA) components. This step consists of differentiation of the time series one or more times in order to transform a non-stationary series into a stationary one. Once the model has been built this process must be reverted to return to the original series which is done by integration.

AR or MA models are built to correct violations of the white noise assumption.[[1]](#footnote-1)

The main difference between ARIMA models and exponential smoothing models is that ARIMA models can also handle auto-correlations within the time series and the irregular component.

**Formal tests for stationarity**

There are formal tests available to determine if a given dataset has stationary characteristics to a sufficient degree to be applied appropriately i.e. to an ARIMA model. These are so called *unit root tests* since they are basic tests to evaluate the characteristics of a dataset in terms of stationarity.

#### Autoregressive models

Autoregressive or AR models take into account relationships between current and historic timestamps, i.e. the value of a variable in one period is assumed to be related to its values in previous periods. Thus an AR(p) model denotes a model with p lags, taking into account p historic timestamps.

The corresponding formula for an AR(p) model is

where is the value at the current timestamp, is a constant, is an error term (irregular component with mean zero and constant variance) and is the coefficient for the lagged variable at time t-i. as residual should have zero mean + constant variance -> irregular c.

An ARMA(p,0) model is an autoregressive model of order p, or AR(p). The formula again depends on the order of the model. For example, an ARMA(2,0) or AR(2) model is defined as

where denotes the stationary time series, denotes the mean of the time series , represents the irregular component with mean zero and constant variance and and are estimated parameters.

AR models are generally used for time series with longer term dependencies between successive data points.

An ARMA(2,0) model is an autoregressive model of order two because in the calculation there are two additional steps of the time series considered ( and ), each weighted with a coefficient (and ). It is auto-regressive since the calculation is based on the series itself and thus considers history values up to the order of the autoregression.

#### Moving average models

A moving average or MA model takes into account the relationship between a data point and the residuals of previous data points, given in number of lags. Thus an MA(q) model denotes a moving average model with q lags, taking into account the last q error terms.

The corresponding formula for a MA(q) model is

An ARMA(0,q) model is a moving average model of order q, or MA(q). The corresponding formula depends on the order of the model. For example, an ARMA(0,1) or MA(1) model is defined as

where denotes the stationary time series, denotes the mean of the time series , represents the irregular component with mean zero and constant variance and is an estimated coefficient. It defines a model of order one since only the latest historical data point is taken into account (. Since the irregular series is added in the calculation, the examined dataset needs to be cleaned of irregular components, i.e. it should have near constant mean and variation.

#### ARMA models

ARMA models are a combined form of both AR and MA models, so they connect both models to one comprehensive model. Depending on the sign of the coefficients in the AR(p) and MA(q) parts the resulting series would either be a smoothed or oscillating version of the original series.

The corresponding formula is defined as

where the first part denotes the AR model with p lags and the second part denotes the MA model with q lags, respectively.

#### Stationarity

In order to apply an ARMA(p,q) model to a time series this series has to be stationary. A stationary series exhibits a constant mean and variance that does not change over time and which does not include any trends.

If a series shows an AR(1) disturbance process

it is defined to be stationary in case two conditions apply:

1. shows white noise characteristics

## Dataset evaluation

Two different datasets were used for these examples, one dataset with time series of hourly interval data, another comprising 5-minute interval time series. Both datasets will be discussed in terms of general behaviour and characteristics.

### Dataset with hourly time intervals

This dataset consists of six time series geographically distributed over three continents (Figure 11).

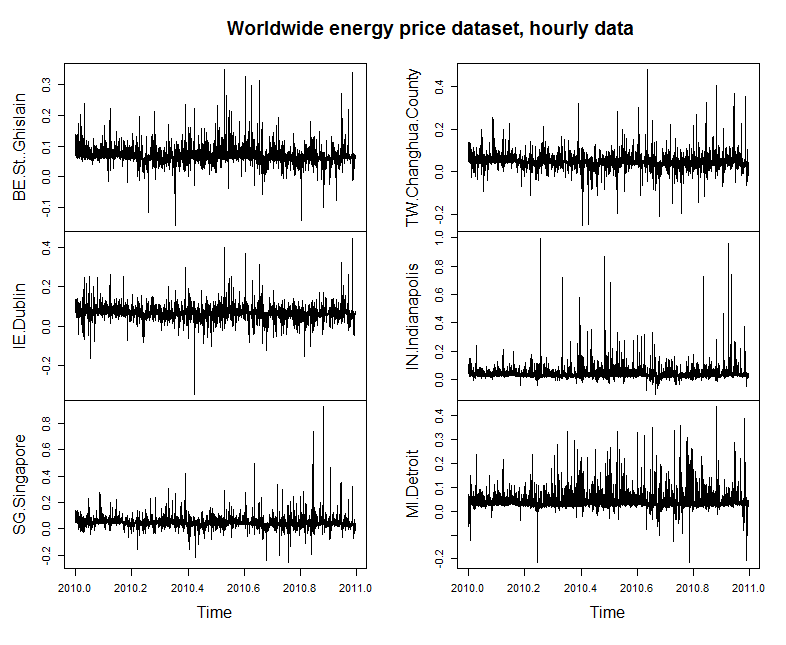


Figure 11: Time series in hourly intervals, worldwide dataset, 2010

Two of these series, namely the Detroit and Indianapolis time series are taken directly from the US energy market MISO, the other series are artificially generated based on the information of the real time series. All time series are shown for the whole year of 2010.

