Temperature Prediction in Madrid: A Deep Dive into Time Series Methods for Accurate Forecasting

Cindy Ly, Haniya N. Raja, Hwajin Seo, Shonak Duggal

University of Waterloo

STAT 443: Forecasting

Reza Ramezan

December 4, 2024

Table of Contents

Introduction

Forecasting temperature accurately has been an important problem for most of human history. Accurate forecasts allow everyone from farmers to corporations to plan accordingly, and in the 21st century with rapid climate change, accurately forecasting temperature is more crucial than ever before.

Our analysis focused on Spain, described as the most climatically diverse country in Europe. We focused on Madrid specifically as it is the largest city by population and is the capital of Spain. We worked with a multivariate time series starting from the beginning of 2015 to the end of 2018. The dataset contains weather features recorded on an hourly basis for several cities in Spain, such as humidity and pressure. The raw dataset contains a total of 17 columns and 178,396 rows, but we will only be extracting the rows for Madrid.

We split the data chronologically 90-10 into training and test sets respectively. Our aim is to forecast the temperature a week into the future. Our data had no missing values.

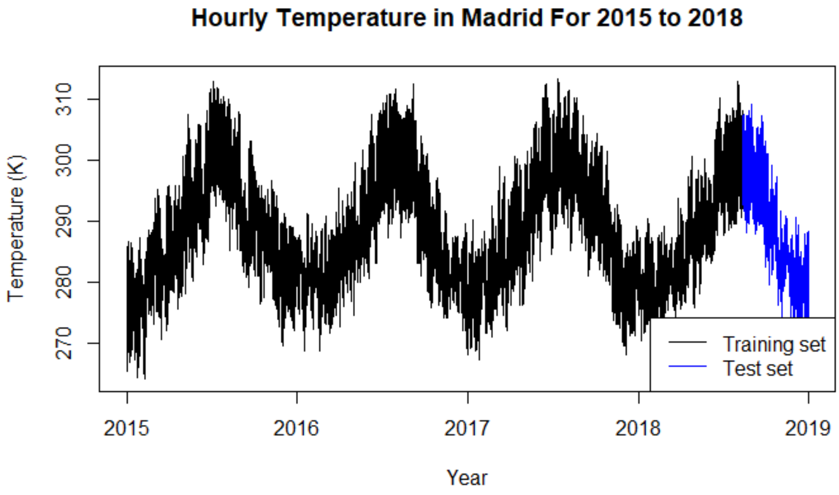


Figure 1. Plot of training and test set

Plotting temperature vs. time, we observe clear seasonality with no trend. The variance appears to be constant. We propose several potential models based on this information.

Regression

Linear model

The linear model is chosen for simplicity and interpretability of the relationship between temperature and various weather features. Below is the summary output. Notice that the model is unstable because β_0 and β_6 have high standard error, thus making prediction intervals wider. In addition, the sign of coefficients are different from what we would expect. For example, the model is implying the decrease in humidity and/or raining 1 hour will increase temperature. We will have to further investigate VIF to determine the presence of multicollinearity and use orthogonal polynomials to eliminate it. But let's first do variable selection.

(Note: for regression only, we reduced data to necessary/relevant observations by using 12AM data to see some graphs/trends more clearly and dropped categorical variables as they lead to an enormous number of indicators.)

Table 1. Regression coefficient estimates, standard errors, and p-values

	Estimate	**SE**	**p-value**
Intercept	292.138586	7.488721	< 2e-16 ***
pressure	0.020376	0.007220	0.00483 **
humidity	-0.317825	0.006770	< 2e-16 ***
wind_speed	-0.147529	0.076594	0.05428 .
wind_deg	-0.002312	0.001197	0.05363 .
rain_1h	1.583949	0.765177	0.03862 *
rain_3h	-12.698573	46.307062	0.78395
clouds_all	0.036738	0.006100	2.16e-09 ***
weather_id	-0.008009	0.001595	5.75e-07 ***

AIC and BIC

As there are many predictors to consider, AIC and BIC are used to select the best subset of predictors to achieve parsimony that penalizes everytime a new predictor is added. AIC's final model included pressure, humidity, wind_speed, wind_deg, rain_1h, clouds_all, weather_id and BIC's final model included pressure, humidity, clouds_all, weather_id.

The sign of coefficients of humidity and rain_1h has not changed. Regardless, we update our linear model such that it contains the predictors chosen from BIC in order to check standard error and p-value. The idea of using BIC here is that for large datasets, BIC favors simpler models compared to AIC.

