Empirical Comparison of the Performance of Structural Time Series Methods in Forecasting Daily Temperature: Case of Johannesburg, South Africa.

By

Thokozani Memela

Supervisor

Dr Alemtsehai Turasie

A research report submitted to the Faculty of Science, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science

January 16, 2019

Contents

List of Figures

List of Tables

Declaration

I, Thokozani Eugen Memela, declare that this research report is my own work. It is being submitted for the Degree of Master of Science to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Dr Alemtsehai Turasie for his guidance, inspiration and support he provided. Thanks must also go to South African Weather Service for providing the data used in this report.

Secondly, I would like to give thanks to my employer WesBank and in particular to my boss, Ms Yoricke Esterhuyse as well as Mr Pierre-Johan Bezuidenhout for their support and for giving me flexible working hours.

Lastly, I would like to thank my partner Ms Sibusisiwe Hadebe and my daughter of 3 years old Ms Amile Memela for their support, patience, understanding and love.

Abstract

The Aim of this study was to compare the forecasting accuracy of exponential smoothing methods and unobserved component methods in forecasting Johannesburg daily temperature. The other objective of this study was to assess the effect of aggregating daily temperature to monthly temperature has on forecasting accuracy of the two structural time series methods. The Johannesburg daily temperature time series used spanned from 01 March 2007 to 31 March 2017. An extension of Holt-Winters model know as TBATS(Trigonometric Fourier representations, Box-Cox transformations, ARMA errors, Trend, and Seasonal component) by De Livera et al. (2011) was found to be more accurate to forecasts Johannesburg daily temperature. This model had high accuracy in Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

The study also found that the two structural time series methods were sensitive to time series classical components present in the data. The monthly temperature data was much smoother than daily temperature data. The two structural time series models used were much accurate in identifying the classical components of a smoother time series. This resulted in much accurate forecast. Holt-Winters additive seasonal model was found to be more accurate to forecast Johannesburg monthly temperature. This model out-performed local linear trend plus Fourier seasonal unobserved component model with high accuracy in Mean Absolute Error (MAE), Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE).

Chapter 1

Introduction

Various statistical methods are proposed to model and forecast time series. Some of these include unobserved component methods, non-linear methods, artificial neural network methods, vector time series methods, etc. Most popular and frequently used include exponential smoothing methods and Autoregressive Integrated Moving Average (ARIMA) methods. Time series methods can be classified into two groups, those that make assumptions about the stochastic process generating the observed time series such as ARIMA methods. Given a non-stationary time series, ARIMA methods tend to stress much importance on stationarity of a time series before a model is proposed and used to generate future forecasts. One can argue that this property of ARIMA methods reduces some information about the initial observed stochastic process. From this observation arises other set of time series methods that decomposes the observed time series into classical components such as, trend, seasonality and cycles. These classical components are then used to find the best model that fit the observed stochastic process (Koopman and Durbin, 2001).

Do unobserved component methods outperform exponential smoothing methods in generating reliable forecasts with application to daily temperature time series? The intent of this study is to empirically compare the forecasting accuracy of the two methods in generating daily temperature forecasts. The main difference between the two methods in generating forecasts is that unobserved component methods tend to treat all observations as equally important to the forecast, while exponential smoothing methods tend to give more weights to recent observations in generating the forecasts (Harvey, 2006). The ideas of exponential smoothing methods seem to have been developed by Robert G. Brown in about 1944. Much research has been seen where exponential smoothing methods have been used to forecast different kinds of time series data. These also extend to temperature or weather related data. Unobserved component model were first introduced in the econometrics and statistics fields by Andrew C. Harvey in 1989. Not much research is seen where unobserved component methods are used in forecasting temperature or weather related time series. Below we highlight some of the work done using these two methods in forecasting time series data.

Noor and Aniza (2017) compared simple exponential smoothing model, double exponential smoothing model and Holt's model to assess which model was suitable to forecast daily temperature level data for three cities in Texas, USA. Model performance was assessed using Sum of Squared Error (SSE), Mean Square Error (MSE) and Root Mean Square Error (RMSE). Double exponential smoothing model was found to be suitable for handling temperature level data as it produced lowest value of MSE and RMSE compared to the other two models. Simple exponential smoothing and Holt's models were found to be sensitive to data that was changing rapidly.

Rajarathinam et al.(2016) used unobserved component model to study trends in area, production and productivity of wheat crops grown in India during the period from 1950 to 2014. Linear regression model was used to test significance of trend component in the area, production and productivity time series data. A local linear trend unobserved component model was used where level component was modelled as deterministic while slope was model as stochastic. Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Percentage Error (MPE) were calculated for the model. Results revealed that the unobserved component model with slope and zero variance seemed to be an appropriate model for observing the trends in the area, production and productivity of wheat crop data. High values of R^2 obtained for the estimated model showed the closeness of the estimates to the actual values.

Motivated by the fact that every farm operation, process of plant growth and development as well as the yield of crop are strongly affected by weather conditions, Murat et al.(2015) used exponential smoothing methods to determine the most suitable exponential smoothing models to be used in generating forecast for air temperature, wind speed, and precipitation time series. A 31 years data for four European countries, Finland, Germany, Spain and Poland was used. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were forecasting accuracy measures used to select an appropriate model. The study found that to predict air temperature and wind speed, seasonal exponential with no trend model was to be preferred. The study also highlighted that it is advisable to use different model performance accuracy measures to assess the overall model forecasting error.

Gundalia and Dholakia (2012) used Holt-Winters exponential smoothing model to predict maximum and minimum temperature for Junagadh region of India. Maximum temperature data used spanned from June 1984 to October 2010. Minimum temperature data used spanned from June 1987 to October 2010. The Holt-Winters model forecasting performance accuracy was measured by Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD). Results of the study showed that maximum temperature exhibited less fluctuations and gave better results as compared to minimum temperature.

1.1 Aim and Objectives

The main aim of this study is to compare the two structural time series methods, that is, unobserved component methods and exponential smoothing methods for their accuracy in generating daily temperature forecasts.

The objectives of this study are to:

- Apply exponential smoothing time series methods to daily temperature data and generate two weeks ahead daily temperature forecasts,
- Apply the unobserved component time series methods to daily temperature data and generate two weeks ahead daily temperature forecasts,
- Apply temporal aggregation to the main temperature time series data which will result into monthly average temperature time series and apply the two structural time series methods to generate monthly average temperature forecasts,
- Compare the forecasting accuracy of the two structural time series methods in forecasting daily temperature and monthly temperature.

1.2 Motivation

Since the early 1960s forecasting has played an increasingly prominent role in all types of organizational planning (Makridakis and Wheelwright, 1989). As the world moves more towards globalization and as developing countries economies are getting stronger each day, it is clear that competition amongst organisations will intensify (Jenkins, 1982). This increases the demand for forecasting as organisations would like to be ahead of time in understanding what the future might look like tomorrow given today. This highlights the need for researchers to continue to search for better and reliable forecasting methods. With the change in technology, computer power and the availability of data increasing as the result of changing world, it is also important to continue forecasting and improving on the current forecasting methods and close any gaps that might be observed in the current forecasting methods.

This study will add to the existing statistical literature on forecasting with structural time series by developing an unobserved component model and exponential smoothing model which best explain the observed time series and providing statistical theoretical justification of the questions posed in the aim and objectives of this study. The study might be of interest to academics, researchers, management and weather forecasters.

1.3 Organization of the Report

The research report is organised as follows. Chapter 2 discusses theoretical background of the structural time series methods used in this study. This chapter starts by describing some basic concepts of time series in the context of proposed methods, then discusses exponential smoothing model as well as unobserved component model used in this study. Model forecasting performance measures that are used to compare the two structural time series models are presented in Section 2.4. Chapter 3 presents daily temperature data and analysis done. Chapter 4 concludes with a summary of findings and discusses possible areas for future research.

Chapter 2

Methodology

In this chapter we discuss theoretical background of the two structural time series methods we use. We begin by discussing basic time series concepts in the context of this study in Section 2.1. Sections 2.2 and 2.3 discusses theoretic background of exponential smoothing and unobserved component methods respectively. We end the chapter in Section 2.4 where we discuss forecasting performance measures that we use to compare the two structural time series methods.

2.1 Basic Time Series Concepts

A time series data arises when sequence of data points for a particular variable of interest are collected over time such as, daily recording of temperature. Mathematically, this is a set of vectors $Y(t)$, $t = 0, 1, 2, ..., T$, where t represents time that has passed. One of the basic assumption of studying time series is that some aspect of the observed past pattern will continue into the future.

A time series data is assumed to be affected by four components, trend, seasonal, cyclical and irregular or error term. The trend component describes observed long term increase or decrease in a time series. A time series is said to be influenced by a seasonal component if there are observed regular periodic fluctuations. A time series is affected by a cyclical component if it exhibits long-term variations that repeat in a reasonably systematic way over time. The irregular component results from other variations of the time series that are not identified as part of trend, seasonal and cyclical components. The effect of these components are generally described by two models, that is, multiplicative and additive models.

$$Multiplicative\ Model : Y(t) = T(t).S(t).C(t).I(t) \tag{2.1}$$

$$Additive\ Model : Y(t) = T(t) + S(t) + C(t) + I(t) \tag{2.2}$$

where $Y(t)$ is the observed time series data and $T(t)$, $S(t)$, $C(t)$, $I(t)$ are trend, seasonal, cyclical and irregular components respectively (Hyndman et al., 2008). Different methods of modelling and forecasting time series are proposed in literature. Structural time series methods namely exponential smoothing methods and unobserved component methods are the ones we use in this study.

2.1.1 State Space Models

State space models are an important tool in many time series analysis. The two structural time series methods that we use in this study can be presented in a state space model. State space model allows for calculation of maximum likelihood estimation of the unknown model parameters and calculation of prediction intervals for the model. Full exposition of state space model can be found in (Hyndman et al., 2008), (Harvey, 1989), (Durbin and Koopman, 2001) and (Commandeur and Koopman, 2007). The linear state space model is defined as ,

$$Y_t = Z_t x_{t-1} + \varepsilon_t\,, \qquad \varepsilon_t \sim NID(0, \sigma^2)\,, \tag{2.3}$$

$$x_t = K_t x_{t-1} + R_t \varepsilon_t\,, \qquad t = 1, \ldots, n \tag{2.4}$$

where Y_t denotes the observed values of the time series at time t and x_t is the state vector of unknown model parameters, ε_t is the error term of Y_t which is assumed to follow a Gaussian white noise process with variance σ^2. By appropriate definitions of the vectors, Z_t, x_t and R_t, and of the matrix K_t, structural time series models that we use in this study are derived as a special case of equation (2.3) and (2.4).

2.2 Exponential Smoothing Methods

Some of the best forecasting methods in the literature of time series forecasting are based on the fundamental concepts of exponential smoothing. The main ideas of exponential smoothing methods are that the value of data degrades over time with newer observations weighted more heavily than older observations in the forecast (Hyndman et al., 2008). When modelling time series data with exponential smoothing methods we usually start with the level L_t and trend T_t. The level and the trend can be combined in different ways giving in particular five future trend models for a time series. The five trend type models can involve a model with only the level component, a model with additive trend, a model with multiplicative trend, a model with additive damped trend and a model with multiplicative damped trend. A full exposition of these can be found in Hyndman et al. (2008). Having chosen the trend model a seasonality component S_t may also be introduced in the model either multiplicative or additive.

As a result of the above, Pegels (1969) firstly defined a classification of exponential smoothing methods and called this classification taxonomy. The taxonomy by Pegel (1969) has since went through different modifications, first by Gardener (1985) then by Hyndman et al. (2002) and more recently by Taylor (2003). These modifications have resulted in a comprehensive list of all possible variants of exponential smoothing methods presented in Table 2.1.

Table 2.1: Taxonomy of Exponential Smoothing methods.

		Seasonal		
		N(None)	A(Additive)	M(Multiplicative)
Trend	N (None)	N,N	N,A	N,M
	A (Additive)	A,N	A,A	A,M
	$\mathrm{A_d}$(Additive Damped)	$\mathrm{A_d}$, N	$\mathrm{A_d}$, A	$\mathrm{A_d}$, M
	M (Multiplicative)	M,N	M,A	M,M
	$\mathrm{M_d}$(Multiplicative Damped)	$\mathrm{M_d}$, N	$\mathrm{M_d}$, A	$\mathrm{M_d}$, M

(N,N) describes the simple exponential smoothing method with no trend and no seasonality. According to Hyndman et al(2008, p12) Double exponential smoothing method (A,N) is a particular case of Holt's method. Triple or Holt-Winters additive and multiplicative exponential smoothing methods are given by (A,A) and (A,M). All other remaining models correspond to models that are less used in literature of exponential smoothing.