What still can be observed is that all time series exhibit some similarities in their distributions. For example, spikes exist from April to June and October to December for nearly each time series.

The corresponding temperature data sets for the same locations are outlined in Figure 12.

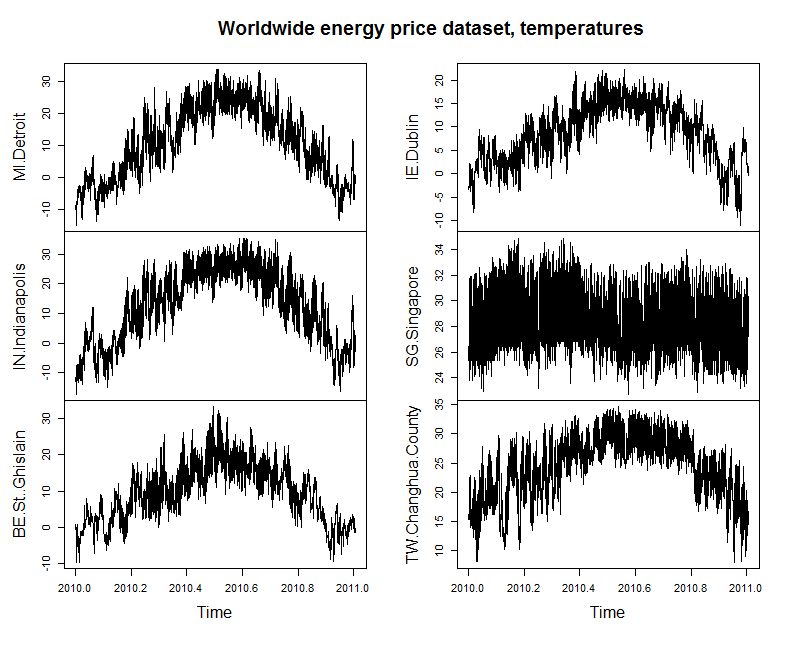


Figure 12: Temperature time series for the worldwide dataset, 2010

Not surprisingly, there is a clearly visible seasonality for the time series of temperatures where temperatures rise during summer and decrease in winter. An exception is the time series of Singapore, where the series stays almost constant throughout the year.

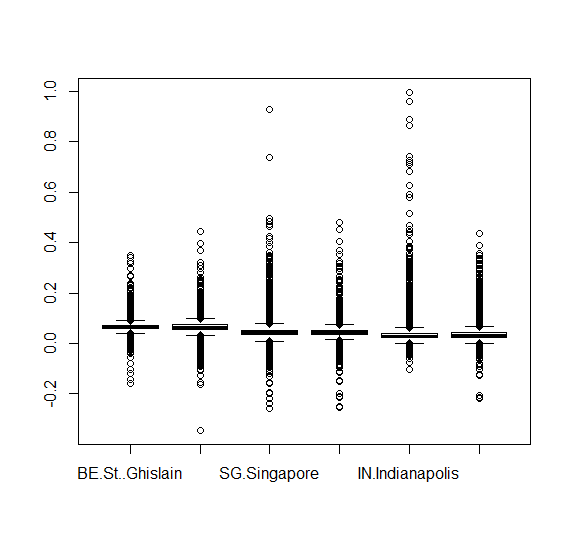
Looking at the boxplots (Figure 13) it can be observed that none of the series exhibits constant variation from the mean which means that the series have to be made stationary before applying prediction models.

Figure 13: Boxplots of the worldwide energy price dataset, 2010

The boxplots for the temperature time series exhibit much more variation due to different climate conditions in different continents.

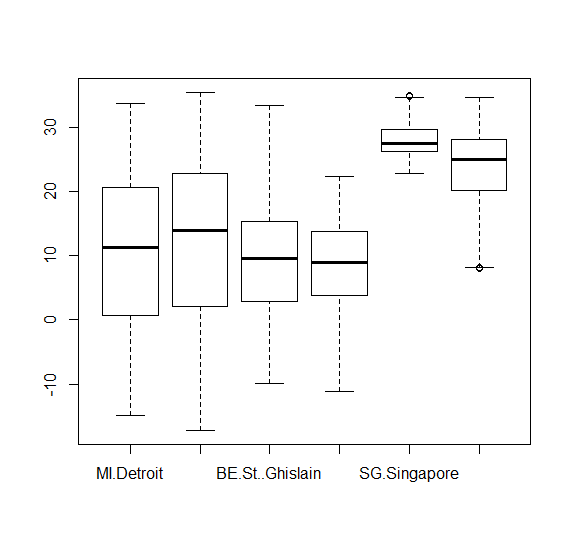


Figure 14: Boxplots of temperatures, worldwide dataset, 2010

### Dataset with 5-minute time intervals

Similar results can be observed for the dataset in 5-minute time intervals. This dataset consists of 15 different energy price time series of 2010 from the US energy markets (Figure 15 and Figure 16).