Table 2. Regression coefficient data after parameter selection

	Estimate	**SE**	**p-value**
Intercept	291.848310	7.404310	< 2e-16 ***
pressure	0.020957	0.007204	0.00368 **
humidity	-0.314981	0.006436	< 2e-16 ***
clouds_all	0.034665	0.005851	3.88e-09 ***
weather_id	-0.009492	0.001446	7.09e-11 ***

Majority of the standard errors are very close to 0, and all the predictors are significant. If we compare the VIF of the original linear model and the updated linear model, improvements have been made in terms of multicollinearity and standard error values. Below is a list of VIF values.

Table 3. VIF values, before and after parameter selection

	pressure	humidity	wind_speed	wind_deg	rain_1h	rain_3h	clouds_all	weather_id
Original	1.044183	1.519783	1.197762	1.062164	1.371627	1.006000	1.722701	1.583784
New	1.032830	1.364739					1.574690	1.292313

LASSO

LASSO is a modification of linear regression that takes into account the number of coefficients and penalizes via tuning parameters. It performs model selection which is useful when there are some factors that are irrelevant when explaining temperature. The graph on the left in Figure 2 shows that β_2,..., β_8 shrink to 0, signalling that they are less important predictors than β_1 and choosing the first predictor as the final model.

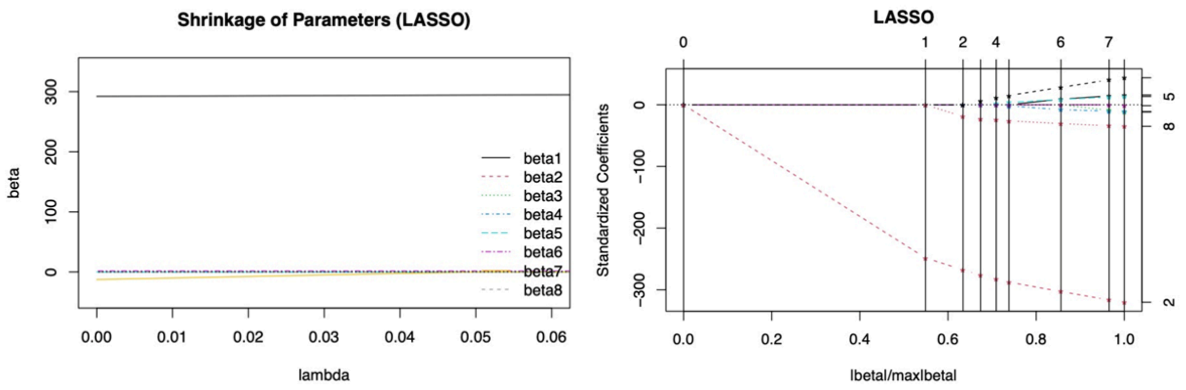


Figure 2. LASSO parameter shrinkage

The graph on the right in Figure 2 and Table 4 exhibits the order in which predictors are introduced to the model. It appears like humidity is the most important predictor, and rain_3h being the least significant.

Table 4. Sequence of parameter selection

	humidity	weather_id	clouds_all	rain_1h	wind_deg	pressure	wind_speed	rain_3h
Var	2	8	7	5	4	1	3	6
Step	1	2	3	4	5	6	7	8

Smoothing / Holt-Winters

Motivation

Smoothing methods estimate the patterns in the data while reducing the noise. It accounts for features of non-stationarity like heteroscedasticity and trend. In particular, the Holt-Winters algorithm is ideal for data that includes trend and seasonality. The raw data has a clear seasonal component, so Holt-Winters should perform well on this data. In addition, the seasonality is roughly constant over time, so we propose the additive Holt-Winters model.

The Model

The additive Holt-Winters method defines a level, trend, and seasonal index in an iterative fashion to forecast h periods ahead. Each component also has a smoothing parameter for precise control of smoothness.

Results

Running the Holt-Winters algorithm on the training set and validating our model using the test set gives us a value of 151.3149 for the APSE. The prediction is fairly accurate compared to the actual data, as shown in the plot below. The prediction becomes less accurate as time goes on, but the test set accounts for 151 days of data, which is much more data than typically required for predicting the weather. The most important thing is the forecast accuracy 1-2 weeks ahead. However, the residual diagnostics (see appendix) show a clear violation of normality, along with seasonality in the ACF so the predictions and prediction intervals are not valid. The data was also fitted to simple exponential smoothing and double exponential smoothing models which both returned APSEs of 406. The multiplicative Holt-Winters model could not be fitted, due to an optimization error. Presumably, because the homoscedastic nature of the data is not suited for this method. Choosing the additive Holt-Winters model is the logical choice.