The nature of daily temperature suggest that temperature is strongly affected by seasonal component. Summer sees high temperature values and winter sees low temperature values. One can also assume that temperature is affected by trend component. This year's summer might be hotter than last year's summer and the same can be said for winter months. Due to the nature of time series in study a Holt-Winters additive model (A,A) and Holt-Winters multiplicative models (A,M) are two exponential smoothing models proposed to describe and forecast daily temperature.

2.2.1 Holt-Winters Models

The two Holt-Winters exponential smoothing models we use in this study are described below,

Multiplicative Seasonality Model:

$$Level : L_t = \alpha \frac{Y_t}{S_{t-m}} + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{2.5}$$

$$Trend : T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1} \tag{2.6}$$

$$Seasonality : S_t = \frac{\delta Y_t}{(L_{t-1} + T_{t-1})} + (1-\delta)S_{t-m} \tag{2.7}$$

$$Forecast : \widehat{Y}_{t+h|t} = (L_t + T_t h)S_{t-m+h_m^{+}{}'} \tag{2.8}$$

Additive Seasonality Model:

$$Level : L_t = \alpha(Y_t - S_{t-m}) + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{2.9}$$

$$Trend : T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1} \tag{2.10}$$

$$Seasonality : S_t = \delta(Y_t - L_{t-1} - T_{t-1}) + (1-\delta)S_{t-m} \tag{2.11}$$

$$Forecast : \widehat{Y}_{t+h|t} = L_t + T_t h + S_{t-m+h_m^{+}{}'} \tag{2.12}$$

where L_t represents the level of the series, T_t represents the trend, S_t is the seasonality and m is the length of seasonality. $\widehat{Y}_{t+h|t}$ is the forecast for h periods ahead, $h_m^{+}{}'$ = $[(h-1) \mod m]+1$. α, γ, δ are smoothing parameters and $0 < \alpha, \gamma, \delta < 1$ (Hyndman et al., 2008, p16).

2.2.2 Forecasting with Exponential Smoothing Methods

According to Chatfield (2013) and Hyndman et al (2008) the main steps that one would follow in setting up a Holt-Winters exponential forecasting model which we employ in forecasting daily temperature in this study are,

- Examine the daily temperature data and select all appropriate exponential smoothing models using significant time series components, that is, level, trend and seasonality.
- By using some heuristic scheme initialize the starting values for $L_0, T_0, S_0, \ldots, S_{-m+1}$,
- Estimate model smoothing parameters α, γ, and δ by minimizing the error sum of squares or other suitable measure,
- Select the best exponential smoothing model for daily temperature time series according to the minimum Akaike Information Criterion (AIC),
- Produce point forecasts using the appropriate model, and
- Assess forecasting accuracy of the estimated model using different forecasting measures.

2.3 Unobserved Component Models

Unobserved component models can be considered to be multiple regression models with time-varying coefficients (Durbin and Koopman, 2001). The basis of unobserved component time series method is that time series can be viewed as decomposition of trend, seasonal, and cyclical components (Commandeur and Koopman, 2007). Unobserved component models tend to give equal weight to both near and far distant observations. One of the important considerations that one needs to take into account when modelling time series data with unobserved component models is that, inefficient and inaccurate forecasts are likely to occur for someone who ignores the salient characteristics of the time series to be forecasted. If one builds a time series model that has no allowance for seasonal variation in it when there is a clear significant seasonality evidence in the observed time series, the forecasting accuracy of such model is likely to be poor. The univariate unobserved component time series model that we use in this study is given by,

$$Y_t = \mu_t + \gamma_t + \epsilon_t \quad , \epsilon_t \backsim NID(0, \sigma_\epsilon^2) \quad t = 1, \ldots, n \tag{2.13}$$

where, μ_t, γ_t and ϵ_t represent trend, seasonality and irregular components, respectively. The error component ϵ_t are normal independent distribution with mean 0 and variance σ^2. The trend μ_t and seasonal γ_t components can be formulated as either stochastic or deterministic. In a stochastic linear process the trend and seasonality are allowed to be flexible and change over time depending on the independent error terms, while in a deterministic modelling approach trend and seasonality are fixed over the entire time span of the series with error terms assumed to be zero (Koopman and Ooms, 2011). This study only look at stochastic modelling of daily temperature observed components.

2.3.1 Trend Component

Modelling of the stochastic trend can be formulated in two ways. One may choose to model the trend as a stochastic local level process or one may choose to model trend by extending the local level model to include a stochastic slope. We unpack these below,

Local Level Model: The stochastic local level component can be modelled as a random walk given by

$$Y_t = \mu_t + \epsilon_t, \quad \epsilon_t \backsim NID(0, \sigma_\epsilon^2) \tag{2.14}$$

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \backsim NID(0, \sigma_\eta^2) \tag{2.15}$$

where $t = 1, \ldots, n$, $\mu_t = l_t$ is the level at time t, ϵ_t is the disturbance at time t and η_t is the level disturbance at time t. The level disturbance series η_t are assumed to be independent and mutually independent of all the disturbances of Y_t.

The Local Lineal Trend Model: This model results from adding a slope component β_t into the local level model discussed above, that is,

$$Y_t = \mu_t + \epsilon_t, \quad \epsilon_t \backsim NID(0, \sigma_\epsilon^2) \quad , \tag{2.16}$$

$$\mu_{t+1} = \mu_t + \beta_t + \eta_t, \quad \eta_t \backsim NID(0, \sigma_\eta^2) \quad , \tag{2.17}$$

$$\beta_{t+1} = \beta_t + \zeta_t, \quad \zeta_t \backsim NID(0, \sigma_\zeta^2) \quad , \tag{2.18}$$

for $t = 1, \ldots, n$. The level disturbance series η_t and ζ_t are assumed to be independent and mutually independent of all the disturbances of Y_t.

2.3.2 Seasonal Component

To account for seasonality in daily temperature modelling using unobserved component model we add seasonality γ_t to either local level model or local linear trend model. One can model seasonality as time varying component, fixed trigonometric component and time varying trigonometric component. In this study we model seasonality as time varying trigonometric component defined per Koopman and Ooms (2011) as,

$$\begin{bmatrix} \gamma_{j,t} \\ \gamma_{j,t+1}^{+} \end{bmatrix} = \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix} \begin{bmatrix} \gamma_{j,t} \\ \gamma_{j,t}^{+} \end{bmatrix} + \begin{bmatrix} \omega_{j,t} \\ \omega_{j,t}^{+} \end{bmatrix}, \quad \begin{bmatrix} \omega_{j,t} \\ \omega_{j,t}^{+} \end{bmatrix} \backsim NID(0, \sigma_\omega^2 I_2) \tag{2.19}$$

with $\lambda_j = 2\pi_j/S$ for $j = 1, ..., S/2$ and $t = 1, ..., n$. the $S - 1$ initial variables $\gamma_{j,1}$ and $\gamma_{j,1}^{+}$ are unknown coefficients. The seasonal error terms $\omega_{j,t}$ and $\omega_{j,t}^{+}$ are assumed to be independent of each other, and are also independent of all other error terms of Y_t (Koopman and Ooms, 2011) and (Harvey, 1989).

2.3.3 Forecasting with Unobserved Component Methods

Once an unobserved component model has been proposed and presented in state space, the next step is for one to estimate the unknown model parameters. The estimations of the state vector in state space model are carried out by using a Kalman filter algorithm which is applied to the observed times series. The main purpose of the Kalman filter is to obtain optimal values of the state at time point t by only considering the observations $\{Y_1, Y_2, \ldots, Y_{t-1}\}$ (Koopman and Ooms, 2011). A full exposition and derivation of the Kalman filter can be found in (Harvey, 1989) and (Durbin and Koopman, 2001). In summary the main steps that we use in this study in setting up a forecasting model using unobserved component are,

- Examine the daily temperature data and select all appropriate unobserved component models based on significant time series components, that is, level, trend and seasonality,
- Estimate the unknown state vector of the model by using the Kalman filter to minimize the mean square linear estimator (MMSLE) together with its mean squared error (MSE),
- Estimate the initial variables for unobserved component model by using a maximum likelihood estimation method (Durbin and Koopman, 2001),
- Select appropriate model based on the minimum Akaike Information Criterion (AIC) and model diagnostics measures,
- Produce point forecasts by continuing the Kalman filter after the end of the observed times series,
- Assess forecasting accuracy of the estimated model using different forecasting error measures.

2.4 Model Diagnostics and Forecasting Error Measures

Forecasting refers to the process of predicting the future outcome of an event. In this section we outline the theoretical background of time series performance measures that we use to compare the two structural time series methods. We also discuss cross validation method and Portmanteau tests that we use in this study.

2.4.1 Diagnostic Checking

Diagnostic checking is an important step in model selection. In the context of structural time series methods we use in this study this involves finding whether the residuals of a selected model satisfy three properties, that is, independent, homoscedasticity and normally distributed. A plot of autocorrelation function (ACF) of the model residual indicate if the residuals of a selected model are independent or not. The ACF is not a sufficient model diagnostic mechanism as it is only able to show linear dependent structures in residuals. In time series analysis we are likely to have different non-linear structures present (Box and Pierce, 1970).

Box-Pierce Test

A portmanteau is the usual test statistics used to test for independence of residuals up to a lag j. Box and Pierce (1970) proposed a portmanteau test statistics that we use in this study to test for independence in model residuals. They defined this as,

$$Q = n \sum_{k=1}^{j} r_k^2 \tag{2.20}$$

where r_k is sample autocorrelation of order k of the residual, j represents the number of lags being tested and n represents number of observations. Under the null hypothesis that the residuals of the selected model are independently distributed, Q follows a χ^2 distribution with $(j - P)$ degrees of freedom. P is the model estimated parameters. The Box-Pierce test performs poorly with small samples.

Ljung-Box Test

Due to disadvantages of Box-Pierce test above Ljung and Box (1978) proposed a test that is able to handle small samples. They defined this as,

$$Q^* = n(n+2)\sum_{k=1}^{j}\frac{r_k^2}{n-k} \tag{2.21}$$

where r_k is sample autocorrelation of order k of the residual, j represents the number of lags being tested and n represent number of observations. Under the null hypothesis that the residuals of the selected model are independently distributed, Q^* follows a χ^2 distribution with $(j - P)$ degrees of freedom. P is the model estimated parameters.

Homoscedasticity

The second most important test on the model residual is homoscedasticity. The test is based on comparing the sum of squares of two independent set of samples. Homoscedasticity of the residuals can be checked with the following,

$$H(h) = \frac{\sum_{t=n-h+1}^{n} e_t^2}{\sum_{t=1}^{h} e_t^2} \tag{2.22}$$

under the hypothesis that model residual are heterescadastic, this statistic follows an F-distribution with (h, h) degrees of freedom, h is defined as the nearest integer to $\frac{n}{3}$ (Commandeur and Koopman, 2007). One can also use the ARCH LM test or Ljung-Box test on squared residuals.

Normality

The least important check is normality in model residual. We check this by using the test statistic,

$$N = n\left(\frac{W^2}{6} + \frac{(X-3)^2}{24}\right) \tag{2.23}$$

where W denotes the skewness of the residuals, and X the kurtosis. The test statistics is tested against a χ^2 distribution with two degrees of freedom (Commandeur and

Koopman, 2007). Another test one may use to test for normality in the residual is the QQ plot. The QQ plot is created by plotting two sets of quantiles against each other. A 45 degree line is used as a reference line. Residuals are normally distributed if there are close to this reference line. QQ plot is not a conclusive test and one need to use this test with caution. One can also use Shapiro-Wilk test or even Kolmogorov-Smirnov test.

2.4.2 Cross Validation Method

The performance of a statistical algorithm relates to how well the algorithm predict on an independent test data. Assessment of this performance is one of the critical steps in the process of modelling and gives guidance to which model to choose. Hastie et al.(2015) defined cross validation method which we use in this study as randomly dividing the dataset into two parts, a training set and test set. The statistical model is fitted on the training set, and the fitted model is then validated on the test set. The resulting test set error rate assessed using some quality measure like mean squared error provides an estimate of the model test error (Hastie et al., 2015). The assumption in this cross validation method is that there is sufficient historical data to use to split the data into training and test set.

2.4.3 Forecast Accuracy Measures

The performance of structural time series methods used in this study is assessed by considering their forecasting accuracy. Below we present accuracy measures discussed in Hyndman et al.(2002) that we use in this study.