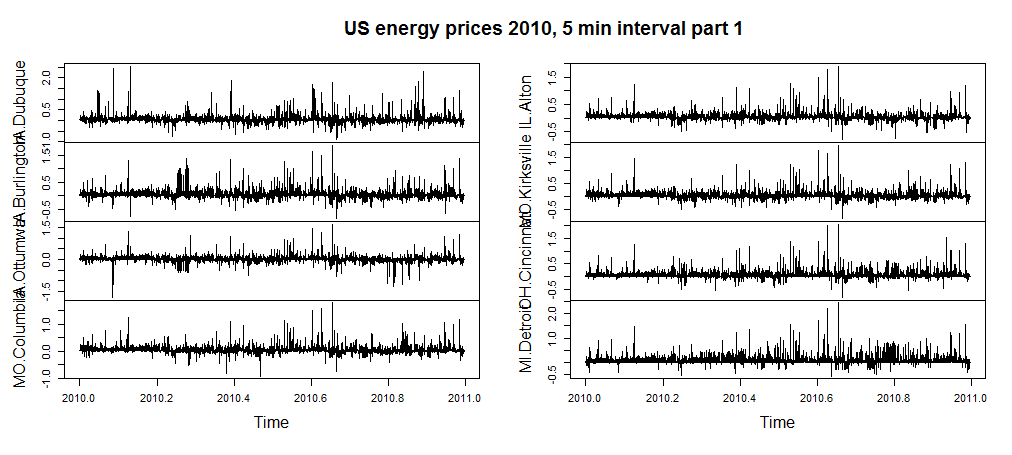


Figure 15: Energy price data, 2010, five minute intervals part 1

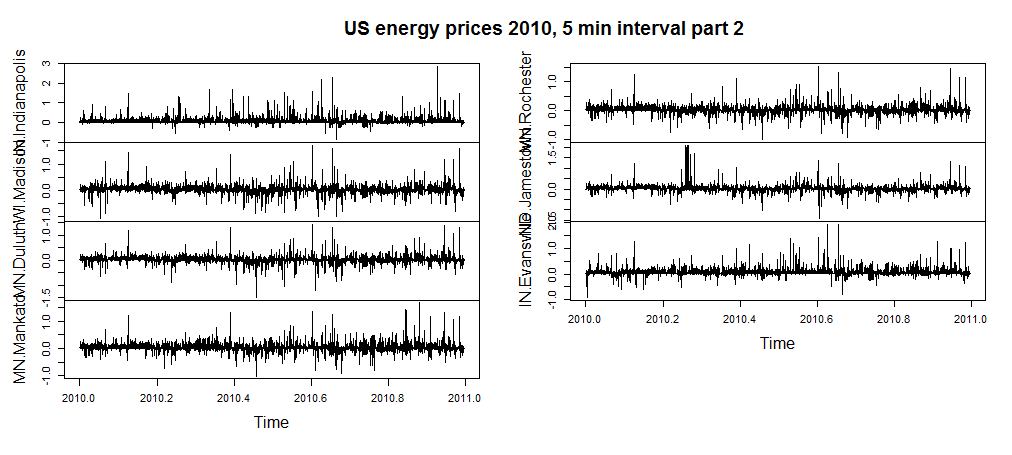


Figure 16: Energy price data, 2010, five minute intervals part 2

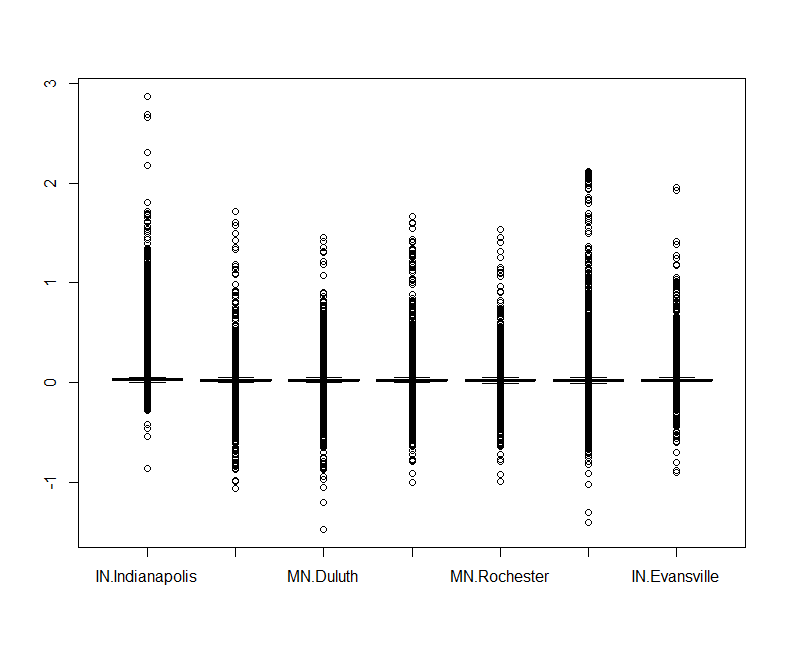
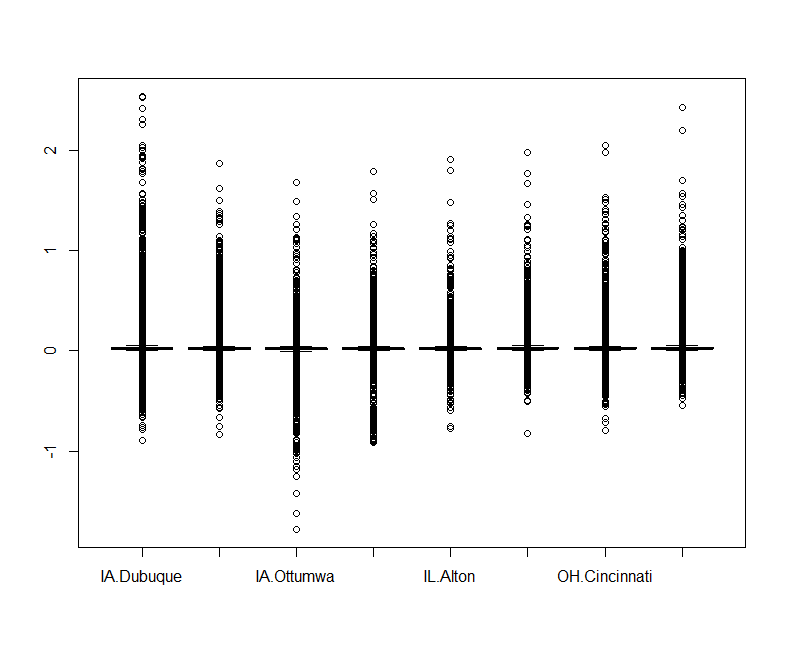
When comparing these datasets there is an obvious similarity concerning the times of the appearance of price spikes. This may be due to the fact that all of these time series are taken from the same energy market. There may be regional differences but one can observe strong similarities as well.