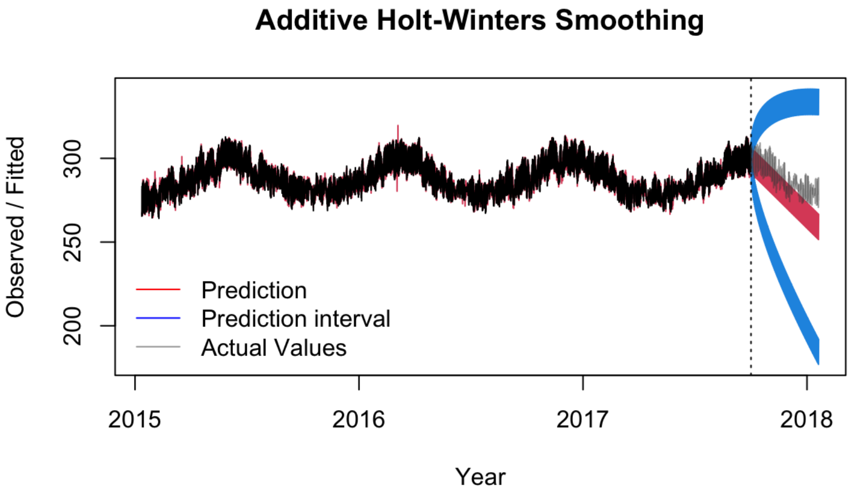


Figure 3. Holt-Winters Forecast

Box-Jenkins

Box & Jenkins are popular models for forecasting time series. They take three forms, namely

1. $ARMA(p, q)$, for stationary time series
2. $ARIMA(p, d, q)$, for non-stationary time series, and
3. $SARIMA(p, d, q) \times (P, D, Q)_s$, which incorporates seasonality.

The models depend on two key functions, the ACF and PACF. Therefore, we start by plotting the ACF and PACF of the training data, and examine the patterns in these graphs.

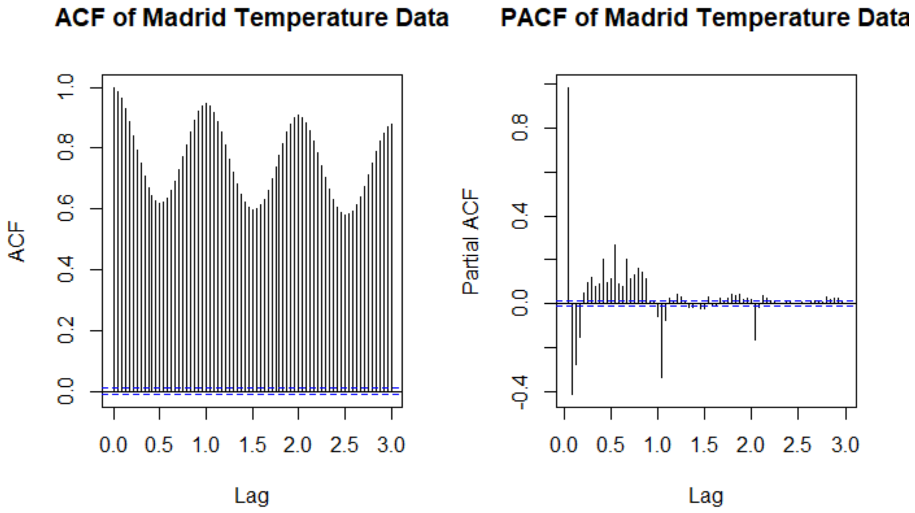
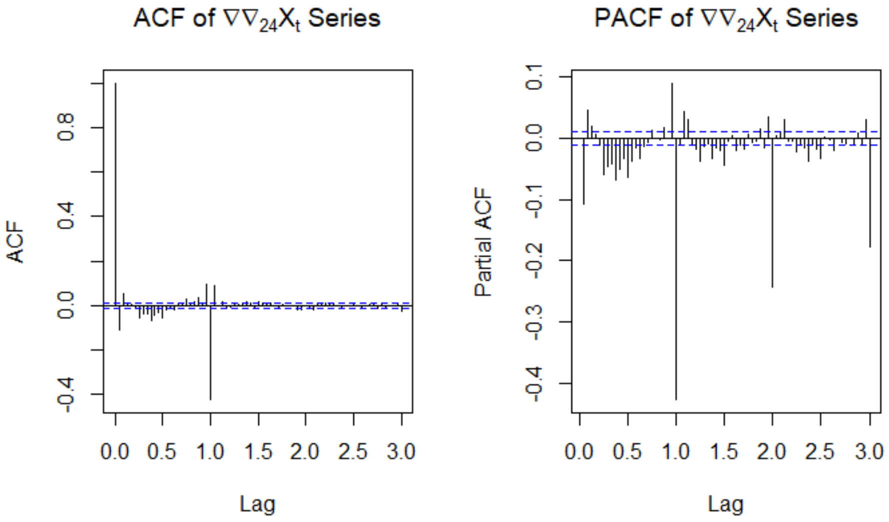


Figure 4. ACF and PACF of training data

The ACF shows clear seasonality, so we propose a SARIMA model with parameters $S = 24$. The PACF appears to cut off after a certain point, which is characteristic of an AR process, so we set the parameter $q = 0$.