Scale Dependent Measures

The one step ahead forecast error for any forecasting procedure is simply defined as $e_t = Y_t - \widehat{Y}_t$ where Y_t is observed time series at time t, $\widehat{Y}_t$ is predicted time series at time t and e_t is the error in prediction at time t between Y_t and $\widehat{Y}_t$. Accuracy measures based on e_t are called scale dependent measures. These are measures which are based on the same scale as the data values.

Mean Error (ME) is calculated as

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t, \tag{2.24}$$

Mean Absolute Error (MAE) is calculated as

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t|, \tag{2.25}$$

Mean Squared Error (MSE) is calculated as

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{2.26}$$

and Root Mean Square Error $RMSE = \sqrt{MSE}$,
where n represent number of observations.

Percentage Measures

Percentage accuracy measures are scale independent. There are mostly used in comparing the performance of time series that are measured on different scales. The two that we use are, Mean Percentage Error (MPE) defined as

$$MPE = \frac{100}{n}\sum_{t=1}^{n} \frac{e_t}{y_t} \tag{2.27}$$

and Mean Absolute Percentage Error (MAPE) defined as

$$MAPE = \frac{100}{n}\sum_{t=1}^{n} \frac{|e_t|}{y_t} \tag{2.28}$$

where n represents the number of observations and y_t is the actual values of the time series.

Chapter 3

Analysis and Results

In statistical modelling data exploration help one gain insight into the data and its underlying structures, understand distribution of data, check for missing values and identify influential data values. We begin this chapter in Section 3.1 where we present the average daily temperature time series that we use. We then discuss data exploration in Section 3.2. We end the chapter by presenting results obtained from both methods for average daily temperature data in Section 3.4 and for average monthly temperature data in Section 3.5.

3.1 Data Description

The average daily temperature data had 3684 observations recorded at the Johannesburg weather station over the period of 01 March 2007 to 31 March 2017. On a daily basis Johannesburg weather station records a minimum temperature value and a maximum temperature value for the day. The average daily temperature value for each day that we use in this study is calculated from these two values. Figure 3.1 below presents Johannesburg average daily temperature time series plot from 01 March 2007 to 31 March 2017. From this figure we see that the average daily temperature data has positive values over the entire sample period. Visual inspection of the data shows evidence of yearly seasonal pattern and upward trend. A total of 30 observations which represent less than 1% of the data were randomly missing. The missing values were replaced by fitting a K Nearest Neighbours algorithm (kNN) to the data. The average daily temperature values range approximately between $2\,^{\circ}\mathrm{C}$ and $29\,^{\circ}\mathrm{C}$. The random fluctuations of the series seem roughly constant in size over time. This behaviour suggest that perhaps an additive model with seasonal component might be appropriate in describing average daily temperature.

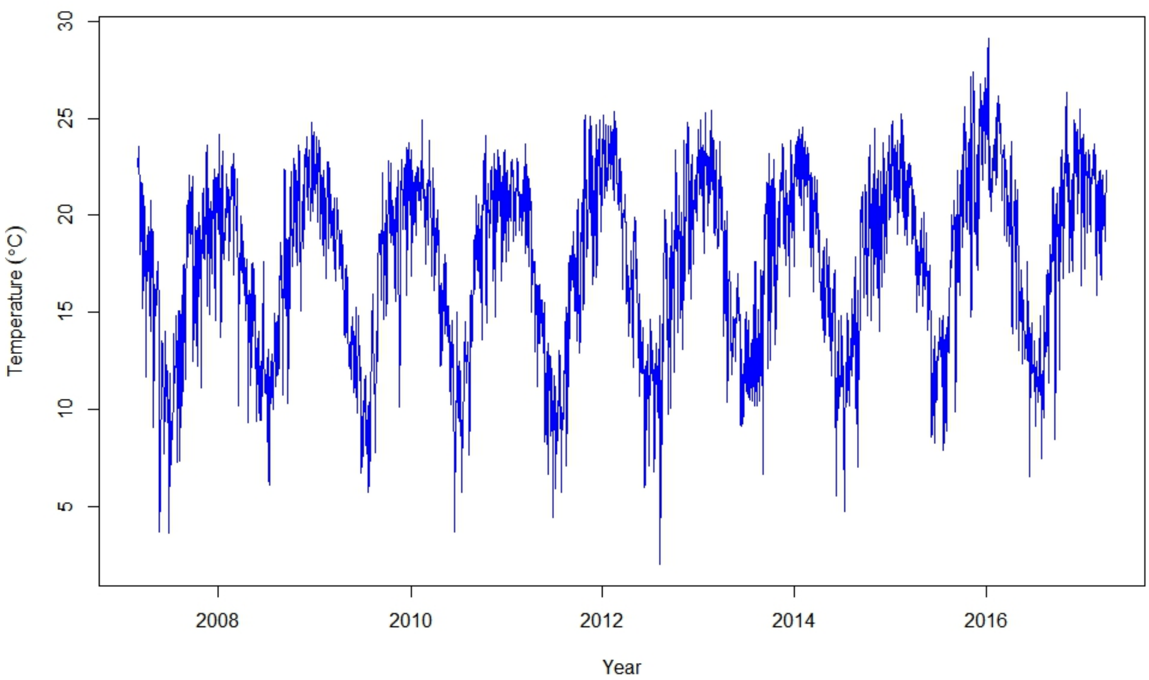

Figure 3.1: Johannesburg Average Daily Temperature from 01 March 2007 to 31 March 2017.

3.2 Data Exploration

We begin data exploration by presenting descriptive statistics for average daily temperature time series in Table 3.1 below. The data ranges between 2°C and 29°C. The mean daily temperature over the period is 17.6°C with median of 18.5°C and a spread of approximately 5°C. The skewness of the data is -0.444 indicating that the distribution of the data is slightly skewed to the left or negatively skewed. For kurtosis, we have 2.352 implying that we have heavier tailed distribution.

Table 3.1: Descriptive Statistics for Johannesburg Average Daily Temperature Data.

N	Mean	Median	St. Dev.	Min	Max	Skewness	kurtosis
3,684	17.612	18.45	4.467	2.000	29.150	-0.444	2.352

We explore the distribution of data over the sampling period and check for the presence of outliers by presenting an annual Box-and-Whisker plot in Figure 3.2 below. The data seem to fluctuate around the median temperature of 18°C. There is an indication of possible one outlying value in year 2010 of approximately 4°C. This value is within the range of our data as explained in Table 3.1 above and is not considered as outlying in this study. Even though in year 2010 this value was out of the range of other values one can accept that in winter temperature can reach 4°C and this was a coldest day in 2010. The plot also seem to suggest that 2017 was the hottest year, on average, over the study period.

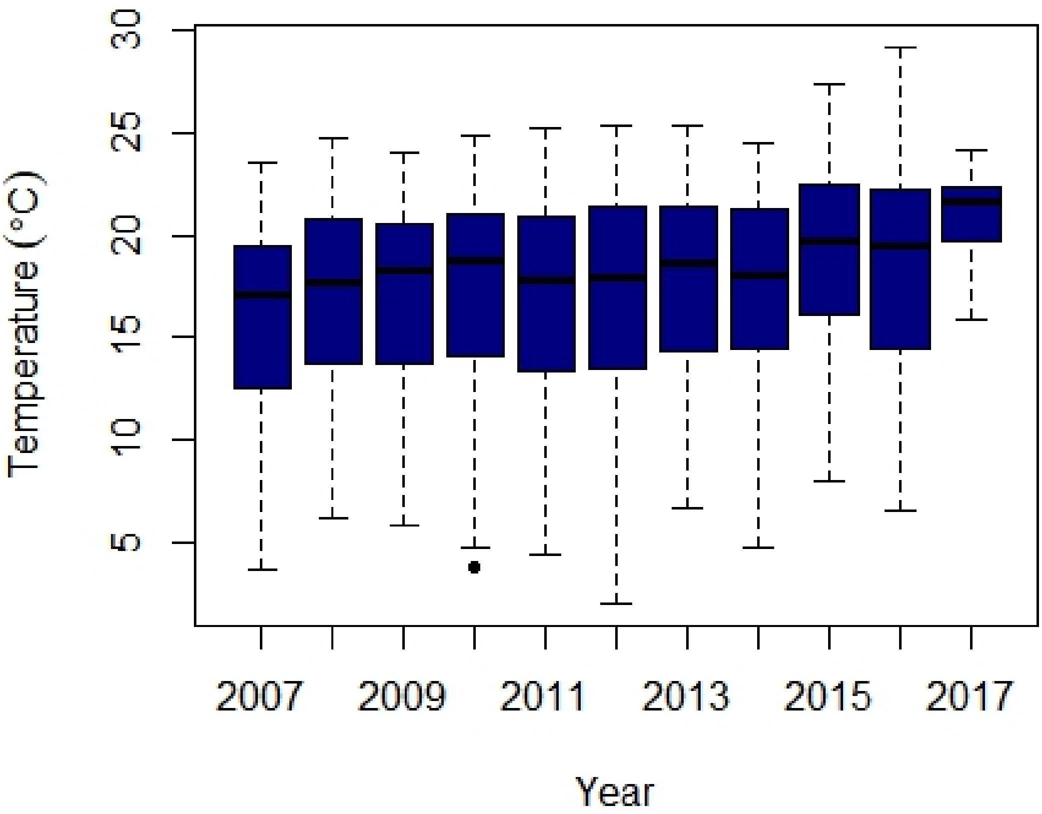

Figure 3.2: Johannesburg Average Daily Temperature Yearly Box-and-Whisker plot from 2007 to 2017.

We next discuss patterns that exist in our data. One of the assumptions made when studying time series data is that it is correlated. In order to assess if daily temperature time series values at different lags are correlated or not and also to understand the nature of correlation identified, we assess the correlation lag plot below for the first 9 lags of our data. Figure 3.3 below clearly shows a positive linear correlation between the current and previous 9 days of daily temperature time series.

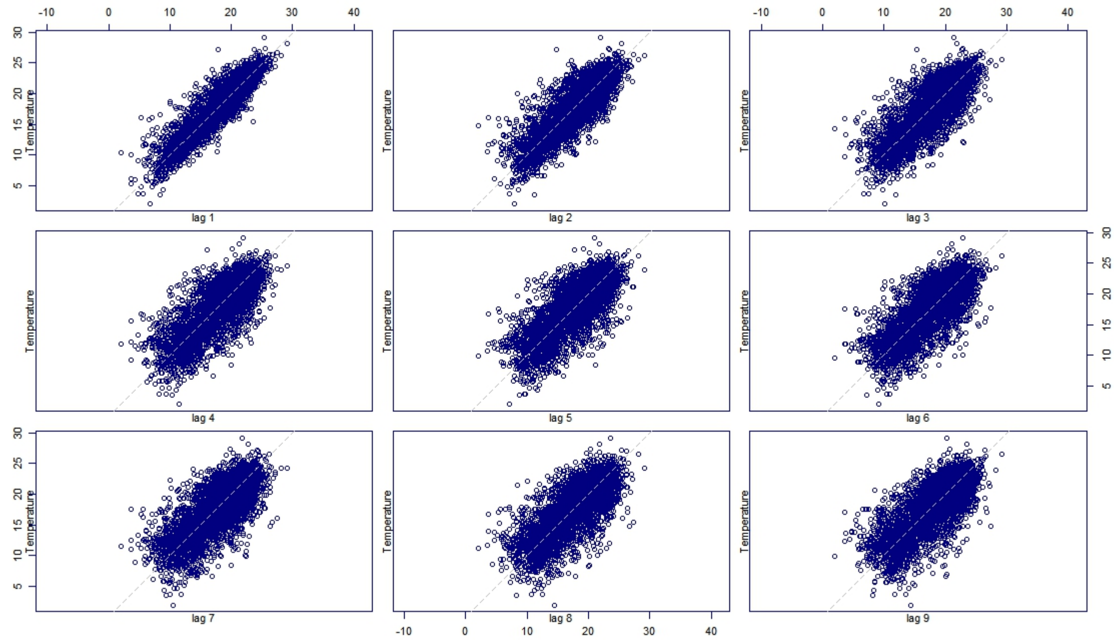

Figure 3.3: Average Daily Temperature Lagged Correlation Plot for the first 9 days.