Figure 17: Boxplot of US energy data, 2010, 5 min int. part 1

Figure 18: Boxplot of US energy data, 2010, 5 min int. part 2

A simple test for stationarity is to print boxplots of the data and observe the variation from the mean. If datapoints are distributed evenly on both sides of the plots the series can be assumed to be stationary. As can be observed from the series in Figure 17 and Figure 18 most of the series exhibit a clearly visible skewness regarding the mean which implies that the series can not be considered stationary.

## Model Verification

Prediction models can be verified whether they provide a good fit to the given data by applying accuracy measurements and investigating certain characteristics of the models.

### Accuracy

Calculating accuracy of forecasting models is a way to see how well the model captures the characteristics of the examined dataset. The key for calculating the accuracy of a model is to investigate the value difference between the actual and forecasted or fitted data points. When a model is fitted to a timeseries (training set) there is a goodness-of-fit measure that compares the actual to the fitted data points within the series. The difference between the actual and fitted data points denotes the errors of in-sample data. When predicting data points outside of the training set the difference between actual and predicted values are called forecast errors of the out-of-sample data. The gaps between the actual and fitted or predicted data points are called residuals or forecast errors, from which accuracy measures can be directly derived.

### Residual characteristics

Residuals or forecast errors should have two characteristics such that the model is ensured to be using all available information from the data:

* Residuals should be uncorrelated such that there is no interdependence between any residuals in the time series
* Residuals should exhibit a zero mean

In addition, for easier calculation of prediction intervals, residuals should exhibit the following two characteristics:

* Residuals are normally distributed over the given timeframe
* Residuals show constant variance

The first two features are crucial for the validity of any forecasting model. If one of these features is false the prediction model can most likely be improved. Compliance with the last two points gives affirmation that the prediction model provides a sufficiently good fit for the dataset, however that does not mean that it can not be improved.

#### Checking Correlation of forecasting errors

In order to check whether residuals are correlated and thus fulfill the first of the above requirements there are two options:

* Plot a correlogram that shows correlations of forecast errors at different intervals (lags) whereby the values should not exceed the significance bounds. If more than one in 20 (5%) of these values clearly exceed the significance bounds the series is said to be correlated.
* Perform a significance test to formally check for evidence of correlation. This test exhibits certain characteristics where the series is more likely to be correlated when the results of the test are below ore above a certain threshold.

##### Correlograms

A correlogram of residuals is obtained by calculating an auto-correlation function on the set of residuals for a minium and maximum amount of lags.

Typically, the autocorrelograms are created based on a maximum of 20 lags as this provides a good estimation and overview about the actual correlations present in the dataset.

Each application of an auto correlation function provides values for each of the 20 lags and significance bounds where correlation values are expected to stay within these bounds for sufficient evidence of non-correlation.

For example, in Figure 19 we see the auto correlogram for the residuals of the ARIMA model applied to two weeks of the Detroit energy data series. On the horizontal axis the lag values are plotted whereas on the vertical axis the output of the autocorrelation function is plotted for each lag.

Significance bounds are plotted as dashed blue lines indicating the lower and upper significance bounds, respectively. It can be observed that two lag values clearly exceed the significance bounds (lags 16 and 18) and another just touches the bounds (lag 5).

From this we can deduce that the ARIMA model applied to this series can possibly improved as the residuals show some correlations. However, 1 in 20 lags are expected to exceed the significance bounds by chance alone, so this model still provides a reasonably good fit for this dataset.

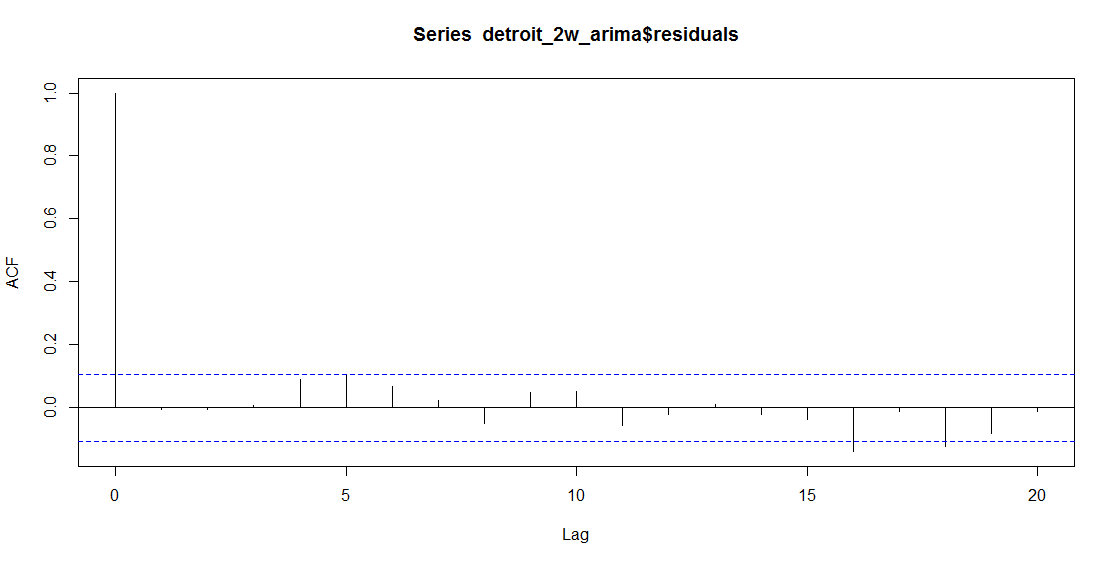


Figure 19: Residual correlations for 2 weeks of detroit energy data, 2010, applied to an ARIMA model

##### Autocorrelation tests

A significance test for autocorrelation is carried out to test if there is significant evidence for the set of residuals being uncorrelated. This can be done by the Ljung-box test which is a statistical test for autocorrelation of a time series.

##### Ljung Box test

Ljung box tests are a common method to testing for evidence of non-zero autocorrelations within a dataset. In our case this is useful to determine if the residuals calculated from the difference between the actual and the forecasted values exhibit some kind of correlation. If this is the case it is highly probable that the forecasting model can be improved as it does not use all available information in the dataset.

In the following we show the results of performing Ljung box tests on a dataset and applied forecasting models. In general, the greater the X-squared distribution value and the smaller the p-value output of the test the greater the existing correlation is in the given dataset and model.

Dataset based on hourly intervals

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Time Range | X^2 Distribution | Degrees of Freedom | P-Value |
| SES |  |  |  |  |
|  | 1 week | 32.6148 | 20 | 0.03717 |
|  | 2 weeks | 30.371 | 20 | 0.06406 |
|  | 1 month | 60.6431 | 20 | 5.661e-06 |
|  | 3 months | 99.645 | 20 | 1.458e-12 |
|  | 6 months | 179.5857 | 20 | < 2.2e-16 |
| Holt’s ES |  |  |  |  |
|  | 1 week | 33.2005 | 20 | 0.03207 |
|  | 2 weeks | 28.0152 | 20 | 0.109 |
|  | 1 month | 53.5716 | 20 | 6.694e-05 |
|  | 3 months | 73.0567 | 20 | 5.728e-08 |
|  | 6 months | 128.2554 | 20 | < 2.2e-16 |
| ARIMA |  |  |  |  |
|  | 1 week | 26.2372 | 20 | 0.1581 |
|  | 2 weeks | 27.7728 | 20 | 0.1149 |
|  | 1 month | 24.7486 | 20 | 0.2112 |
|  | 3 months | 22.8967 | 20 | 0.2939 |
|  | 6 months | 20.8204 | 20 | 0.4078 |

Here a dataset of hourly timed intervals between datapoints is examined where each model is tested for various time ranges (1 week up to 6 months). Each calculation includes the value of the X^2 Distribution, the degrees of freedom and the p-value, indicating the result of the test.

X^2 Distribution and p-value are strictly correlated where the p-value rises when the X^2 Distribution value decreases. Degrees of freedom denote the number of lags that are taken into account.

Interestingly, SES and Holt’s ES models exhibit increasing evidence of correlation (smaller p-value) when investigating a longer time range in the dataset whereas the ARIMA model shows the exact reverse behaviour. This is due to the fact that ARIMA models already account for correlations within the dataset and generally provide a better fit for longer term observations.

Next, a dataset based on 5 minute interval data is examined and the Ljung box tests is applied. Results are shown below.

Dataset based on 5 minute intervals

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Time Range | X^2 Distribution | Degrees of Freedom | P-Value |
| SES |  |  |  |  |
|  | 1 week | 141.7276 | 20 | < 2.2e-16 |
|  | 2 weeks | 203.9125 | 20 | < 2.2e-16 |
|  | 1 month | 446.8861 | 20 | < 2.2e-16 |
|  | 3 months | 1263.304 | 20 | < 2.2e-16 |
|  | 6 months | 2825.756 | 20 | < 2.2e-16 |
| Holt’s ES |  |  |  |  |
|  | 1 week | 138.2416 | 20 | < 2.2e-16 |
|  | 2 weeks | 197.0041 | 20 | < 2.2e-16 |
|  | 1 month | 427.3847 | 20 | < 2.2e-16 |
|  | 3 months | 1205.281 | 20 | < 2.2e-16 |
|  | 6 months | 2810.618 | 20 | < 2.2e-16 |
| ARIMA |  |  |  |  |
|  | 1 week | 38.8164 | 20 | 0.007027 |
|  | **2 weeks** | **28.7566** | **20** | **0.09266** |
|  | 1 month | 30.3909 | 20 | 0.06376 |
|  | 3 months | 37.1348 | 20 | 0.01127 |
|  | 6 months | 67.6991 | 20 | 4.306e-07 |

Since there is much more data available in this dataset the correlation increases dramatically. Also, the minimum X^2 Distribution value for ARIMA models can now be observed at the 2 weeks interval, which sums up to the same amount of data when sampled hourly for six months.