We end data exploration by decomposing average daily temperature time series into seasonal, trend and irregular components using classical decomposition method. Classical decomposition method is used because seasonality seems to be constant year on year in daily temperature series. This is the fundamental assumption made when using classical decomposition method (Hyndman and Athanasopoulos, 2014). Figure 3.4 presents daily temperature decomposition. The plot label observed present original average daily temperature time series without decomposition. Trend component is presented in the second row labelled trend. The average behaviour of the trend over time is slowly increasing from the start of the series till about year 2015 where we see a significant upward rise which does not last long and then a downward movement is observed after that. Overall, trend component suggests upward trend in our data. Seasonal component is shown in the third row labelled seasonal. Seasonal component suggests a strong yearly seasonality exist in our data. The last row in Figure 3.4 presents random component which is neither systematic nor predictable.

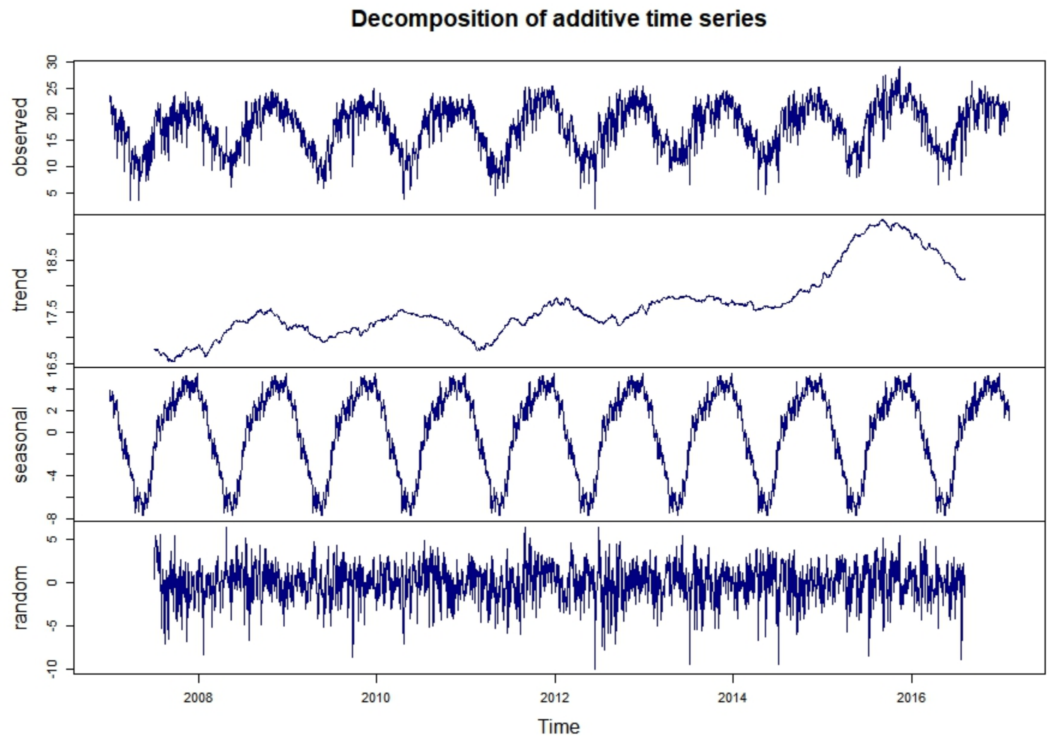

Figure 3.4: Classical Decomposition of Johannesburg Average Daily Temperature from 01 March 2007 to 31 March 2017.

3.3 Model Identification and Fitting

The average daily temperature time series is divided into two samples, 80% of data is used for model training and 20% of data is used for model testing. The model parameters are estimated using the training sample and the test sample is used to validate models. We first present exponential smoothing models and unobserved component models results for daily temperature data and then present respective results for monthly temperature data. R open source statistical software is used in modelling the two temperature time series in this study.

3.4 Daily Temperature Modelling Results

3.4.1 Exponential Smoothing Model Estimation

Holt-Winters additive seasonal model and Holt-Winters multiplicative seasonal model are fitted to the daily temperature data. Holt-Winters additive seasonal model is preferred when seasonal variation in the data are roughly constant through time, while the multiplicative seasonal method is preferred when seasonal variations are changing proportional to the level of the series (Hyndman and Athanasopoulos, 2014)). Data exploration in Section 3.2 shows strong evidence of constant seasonality in daily temperature. This observation leads us to assume that a more appropriate exponential smoothing model that might describe Johannesburg daily temperature is Holt-Winters additive seasonal model. Holt-Winters multiplicative seasonal model is included for comparison and it is assumed that additive seasonal model will outperform multiplicative model in forecasting accuracy. We start by presenting results of estimated model parameters in Table 3.2 below.

Table 3.2: Estimates of Exponential Smoothing Model Parameters for Average Daily Temperature.

Parameter	H-W Add	H-W Mult
Alpha (α)	0.7565	0.6190
Beta (β)	0	0
Gamma (γ)	1.0000	0.6574

Table 3.2 above presents estimated model parameters. Smoothing is controlled by three parameters alpha (α), beta (β) and gamma (γ) in Holt-Winters exponential smoothing models. The estimated values of alpha, beta and gamma for the Holt-Winters additive seasonal model are 0.7565, 0, 1, respectively. Beta is 0 suggesting that the slope of the trend in our data does not change with time and is set to equal to its initial value. The estimate of the seasonality is high at 1 indicating that seasonal estimate at current time

point is just based upon very recent observations.

Holt-Winters multiplicative seasonal model parameters α, β and γ estimates are 0.6190, 0 and 0.6574 respectively. Beta is 0 suggesting that the slope of the trend in our data does not change with time and is set to equal to its initial value. Gamma is low at 0.6574 compared to Holt-Winters additive seasonal model indicating that seasonal estimate at current time point for this model is based upon both recent and far distant observations.

3.4.2 Exponential Smoothing Model Diagnostics

To assess if the fitted models are adequate to daily temperature we carry out residual analysis. This is done by checking whether the model residuals meet the major assumptions such as having constant variance, uncorrelated to each other and normally distributed. Figure 3.5 presents the residual plot, residual normal QQ plot and residual autocorrelation plot for the two exponential smoothing models. The first row in Figure 3.5 shows residual plot for the two models. This row suggests that residuals are scattered around mean zero. The second row shows the residual normal QQ plots. A 45 degree reference line is plotted with residual plot. This row suggests that it is plausible to assume that residuals for both models are normally distributed since the residual normal QQ plot for both models are close to 45 degree reference line. The autocorrelation plots in the third row shows positive auto-correlated residuals, with some of the lags showing significant correlations for both models.

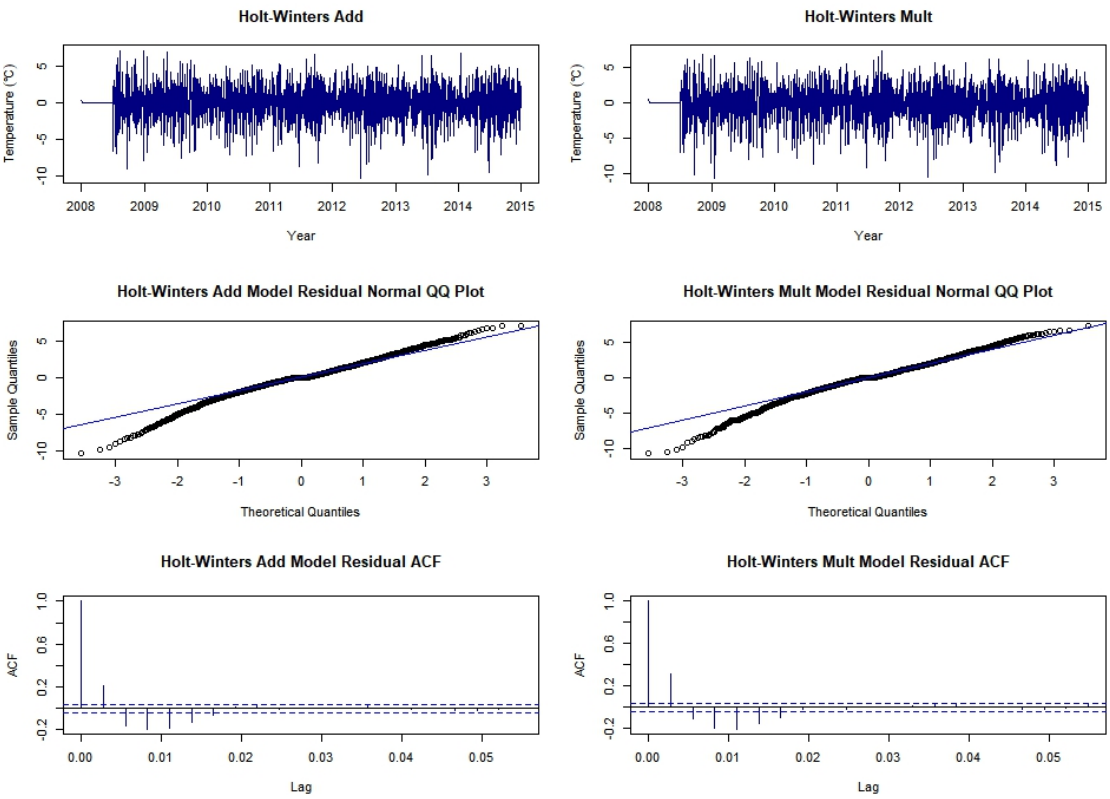

Figure 3.5: Residual plot, residual normal QQ plot and residual autocorrelation plot of Holt-Winters models for average daily temperature.

We assess results obtained from the Ljung-Box test up to 730 lags in Table 3.3 below. It is common practice to use a Ljung-Box test to assess if model residuals are correlated or not. However, not much literature exists on how to choose the number of lags for the test (Hyndman and Athanasopoulos, 2014). Through this test we want to ensure that the number of lags used is large enough to capture any meaningful and troublesome correlations. We use Hyndman and Athanasopoulos (2014) recommendations of using $h = 2m$ lags for seasonal data. Where m is the seasonality period in the time series and h is total number of lags used. For seasonal data Hyndman and Athanasopoulos (2014) claimed that it is common to have correlations at multiples of the seasonal lag remaining in the residuals, so it is advisable to include at least two seasonal lags in Ljung-Box test. This is the reason we use 730 lags for the Ljung-Box test for daily temperature models. Table 3.3 below suggests that one would reject the null hypothesis that model residuals autocorrelations up to lag 730 equal to zero since the p-values are below 0.05 significant level. Model residuals for both Holt-Winters models are correlated.

Table 3.3: Ljung-Box Test Result up to 730 lags for Daily Temperature Holt-Winters models.

Parameter	Chi-squared	df	p-value
H-W Add	2058	729	$2.2e$-16
H-W Mult	2556	729	$2.2e$-16

One major problem of both seasonal Holt's Winters methods fitted to the data is the issue of autocorrelation in residuals suggested by Table 3.3 above. According to Hyndman and Koehler (2016) the reason for this is that traditional Holt Winters models are generally designed to handle short seasonal periods such as 12 for monthly data and 4 for quarterly data where as in this study we have high frequency annual seasonal period. De Livera et al. (2011) proposed a modification to the Holt-Winters exponential smoothing methods in order to include a wide variety of seasonal patterns and solve the problem of correlated errors. The proposed method is called TBATS (Trigonometric Fourier representations, Box-Cox transformations, ARMA errors, Trend, and Seasonal component). The model includes a Box-Cox transformation which is used to stabilize the variance in the data, replace the error term with an ARMA(p, q) process and replace the seasonal component of the model with a trigonometric representation based on Fourier series that incorporates multiple seasonality. We fit the model to daily temperature series and presents model diagnostics results below.

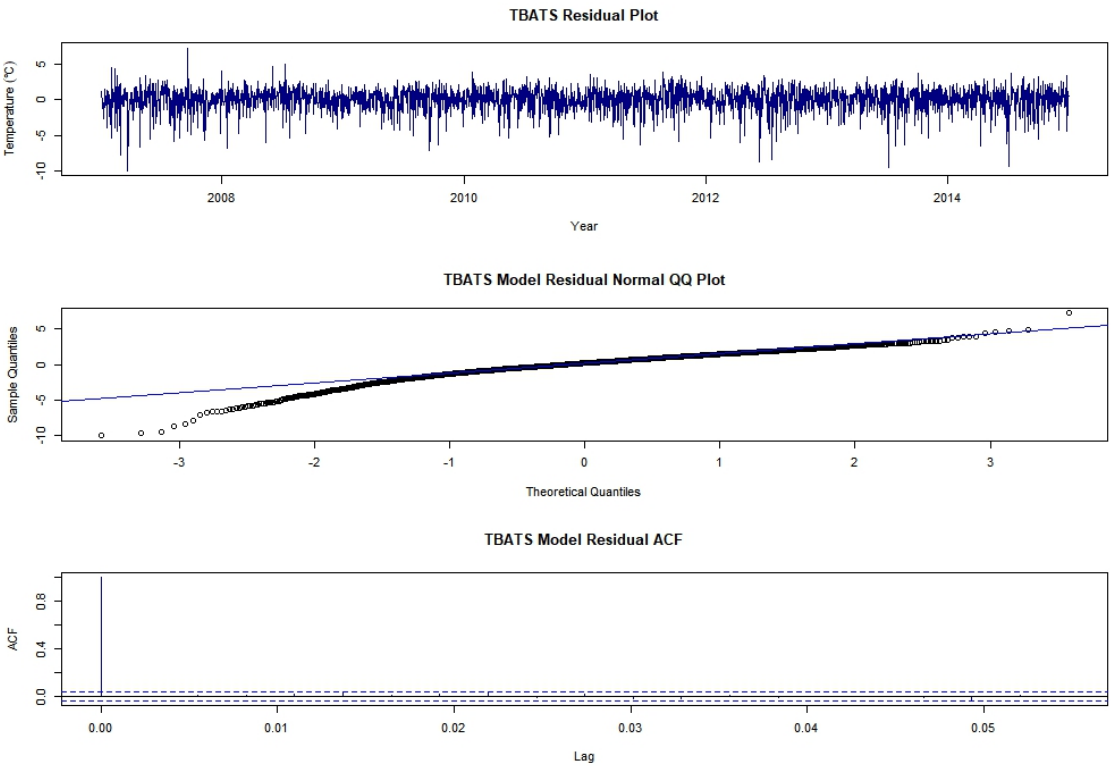

Figure 3.6: Residual plot, residual normal QQ plot and residual autocorrelation plot of TBATS models for average daily temperature.

Table 3.4: Ljung-Box Test Result up to 730 lags for Daily Temperature TBATS model.

Parameter	Chi-squared	df	p-value
TBATS	665.91	729	0.954

To assess if the TBATS model is adequate to daily temperature we carry out residual analysis in Figure 3.6 above. This Figure suggests that residuals of the model meet the major assumptions such as having constant variance, uncorrelated to each other and normally distributed. Table 3.4 above presents Ljung-Box test up to 730 lags. The test suggests that one would fail to reject the null hypothesis that model residuals autocorrelations up to lag 730 equal to zero since the p-value is greater than 0.05 significant level. We conclude that TBATS model is appropriate model to forecast Johannesburg daily temperature.

3.4.3 Exponential Smoothing Model Accuracy

Training sample in modelling is not necessarily an indicator of model performance but rather testing or validation sample. We present testing sample model accuracy performance measures in Table 3.5 below. TBATS model outperform Holt-Winters additive seasonal model and Holt-Winters multiplicative seasonal model with high accuracy in all performance measures MAE, RMSE, MPE and MAPE. This confirms our conclusion made in Section 3.4.2 above that TBATS is appropriate model to describe Johannesburg daily average temperature.

Table 3.5: Exponential Smoothing Model Accuracy Measures for Daily Temperature Test Sample.

METHOD	ME	MAE	RMSE	MPE	MAPE
H-W Add	-0.563	4.292	5.375	-10.187	27.756
H-W Mult	-0.322	4.515	5.606	-8.520	28.766
TBATS	0.843	4.031	1.705	-0.974	24.515

3.4.4 Unobserved Component Model Estimation

The model building process for unobserved component model involves understanding which components per data exploration in Section 3.2 are significant to be included in the model. The daily temperature data is highly influenced by seasonality and slightly upwards trend. Seasonality is then modelled as Fourier or trigonometric model. We test if including trend component as either local level model or local linear trend model to the Fourier seasonal model is significant or not. This is done by constructing confidence interval test for trend component in daily temperature. Table 3.6 below presents 95% confidence interval results for trend component. Table 3.6 suggests that local level model and local linear trend model are not significant to be added to the Fourier seasonal model as the confidence intervals for these components include zero in them. Trend component is then dropped from the model resulting to an unobserved component model with seasonality component only which we model as Fourier seasonal form.

Table 3.6: 95% Confidence Interval Test for Trend in Daily Temperature.

Trend Component	Lower Bound	Upper Bound
Local Level	-22.3927	34.8582
Local Linear Trend	-40.6156	31.6483

3.4.5 Unobserved Component Model Diagnostics

We assess if the seasonal Fourier model is adequate to daily temperature data by assessing if the model residual meet the major assumptions such as constant variance, uncorrelated to each other and normally distributed. Figure 3.7 presents the residual plot, residual normal QQ plot and residual autocorrelation plot for Fourier seasonal unobserved component model. The first row in Figure 3.6 suggests a roughly constant variance for the model. Second row suggests that the residuals are approximately normally distributed. Although there is a gap between the normal plot and 45 degree reference line at the beginning and end of the panel, it is plausible to accept that residuals for the model are normally distributed. The autocorrelation function in the third row shows positive auto-correlated residuals, with some of the lags showing significant correlations. Furthermore, Table 3.7 Ljung-Box test results confirm that the model residual up to 730 lags are uncorrelated since the p-value obtained is greater than 0.05 significant level.

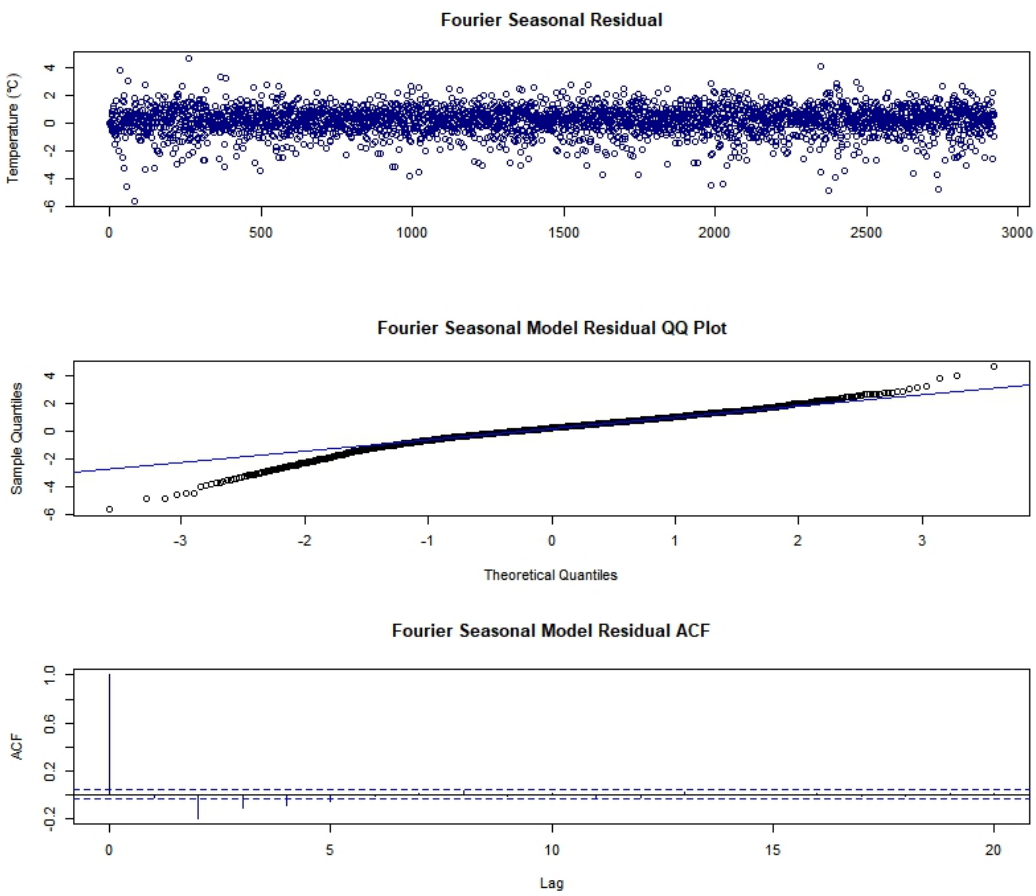

Figure 3.7: Residual plot, residual normal QQ plot and residual autocorrelation plot of Fourier Seasonal Unobserved Component Model for average daily temperature.

Table 3.7: Ljung-Box Test Result up to 730 lags for Fourier Seasonal Unobserved Component Model for Daily Temperature.

Parameter	Chi-squared	df	p-value
Fourier Seasonal	605.2	729	0.82

3.4.6 Unobserved Component Model Accuracy

Table 3.8 below presents out of sample model performance results for Fourier seasonal unobserved component model.

Table 3.8: Fourier Seasonal Model Test Sample Accuracy Measures for Daily Temperature.

METHOD	ME	MAE	RMSE	MPE	MAPE
Fourier Seasonal	-1.536	4.681	5.730	-2.723	27.744

3.4.7 Daily Temperature Modelling Conclusion

Which model between TBATS model and Fourier seasonal unobserved component model is more accurate in forecasting Johannesburg daily temperature? We end this section by comparing the forecasting measures for both models using Table 3.5 and Table 3.8. These tables suggests that TBATS model outperform Fourier seasonal unobserved component model with high accuracy in MAE, MPE and MAPE. In conclusion, TBATS model is more appropriate model to forecast Johannesburg daily temperature.

Figure 3.8 below shows TBATS model forecasting performance. Both prediction intervals in Figure 3.8 below appears to be smaller at the beginning of the forecasts and much wider as we forecasts ahead into the future.

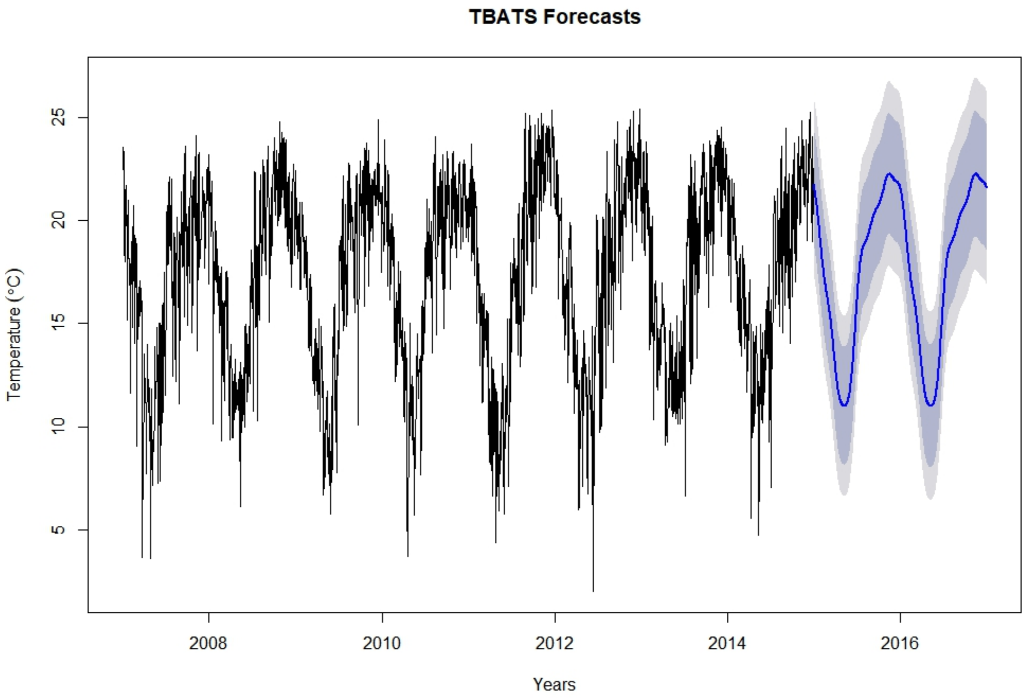

Figure 3.8: TBATS model forecasts for Daily Temperature Test Sample (Jan 2015 to Mar 2017).

3.5 Monthly Temperature Modelling Results

One of the other objective of this study is to assess the effect of temporal aggregation to the accuracy of the two structural time series methods we use. In order to study this, average daily temperature time series is aggregated to average monthly temperature time series. Figure 3.9 presents classical decomposition of Johannesburg average monthly data. From the third row of Figure 3.9 we see that average monthly temperature is strongly affected by yearly seasonal component. Second row of the figure suggests slightly upwards trend. The components of the monthly temperature in this figure compared to daily temperature of Figure 3.4 seems to be much smoother.