Still, ARIMA models show much less evidence for correlations than the exponential smoothing models. The former provide a better fit for this data, however there might still be other models showing a better fit.

###### Ljung Box test explanation

What has to be noted for this test is that it takes an average of the lags up to the maximum stated lag[[2]](#footnote-2). Suppose there is only little correlation within the dataset except for lag 8, tests with maximum lags up to 7 will produce very low correlation results, estimating a good fit of the current model.

Tests with a maximum lag of 8 will show large correlation and decrease in power with increasing value of the maximum lag. Therefore it is adviced to produce a correlogram to also visually check correlations between data points. [[3]](#footnote-3)

For a more compact view of the results a summary is provided below for both error distribution and p-value results on the different datasets.

X^2 Distribution of Forecast Errors, hourly data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 1 week | 2 weeks | 1 month | 3 months | 6 months |
| SES | 32.6148 | 30.371 | 60.6431 | 99.645 | 179.5857 |
| Holt’s ES | 33.2005 | 28.0152 | 53.5716 | 73.0567 | 128.2554 |
| ARIMA | 26.2372 | 27.7728 | 24.7486 | 22.8967 | 20.8204 |

X^2 Distribution of Forecast Errors, 5 minute data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 1 week | 2 weeks | 1 month | 3 months | 6 months |
| SES | 141.7276 | 203.9125 | 446.8861 | 1263.304 | 2825.756 |
| Holt’s ES | 138.2416 | 197.0041 | 427.3847 | 1205.281 | 2810.618 |
| ARIMA | 38.8164 | 28.7566 | 30.3909 | 37.1348 | 67.6991 |

P-Value results, hourly data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 1 week | 2 weeks | 1 month | 3 months | 6 months |
| SES | 0.03717 | 0.06406 | 5.661e-06 | 1.458e-12 | < 2.2e-16 |
| Holt’s ES | 0.03207 | 0.109 | 6.694e-05 | 5.728e-08 | < 2.2e-16 |
| ARIMA | 0.1581 | 0.1149 | 0.2112 | 0.2939 | 0.4078 |

P-Value results, 5 minute data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 1 week | 2 weeks | 1 month | 3 months | 6 months |
| SES | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 |
| Holt’s ES | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 |
| ARIMA | 0.007027 | 0.09266 | 0.06376 | 0.01127 | 4.306e-07 |

### Accuracy measures

#### Root mean square error

The root mean square error is an absolute error measure that is based on absolute values of forecast errors. As the name implies, the sum of squared errors is calculated from which the rooted mean is derived.

The definition is

where T denotes the time range, the actual and the predicted value. This provides for a good estimate of the performance of the applied model.

The root mean square errors of the previously discussed datasets are shown below.

RMSE, hourly data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 1 week | 2 weeks | 1 month | 3 months | 6 months |
| SES | 0.02791716 | 0.02640844 | 0.0207025 | 0.01912534 | 0.02320032 |
| Holt’s ES | 0.02829608 | 0.02672659 | 0.02091181 | 0.01931068 | 0.02335705 |
| ARIMA | 0.02651722 | 0.02548897 | 0.01969798 | 0.01843001 | 0.02242911 |

RMSE, 5 minute data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | 1 week | 2 weeks | 1 month | 3 months | 6 months |
| SES | 0.03732016 | 0.0360219 | 0.03147395 | 0.02916766 | 0.03613269 |
| Holt’s ES | 0.03740848 | 0.03608527 | 0.03151905 | 0.02919611 | 0.03615595 |
| ARIMA | 0.03564885 | 0.03450666 | 0.0301098 | 0.02780282 | 0.03412708 |