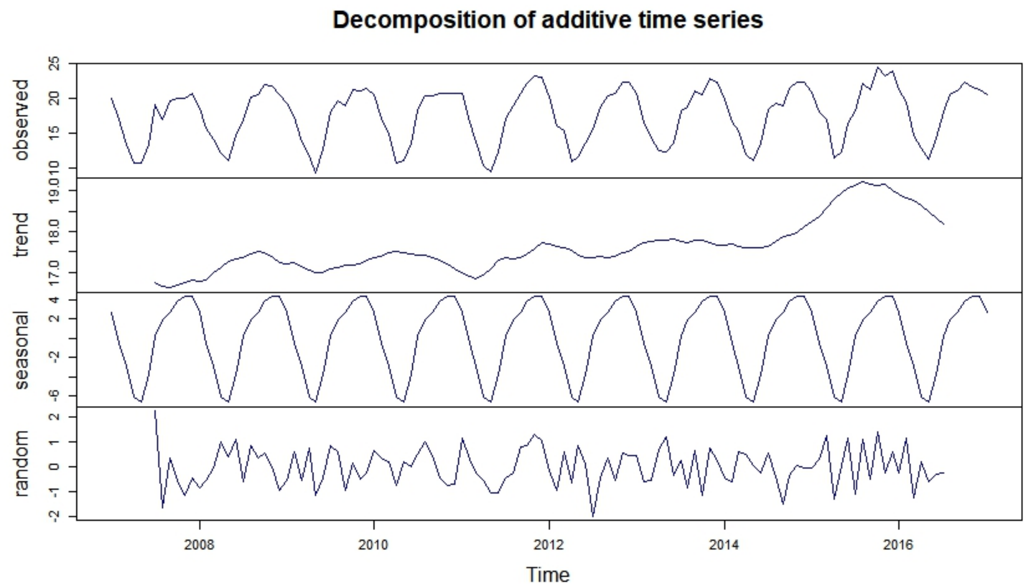

Figure 3.9: Decomposition of Johannesburg Average Monthly Temperature from March 2007 to March 2017 Using Classical Decomposition Method.

3.5.1 Exponential Smoothing Model Estimation

Holt-Winters additive seasonal model and Holt-Winters multiplicative seasonal model are fitted to the average monthly temperature data. Table 3.9 presents parameter estimates for the two models. This table suggests that both models accounted for trend effect in them. The estimate of trend smoothing parameter (β) is not zero as what we saw in parameter estimates for daily temperature in Section 3.4.1. Interestingly both models estimate trend smoothing parameter to be 0.0691. Holt-Winters additive seasonal model estimate for (α) and (γ) are 0.0616 and 0.4566. Holt-Winters multiplicative seasonal model estimate (α) and (γ) to be 0.0360 and 0.4208 respectively.

Table 3.9: Estimates of Exponential Smoothing Model Parameters for Average Monthly Temperature.

Parameter	H-W Add	H-W Mult
Alpha (α)	0.0616	0.0360
Beta (β)	0.0691	0.0691
Gamma (γ)	0.4566	0.4208

3.5.2 Exponential Smoothing Model Diagnostics

We assess if Holt-Winters models are adequate to monthly temperature data by assessing if the model residuals have constant variance, normally distributed and uncorrelated to each other using Figure 3.10 below. The first and second rows of Figure 3.10 suggests that constant variance and normality assumptions seems to hold for both models. The third row suggests that residuals are uncorrelated. We assess results from Ljung-Box test up to 24 lags in Table 3.10. This table suggests that both Holt-Winters seasonal models have uncorrelated residuals since p-values are greater than 0.05 significant level. Both Holt-Winters models have residuals that resemble white noise.

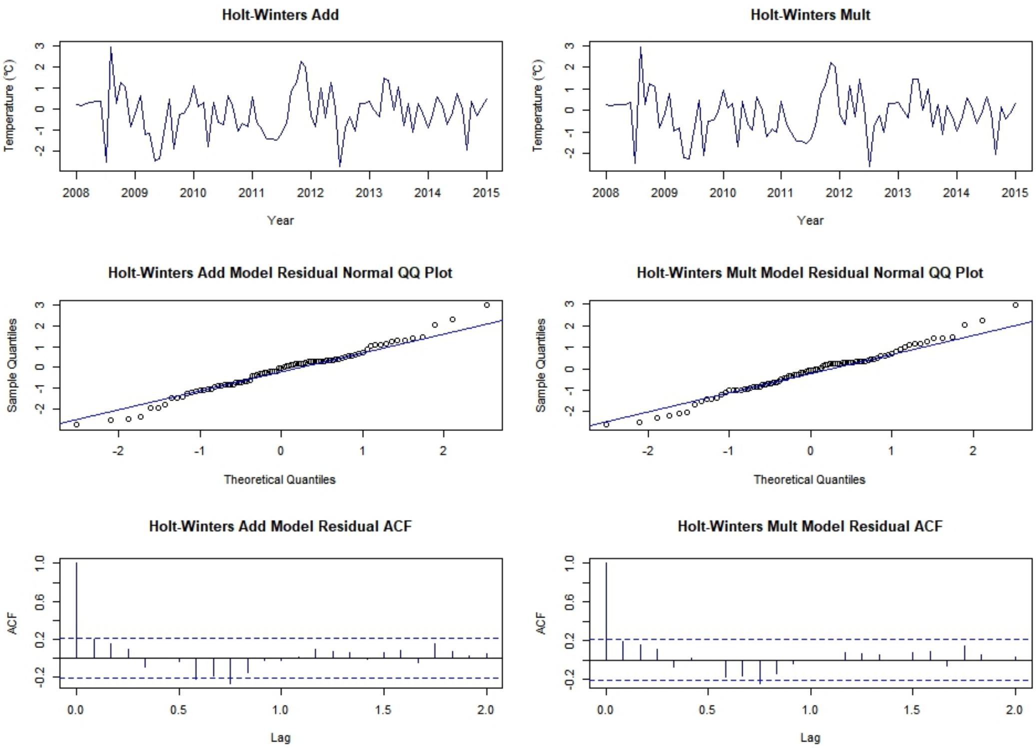

Figure 3.10: Residual plot, residual normal QQ plot and residual autocorrelation plot of Holt-Winters models for average monthly temperature.

Table 3.10: Ljung-Box Test Result up to 24 lags for Monthly Temperature Holt-Winters models.

Parameter	Chi-squared	df	p-value
H-W Add	30.919	23	0.1248
H-W Mult	26.795	23	0.2648

3.5.3 Exponential Smoothing Model Accuracy

We present out of sample model performance accuracy measures in Table 3.11 below. The table suggests that Holt-Winters additive seasonal model outperform Holt-Winters multiplicative seasonal model with high accuracy in MAE and MAPE. Holt-Winters additive seasonal model seem to be more appropriate exponential smoothing model to describe Johannesburg monthly temperature.

Table 3.11: Exponential Smoothing Model Accuracy Measures for Monthly Temperature Test Sample.

METHOD	ME	MAE	RMSE	MPE	MAPE
H-W Add	0.728	3.445	4.284	0.409	20.010
H-W Mult	0.646	3.493	4.312	-0.030	20.318

3.5.4 Unobserved Component Model Estimation

A step by step modelling process is again used to determining which components are significant to be added to our monthly unobserved component model. Figure 3.9 suggests that monthly temperature data is strongly affected by yearly seasonality. Seasonal effect is then modelled as Fourier or trigonometric seasonal form. The next question is how trend should to be modelled? Figure 3.9 also suggests an upwards trend in our monthly data. We use a 95% confidence interval test to confirm if indeed trend is significant to be added to our Fourier seasonal model. Trend can be modelled in two ways, either as a local level model or as a local linear trend model. Table 3.12 presents 95% confidence interval test for inclusion of trend as either local level or local linear model to Fourier seasonal model. Table 3.12 suggests that trend should be modelled as local linear trend component as the confidence interval results for this component does not include zero in it. This confirms that trend is significant in our monthly temperature data. A local linear trend plus Fourier seasonal unobserved component model is proposed to model Johannesburg monthly temperature data.

Table 3.12: 95% Confidence Interval Test for Trend in Monthly Temperature.

Slope Component	Lower Bound	Upper Bound
Local Level	-1.4081	0.000
Local Linear Trend	0.0039	0.0147

3.5.5 Unobserved Component Model Diagnostics

We assess if local linear trend plus Fourier seasonal unobserved component model is adequate to monthly temperature data by assessing if the model residuals have constant variance, are normally distributed and uncorrelated to each other using Figure 3.11 below. The first row of this figure suggests that model residuals have constant variance. Second row suggests that residuals are normally distributed with the residual line very close to the 45 degree reference line. The third row suggests that model residuals are correlated as there are two significant autocorrelation spikes at lag 9 and 10. We confirm this by also assessing the Ljung-Box portmanteau results up to 24 lags in Table 3.13. The table suggests that model residuals are uncorrelated as p-value is greater than 0.05 significant level.

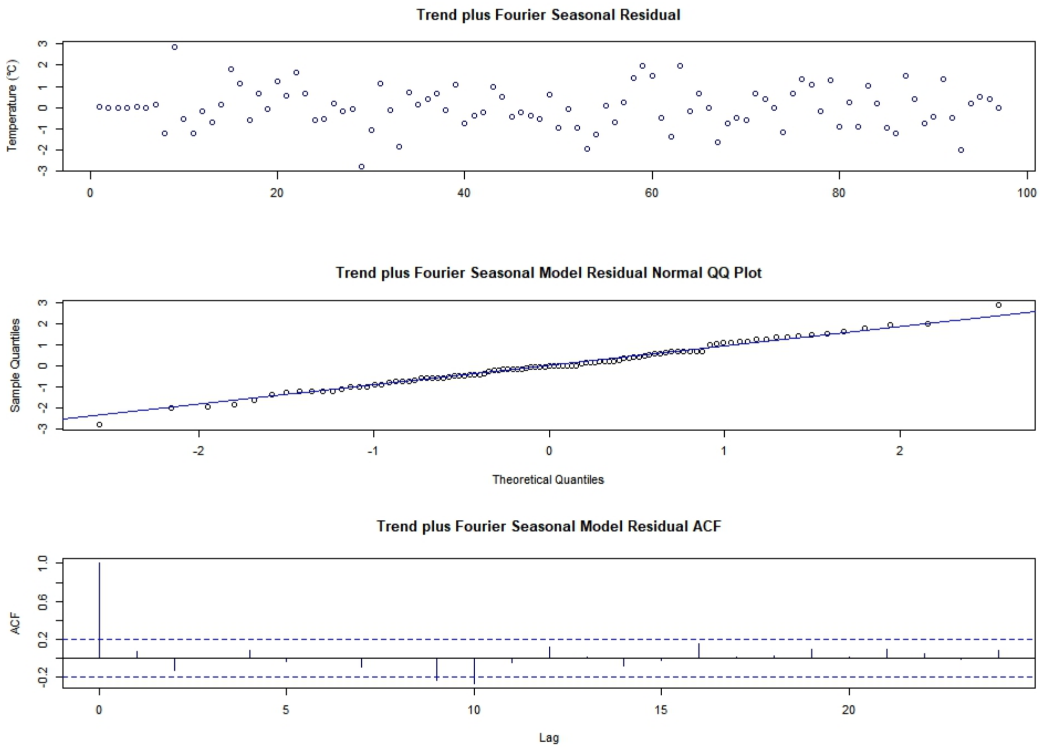

Figure 3.11: Residual plot, residual normal QQ plot and residual autocorrelation plot of local linear trend plus Fourier seasonal model for average monthly temperature.

Table 3.13: Ljung-Box Test Result up to 24 lags for Local Linear Trend plus Fourier Seasonal Model

Method Residuals	Chi-squared	df	p-value
LinearTrend+FourierSeason	26.44	23	0.2805

3.5.6 Unobserved Component Model Accuracy

Table 3.14 below presents out of sample model performance results for local linear trend plus Fourier seasonal unobserved component model.

Table 3.14: Local linear Trend plus Fourier Seasonal Model Accuracy Measures for Daily Temperature Test Sample.

METHOD	ME	MAE	MPE	MAPE
Linear Trend + Fourier Seasonal	-0.937	4.823	0.882	27.500

3.5.7 Monthly Temperature Modelling Conclusion

Holt-Winters additive seasonal model is more appropriate model to forecast Johannesburg monthly temperature. Assessing results of Table 3.11 and Table 3.14 this model outperform local linear trend plus Fourier seasonal unobserved component model with high accuracy in MAE, MPE and MAPE. Figure 3.12 below shows Holt-Winters additive seasonal model forecasting performance for monthly temperature. The forecast intervals are much small across the forecasting horizon.

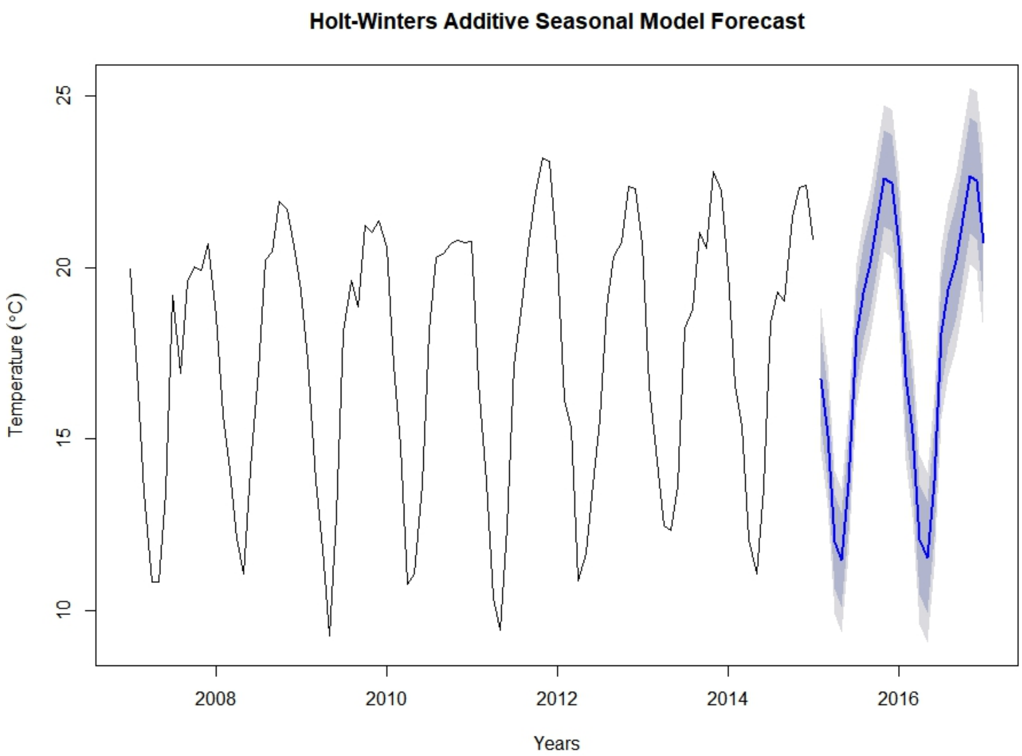

Figure 3.12: Holt-Winters additive seasonal model forecast for Monthly Temperature Test Sample (Jan 2015 to Mar 2017).

Chapter 4

Conclusion

4.1 Summary of Findings

The main objective of this study has been to compare the forecasting accuracy of the two structural time series methods, that is, exponential smoothing methods and unobserved component methods in forecasting Johannesburg daily temperature. The other objective of this study has been to understand what effect does aggregating daily temperature data to monthly temperature data has on forecasting accuracy of the two structural time series methods.

We found that both Holt-Winters seasonal models were not adequate to model and forecasting Johannesburg daily temperature as both models failed one of the major assumption on correlation in residuals. An extension of Holt Winters by De Livera et al. (2012) model know as TBATS (Trigonometric Fourier representations, Box-Cox transformations, ARMA errors, Trend, and Seasonal components) was found to be more appropriate model to describe and forecasts Johannesburg daily temperature. Holt Winters additive seasonal model was found to be more appropriate model to forecasts Johannesburg average monthly temperature. The two structural time series models used in this study were able to capture all the significant components of a smoother monthly temperature time series than high frequency daily temperature time series. We see evidence of this where the two structural time series models identify trend component as significant in monthly temperature but not significant in daily temperature. This shows that the two structural time series models are sensitive to classical components present in the data.

4.2 Further Research

The study only considered univariate structural time series models to describe Johannesburg daily temperature. One might argue that this might be a limiting factor in determining a more accurate daily temperature model as today's temperature might be strongly affected by yesterday's humidity and atmospheric pressure. This study may be extended by,

- Considering a multivariate extension of the structural time series methods presented in this study by adding other time series data that have influence on daily temperature such as humidity and atmospheric pressure,
- Assess the forecasting performance of other time series methods especially those that has gained immense popularity in the last few years such as Artificial Neural Network methods. Compare results obtained from this method with the results of this study.

References

Box G.E. and Pierce D.A., (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332): 1509−1526.

Chatfield C., (2013). The Analysis of Time Series: An Introduction. *CRC press*.

Commandeur J.F. and Koopman S.J., (2007). State Space Time Series Analysis. *Oxford University Press*.

De Livera A.M.,Hyndman R.J. and Snyder R.D., (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496): 1513−1527.

Durbin J. and Koopman S.J., (2001). Time series analysis by state space methods. *OUP Oxford,* (38):1−157.

Gardner E.S., (1985). Exponential smoothing: The state of the art. *Journal of Forecasting,* 4(1):1−28.

Gundalia M.J. and Dholakia M.B., (2012). Prediction of maximum/minimum temperatures using Holt Winters Method with Excel Spread Sheet for Junagadh Region. *International Journal of Engineering Research & Technology,* 1(6):1−8.

Harvey A.C., (1989). Forecasting, Structural Time Series and the Kalman Filter. *Cambridge University Press*, 1−100.

Harvey A.C., (2006). Forecasting with unobserved components time series models. *Handbook of Economic Forecasting,* (1):216−227.

Hastie T., Witten D., James G. and Tibshirani R., (2015). An introduction to Statistical Learning with applications in R. *Springer*.

Hyndman R.J., Koehler A.B., Snyder R.D. and Grose S., (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, (18):439−454.

Hyndman R.J. and Koehler A.B., (2006). Another look at measures of forecast accuracy. *international Journal of Forecasting,* 22(4):679−688.

Hyndman R., Koehler A.B., Ord J.K. and Snyder R.D., (2008). Forecasting with exponential smoothing: the state space approach. *Springer Science and Business Media*.

Hyndman R. and Athanasopoulos G., (2014). Forecasting: principles and practice. *OTexts*.

Jenkins M.G., (1982). Some practical aspects of forecasting in organizations. *Journal of Forecasting,* (1):3−21.

Koopman S. and Ooms M., (2011). Forecasting economic time series using unobserved components time series models. *Citeseer*.

Ljung G.M. and Box G.E., (1978). On a measure of lack of fit in time series models. *Biometrika,* 65(2):297−303.

Makridakis S. and Wheelwright S.C., (1989). Forecasting Methods for Management for the 21st Century. *Management Sciences,* (12):329−352.

Muhamad N.S. and Din A.M., (2017). Exponential smoothing techniques on daily temperature level data. *International Conference on Computing and Informatics.*

Murat M., Malinowska I., Hoffmann H. and Baranowski P., (2016). Statistical modelling of agrometeorological time series by exponential smoothing. *International Agrophysics.* 30(1):57−65.

Pegels C.C., (1969). Exponential forecasting: some new variations. *Management Sciences,* (15):311−315.

Rajarathinam A., Vetriselvi R. and Balamurugan D., (2016). Unobserved Components model for forecasting Wheat Production. *Int. J. Agricult. Stat. Sci,* 12(1):161−167.

Taylor J.W., (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting,* 19(4):715−725.