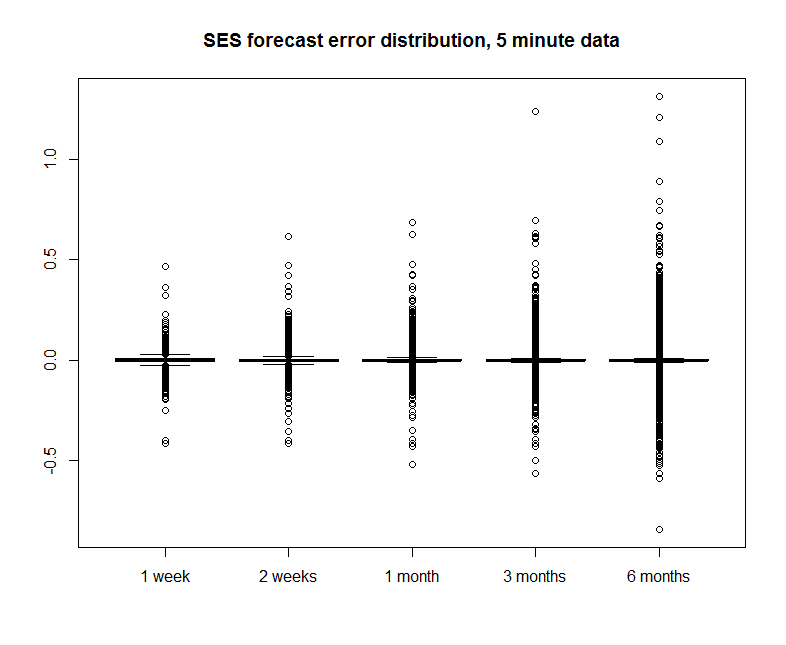
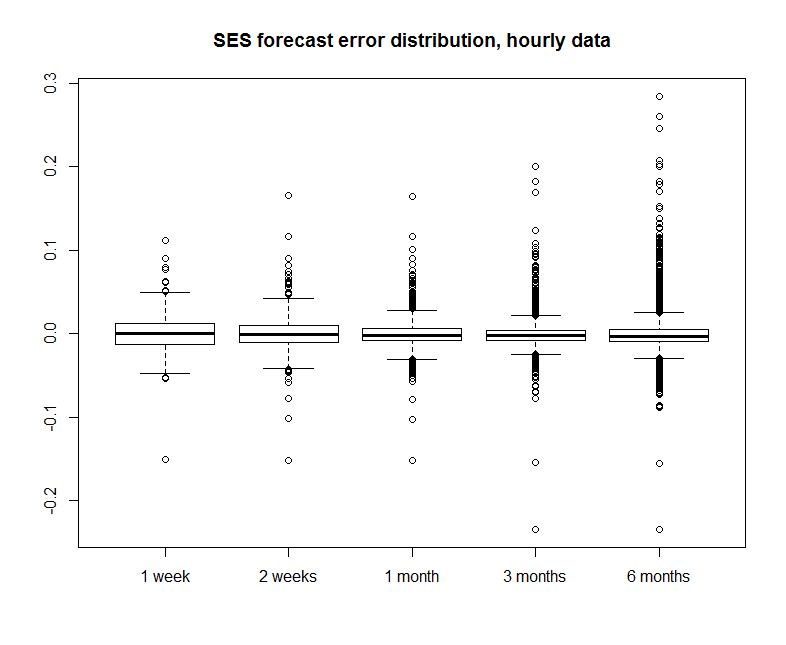
Thus, there is also evidence that ARIMA models outperform exponential smoothing models for both datasets.

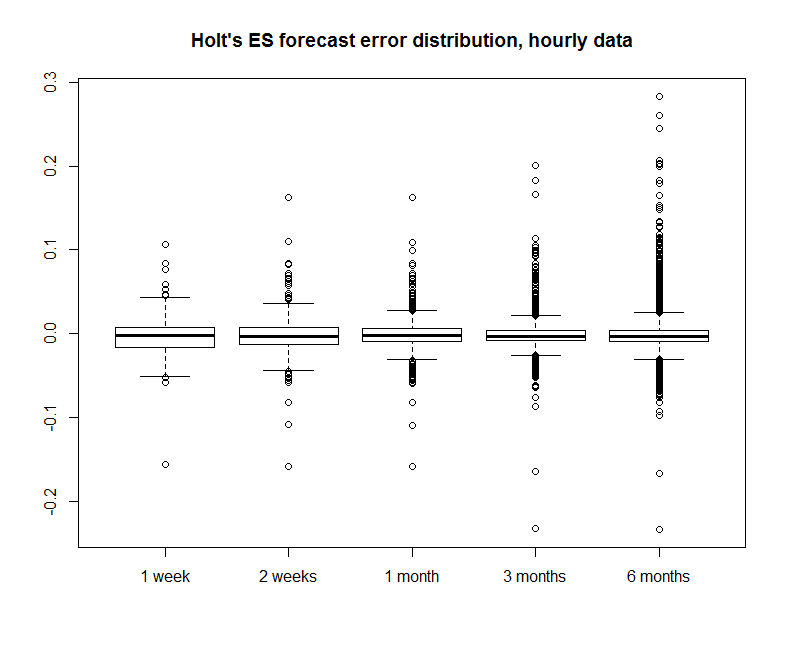
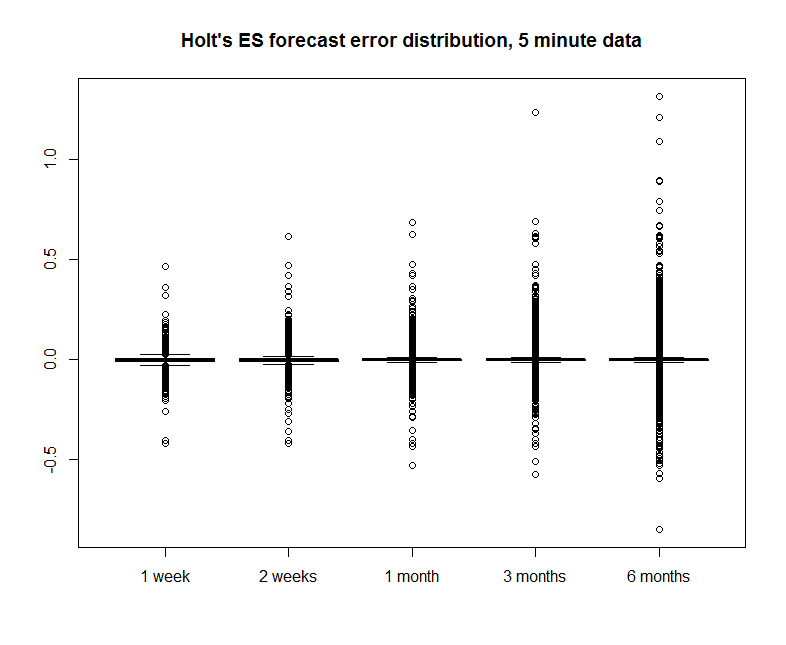
#### Forecast error distribution

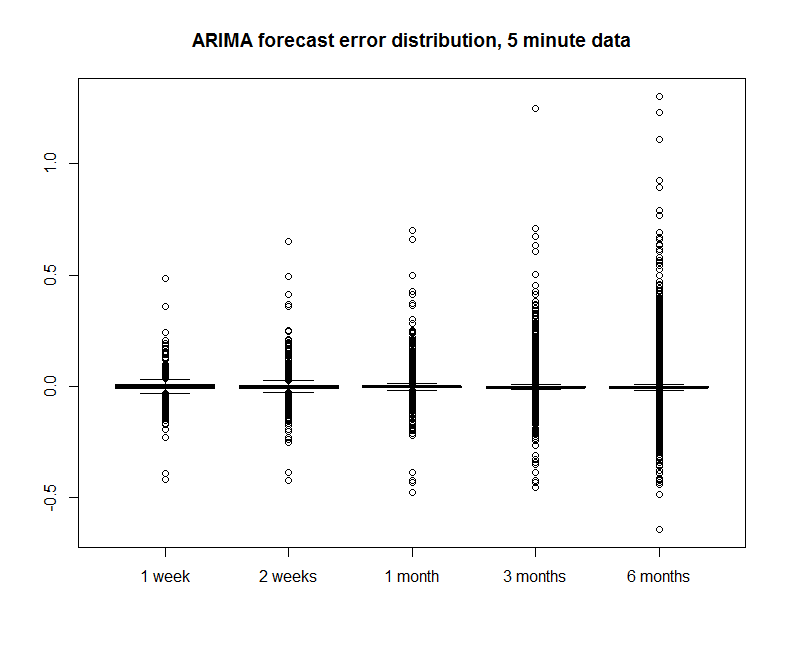
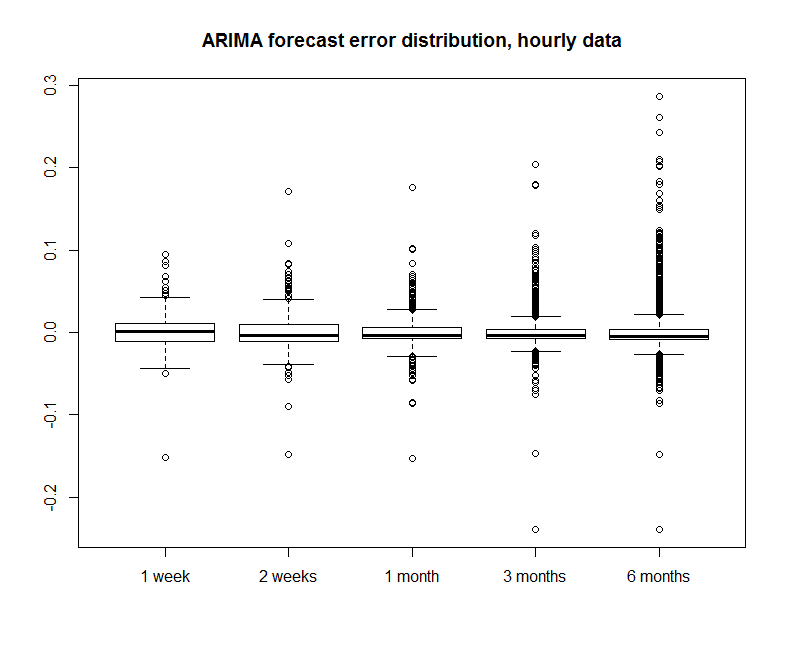
For an overview of the distribution of forecast errors the corresponding boxplots are shown below for each model and dataset.

What can be deduced from the results is that the distribution of the forecast errors is biased which is clearly observed for the hourly time series data. For 5 minute datasets the plots are more evenly distributed around the mean, though variation is still present.

Therefore all of the applied models can possibly be improved to provide a better fit for this data.

SES forecast error distributions

Holt’s ES forecast error distributions

ARIMA forecast error distribution

1. Introduction to ARIMA time series modeling: <https://www.youtube.com/watch?v=Y2khrpVo6qI> [↑](#footnote-ref-1)
2. Source: <https://stat.ethz.ch/pipermail/r-help/2004-April/049548.html> [↑](#footnote-ref-2)
3. Further information and papers regarding this topic may be found here:

   Burns Statistics

   patrick at burns-stat.com

   +44 (0)20 8525 0696

   <http://www.burns-stat.com> [↑](#footnote-ref-3)