Appendix A

R Code

This section contains the R code used to generate the output results in the research report.

```
# Global settings
setwd("~/R/R-3.4.2/R Codes/Research Report")
library(readxl)
library(dplyr)
library(tidyr)
library(stargazer)
library(forecast)
library(dlm)
library(KFAS)
library(zoo)

# 1. ANALYSIS OF AVERAGE DAILY TEMPERATURE

# Data importation
tempdata = read_excel("Daily_Temperature.xlsx")
tempdata = as.data.frame(tempdata)
Temperature = tempdata %>%
separate(Date, c("Year", "Month", "Day"), sep = "-", remove = FALSE)
Temperature$Year = as.integer(Temperature$Year)
Temperature$Month = as.integer(Temperature$Month)
Temperature$Day = as.integer(Temperature$Day)
```

```r
#Missing value check and imputation
sum(is.na(Temperature$Temp))
Temperature$Temp[is.na(Temperature$Temp)]= mean(Temperature$Temp, na.rm = TRUE)

# Removing Leap days
Temperature = Temperature[!(Temperature$Month == "02" & Temperature$Day == "29")
    ,]

#summary statistics
stargazer(data.frame (Temperature[,5]) , type ="latex")
median(Temperature$Temp)

#Time series plot
plot(Temperature$Date, Temperature$Temp, type = "l", col = "blue", xlab = "Year"
    ,
ylab = expression(Temperature ~( degree *C)))

# Outlier Detection
boxplot ( Temp ~ Year , data = Temperature , xlab =" Year ", ylab =
expression ( Temperature ~( degree *C)), pch =20 , col =" navy ", mgp =
c(2, 1, 0))

#Lagged correlation plot
par(mfrow = c(3,3))
lag.plot(Temperature$Temp, lags = 9, do.lines = FALSE, col = "navy")
par(mfrow = c(1,1))

# Classical Decomposition of Series to components
TTemperature = ts( Temperature$Temp , start =c (2007 ,03 ,01) , frequency =365)
plot(decompose( TTemperature) , col =" navy ", mgp = c(2, 1, 0))

#Training and Testing Samples
Sample = between(Temperature$Year ,2007, 2015)
TrainTemperature = Temperature[Sample,]
TestTemperature = Temperature[!Sample,]
Traindata = ts( TrainTemperature ["Temp"], start = 2007, end = 2015, frequency =
     365)
Testdata = ts( TestTemperature ["Temp"], start = 2016, end = 2017, frequency =
    365 )

# Exponential Smoothing Method for Daily Temperature
```

```r
HW_add_fit = HoltWinters(Traindata, seasonal = "additive")
HW_mult_fit = HoltWinters(Traindata, seasonal = "multiplicative")

# Model Parameter Estimate
par2 = c(HW_add_fit$alpha,HW_add_fit$beta, HW_add_fit$gamma)
par3 = c(HW_mult_fit$alpha,HW_add_fit$beta,HW_mult_fit$gamma)

# Model Residuals
Res_HW_add = residuals(HW_add_fit)
Res_HW_mult = residuals(HW_mult_fit)

# Model Diagnostic
par(mfrow =c(3,2))
plot(Res_HW_add, ylab = expression ( Temperature ~( degree *C)), xlab = "Year",
    main = "Holt-Winters Add", col = "navy")
plot(Res_HW_mult, ylab = expression ( Temperature ~( degree *C)), xlab = "Year",
     main = "Holt-Winters Mult", col = "navy")
qqnorm(Res_HW_add, main = "Holt-Winters Add Model Residual Normal QQ Plot")
qqline(Res_HW_add, col="blue")
qqnorm(Res_HW_mult, main = "Holt-Winters Mult Model Residual Normal QQ Plot")
qqline(Res_HW_mult, col="blue")
acf(Res_HW_add, main = "Holt-Winters Add Model Residual ACF", lag.max = 20, col
    = "navy", plot = TRUE)
acf(Res_HW_mult, main = "Holt-Winters Mult Model Residual ACF", lag.max = 20,
    col = "navy", plot = TRUE)
par(mfrow = c(1,1))

# Ljung-Box Test
Box.test (Res_HW_add , lag =730 , fitdf =1, type = "Ljung-Box")
Box.test (Res_HW_mult , lag =730 , fitdf =1, type = "Ljung-Box")

# Validation Sample model Accuracy Measures
HW_add_V = accuracy(forecast(HW_add_fit), Testdata)[2,]
HW_mult_V = accuracy(forecast(HW_mult_fit), Testdata)[2,]
Comparison_V = data.frame(rbind(HW_add_V, HW_mult_V))
models_V = data.frame(METHOD = c("H-W Add", "H-W Mult"))
accuracy_measures_V = cbind(models_V, Comparison_V)
accuracy_measures_V = accuracy_measures_V[, -c(8:9)]
row.names(accuracy_measures_V) = NULL
accuracy_measures_V
stargazer(accuracy_measures_V, type = "latex", summary = FALSE)
```

```
#  Other Complex Exponential smoothing models (TBATS)
# model fitting
Robust_tbats = tbats(Traindata)
# Model Parameter Estimate 1
par4 = c(Robust_tbats$alpha, Robust_tbats$beta, Robust_tbats$lambda)
# Training Sample Model Accuracy Measures
Robust_tbats_T = accuracy(forecast(Robust_tbats))[1,]
# Validation Sample model Accuracy Measures
Robust_tbats_V = accuracy(forecast(Robust_tbats), Testdata)[2,]
# Model Diagnostic Testing
Res_Robust_tbats = residuals(Robust_tbats)
# Model Diagnostic
par(mfrow =c(3,1))
plot(Res_Robust_tbats, ylab = expression ( Temperature ~( degree *C)), xlab = "
    Year",main = "TBATS Residual Plot", col = "navy")
qqnorm(Res_Robust_tbats, main = "TBATS Model Residual Normal QQ Plot")
qqline(Res_Robust_tbats, col="blue")
acf(Res_Robust_tbats, main = "TBATS Model Residual ACF", lag.max = 20, col = "
    navy", plot = TRUE)
par(mfrow = c(1,1))
# Model Residual test Autocorrelation by Box-Ljung
Box.test (Res_Robust_tbats , lag =730 , fitdf =1, type = "Ljung-Box")
plot(forecast(Robust_tbats), main = "TBATS Forecasts",  xlab = "Years", ylab =
    expression ( Temperature ~( degree *C)))

# Unobserved Component Model for Daily Temperature

Buildmod_SLTr = function (par)
{
mod <- dlmModTrig(s=365,q=2)
V( mod) <- exp(par[1])
diag(W( mod))[1] <- exp(par[2])
diag(W( mod))[2:3] <- exp(par[3])
return( mod)
}
SLTr_Par_est = dlmMLE(Traindata, rep(0,3), Buildmod_SLTr)
S_Level_TrSeas= Buildmod_SLTr(SLTr_Par_est$par)

# Filtering and Smoothing
S_Level_TrSeas_Filtered = dlmFilter(Traindata, S_Level_TrSeas)
```

```
S_Level_TrSeas_Smoothed = dlmSmooth(S_Level_TrSeas_Filtered)

#Residuals
Res_S_Level_TrSeas = c(residuals(S_Level_TrSeas_Filtered, sd = F))

# 95% Trend Confidence Intervals
smooth<-dlmSmooth(Traindata, S_Level_TrSeas)
filt = dlmFilter(Traindata, S_Level_TrSeas )
U.S <- smooth$U.S
D.S <- smooth$D.S
V <- dlmSvd2var(U.S, D.S)
vars <- c()

for (i in 1:(length(S_Level_TrSeas) + 1))
{
vars[i] <- V[[i]][2, 2]
}
pl <- smooth$s[,2] + qnorm(0.05, sd = sqrt(vars))
pu <- smooth$s[,2] + qnorm(0.95, sd = sqrt(vars))
(ymin <- min(pl))
(ymax <- max(pu))

# 95% Level Confidence Intervals
U.C <- filt$U.C
D.C <- filt$D.C
U.R <- filt$U.R
D.R <- filt$D.R
U.S <- smooth$U.S
D.S <- smooth$D.S
n.obs <- length(smooth$s[-1, 1])
dlmWfp <- S_Level_TrSeas
Cov.filter <- dlmSvd2var(U.C, D.C)
Cov.smooth <- dlmSvd2var(U.S, D.S)
Cov.one.ahead <- dlmSvd2var(U.R, D.R)
n <- n.obs
T <- dlmWfp$G
Cov.diff <- NULL
Cov.diff[[n]] <- Cov.smooth[[n+1]]
for (i in (n-1):1)
{ c.diff <- Cov.filter[[i+1]] %*%
t(T) %*%
```

```
solve(Cov.one.ahead[[i+1]]) %*%
Cov.diff[[i+1]]
Cov.diff[[i]] <- c.diff
}
x2 <- smooth$s[-1, 1]
trend.diffs <- x2 - x2[n.obs]
v <- c()
for(i in 1:n.obs)
{ v[i] <- Cov.smooth[[n.obs + 1]][1, 1] + Cov.smooth[[i +
1]][1, 1] - 2 * Cov.diff[[i]][1, 1]
}
trend.diffs <- ts(trend.diffs, start = c(2007, 1), frequency =
1)
pl <- trend.diffs + qnorm(0.05, sd = sqrt(v))
pu <- trend.diffs + qnorm(0.95, sd = sqrt(v))
(ymin <- min(pl))
(ymax <- max(pu))

# Model Diagnostics
par(mfrow =c(3,1))
plot(Res_S_Level_TrSeas, ylab = expression ( Temperature ~( degree *C)), xlab =
    "", main = "Fourier Seasonal Residual", col = "navy")
qqnorm(Res_S_Level_TrSeas, main = "Fourier Seasonal Model Residual QQ Plot")
qqline(Res_S_Level_TrSeas, col="blue")
acf(Res_S_Level_TrSeas, main = "Fourier Seasonal Model Residual ACF", lag.max =
    20, col = "navy", plot = TRUE)
par(mfrow = c(1,1))

# Residual Box Test and Ljung-Box
Box.test (Res_S_Level_TrSeas , lag =730 , fitdf =1, type = "Ljung-Box")

# Out of Sample Accuracy measures
forecast_S_Level_TrSeas = dlmForecast(S_Level_TrSeas_Filtered, nAhead = 366)
forecast_S_Level_TrSeas = c(S_Level_TrSeas_Filtered$f)
Testdata = c(Testdata)
V_Res_S_Level_TrSeas = forecast_S_Level_TrSeas - Testdata
(ME = round(mean(V_Res_S_Level_TrSeas),8))
(MAE =  mean(abs(V_Res_S_Level_TrSeas)))
(RMSE = sqrt(mean(V_Res_S_Level_TrSeas^2)))
(MPE = 100*(mean((V_Res_S_Level_TrSeas) / as.numeric(Testdata))))
(MAPE = 100*(mean(abs(V_Res_S_Level_TrSeas) / as.numeric(Testdata))))
```

```
# Gnerate Future Forecast with Appropriate Model
plot(forecast(HW_add_fit), main = "Holt-Winters Additive Seasonal Model Forecast
    ",  xlab = "Years", ylab = expression ( Temperature ~( degree *C)))

# 2. ANALYSIS OF AVERAGE MONTHLY TEMPERATURE

#Calculate Monthly Temperature means
Temp1 = xts(Temperature, order.by = as.Date(Temperature$Date), "%m/%d/%Y")
monthly_endpoints = endpoints(Temp1, on ="month")
tempdata_1 = period.apply(Temp1$Temp, INDEX = monthly_endpoints, FUN = mean)
tempdata_2 = as.data.frame(tempdata_1)
Date = index(tempdata_1)
tempdata_3 = cbind(Date,tempdata_2 )
Temperature = tempdata_3 %>%
separate(Date, c("Year", "Month", "Day"), sep = "-", remove = FALSE)
Temperature$Year = as.integer(Temperature$Year)
Temperature$Month = as.integer(Temperature$Month)
Temperature$Day = as.integer(Temperature$Day)

# Decomposition of Time Series to Classical Components
TTemperature = ts( Temperature$Temp , start =2007,end = 2017, frequency =12)
plot(decompose( TTemperature) , col =" navy ",mgp = c(2, 1, 0))

#Training and Test Samples
Sample = between(Temperature$Year ,2007, 2015)
TrainTemperature = Temperature[Sample,]
TestTemperature = Temperature[!Sample,]
Traindata = ts( TrainTemperature ["Temp"], start = 2007, end = 2015, frequency =
     12)
Testdata = ts( TestTemperature ["Temp"], start = 2016, end = 2017, frequency =
    12 )

#Exponential Smoothing Method for Monthly Temperature
HW_add_fit = HoltWinters(Traindata, seasonal = "additive")
HW_mult_fit = HoltWinters(Traindata, seasonal = "multiplicative")

# Model Parameter Estimate
par2 = c(HW_add_fit$alpha,HW_add_fit$beta, HW_add_fit$gamma)
par3 = c(HW_mult_fit$alpha,HW_add_fit$beta,HW_mult_fit$gamma)
```

```
# Model Diagnostic Testing
Res_HW_add = residuals(HW_add_fit)
Res_HW_mult = residuals(HW_mult_fit)

# Model Residual constant variance test
par(mfrow =c(3,2))
plot(Res_HW_add, ylab = expression ( Temperature ~( degree *C)), xlab = "Year",
    main = "Holt-Winters Add", col = "navy")
plot(Res_HW_mult, ylab = expression ( Temperature ~( degree *C)),  xlab = "Year"
    , main = "Holt-Winters Mult", col = "navy")
qqnorm(Res_HW_add, main = "Holt-Winters Add Model Residual Normal QQ Plot")
qqline(Res_HW_add, col="blue")
qqnorm(Res_HW_mult, main = "Holt-Winters Mult Model Residual Normal QQ Plot")
qqline(Res_HW_mult, col="blue")
acf(Res_HW_add, main = "Holt-Winters Add Model Residual ACF", lag.max = 20, col
    = "navy", plot = TRUE)
acf(Res_HW_mult, main = "Holt-Winters Mult Model Residual ACF", lag.max = 20,
    col = "navy", plot = TRUE)
par(mfrow = c(1,1))

# Residual Box Test and Ljung-Box
Box.test (Res_HW_add , lag =24 , fitdf =1, type = "Ljung-Box")
Box.test (Res_HW_mult , lag =24 , fitdf =1, type = "Ljung-Box")

# Validation Sample model Accuracy Measures
HW_add_V = accuracy(forecast(HW_add_fit), Testdata)[2,]
HW_mult_V = accuracy(forecast(HW_mult_fit), Testdata)[2,]
Comparison_V = data.frame(rbind(HW_add_V, HW_mult_V))
models_V = data.frame(METHOD = c("H-W Add", "H-W Mult"))
accuracy_measures_V = cbind(models_V, Comparison_V)
accuracy_measures_V = accuracy_measures_V[, -c(8:9)]
row.names(accuracy_measures_V) = NULL
accuracy_measures_V
stargazer(accuracy_measures_V, type = "latex", summary = FALSE)

# Unobserved Component Model for Monthly Temperature

Buildmod_SLTr_1 = function (par)
{
mod_1 <- dlmModPoly(order=2) + dlmModTrig(s=12,q=2)
V(mod_1) <- exp(par[1])
```

```r
diag(W(mod_1 ))[1] <- exp(par[2])
diag(W(mod_1 ))[2:3] <- exp(par[3])
diag(W(mod_1 ))[3:4] <- exp(par[4])
return(mod_1 )
}
SLTr_Par_est_1 = dlmMLE(Traindata, rep(0,4), Buildmod_SLTr_1)
S_Level_TrSeas_1 = Buildmod_SLTr_1(SLTr_Par_est_1$par)

# Filtering AND Smoothing
S_Level_TrSeas_Filtered_1 = dlmFilter(Traindata, S_Level_TrSeas_1)
S_Level_TrSeas_Smoothed_1 = dlmSmooth(S_Level_TrSeas_Filtered_1)

#Residual
Res_S_Level_TrSeas_1= c(residuals(S_Level_TrSeas_Filtered_1, sd = F))

# Model Diagnostics
par(mfrow =c(3,1))
plot(Res_S_Level_TrSeas_1, ylab = expression ( Temperature ~( degree *C)), xlab
    = "", main = "Fourier Seasonal Residual", col = "navy")
qqnorm(Res_S_Level_TrSeas_1, main = "Fourier Seasonal Model Residual QQ Plot")
qqline(Res_S_Level_TrSeas_1, col="blue")
acf(Res_S_Level_TrSeas_1, main = "Fourier Seasonal Model Residual ACF", lag.max
    = 20, col = "navy", plot = TRUE)
par(mfrow = c(1,1))

# Residual Box Test and Ljung-Box
Box.test (Res_S_Level_TrSeas_1 , lag =24 , fitdf =1, type = "Ljung-Box")

# Validation  sample Model Accuracy Measures
forecast_S_Level_TrSeas_1 = dlmForecast(S_Level_TrSeas_Filtered_1, nAhead = 13)
forecast_S_Level_TrSeas_1 = c(forecast_S_Level_TrSeas_1$f)
Testdata = c(Testdata)
V_Res_S_Level_TrSeas_1 = forecast_S_Level_TrSeas_1 - Testdata
(ME = round(mean(V_Res_S_Level_TrSeas_1),8))
(MAE =  mean(abs(V_Res_S_Level_TrSeas_1)))
(RMSE = sqrt(mean(V_Res_S_Level_TrSeas_1)))
(MPE = 100*(mean((V_Res_S_Level_TrSeas_1) / as.numeric(Testdata))))
(MAPE = 100*(mean(abs(V_Res_S_Level_TrSeas_1) / as.numeric(Testdata))))

# Generate Future Forecast with Appropriate Model
```

```
plot(forecast(HW_add_fit), main = "Holt-Winters Additive Seasonal Model Forecast
    ",  xlab = "Years", ylab = expression ( Temperature ~( degree *C)))
```