Our next step is to difference the data to achieve stationarity. After regular differencing once ($d = 1$) and differencing at lag 24 once ($D = 1$), we find the ACF and the PACF of the differenced data is as shown.



ACF of $\nabla\nabla_{24}X_t$ Series

PACF of $\nabla\nabla_{24}X_t$ Series

Figure 5. ACF and PACF of differenced data

Studying the peaks in these two graphs gives us the possible values for the remaining parameters. After fitting all possible models, we found the best model to be $SARIMA(0,1,0) \times (3,1,1)_{24}$, which had an APSE of 40.6.

When performing residual diagnostics, we found that none of the SARIMA models had uncorrelated, Normally distributed residuals (See Appendix 3 for the residual diagnostics of $SARIMA(0,1,0) \times (3,1,1)_{24}$. All models had similar residuals). We attempted to correct this by fitting models to transformed data (log, squared, square root, exponential), but this did not have any effect on the residuals.

XGBoost

Motivation

We observe that there are 36,367 total observations for 9 x-variables, and most of the variables had similar or the same values for the majority of the data, except a handful of observations. Decision tree-based models such as XGBoost (Extreme Gradient Boosting) tend to perform better on data of this nature. Based on our observations, we try using XGBoost on the data.

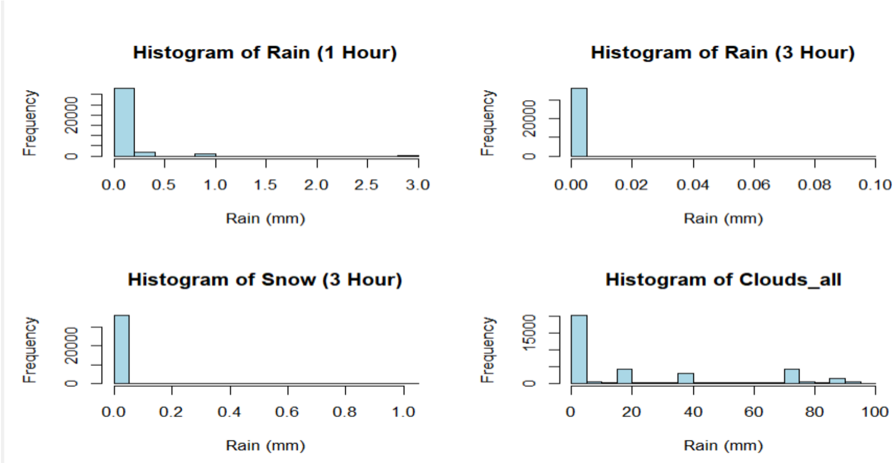


Figure 6. Precipitation histograms

Pre-processing/Feature Engineering

The model uses the entire dataset lagged by 1 week as well as the weather features lagged by 1 week as engineered features (since we're trying to predict the temperature a week ahead, given the conditions right now) as well as the target series (Temperature) lagged by 2 weeks as well.

Model Explanation

XGBoost works by combining many simple decision tree models through a process called gradient boosting. We start with an initial guess, usually, the mean of the target variable, calculate the residuals and then fit another tree to them, one at a time, to correct the errors from the previous trees, improving accuracy step by step.

The model runs for a set number of rounds, allowing it to refine its predictions gradually. Each tree has a limited depth, which allows it to capture intricate patterns without becoming too complex, helping combat overfitting. The learning rate controls how much each new tree affects the overall model. It is a form of penalisation that avoids overfitting and helps the model generalize new data better.

XGBoost focuses on minimizing the difference between actual and predicted values by optimizing the mean squared error. Parameters like tree depth and learning rate regulate the bias-variance trade-off both at the individual Tree level as well as the whole ensemble level.

Model Tuning

The tuning process for this model was a little out of the ordinary as with a time series you can’t really use Cross Validation as there'd be a ‘gap’ in the middle (assuming the test set is not the first or the last set).

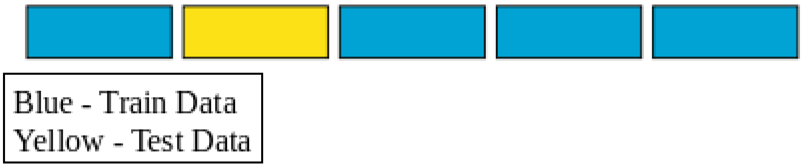


Figure 7. Cross-validation diagram

So I employed a grid search and for every combination of parameters, I trained the model on 90% of the data, predicted on the test set of 10%, calculated the MSE and stored it.

Results

The optimal model had test set MSE of 10.79242 with:

- max_depth = 3, (how many splits in each Tree)
- eta = 0.2, (Learning Rate ie penalisation factor λ)
- nrounds = 25, (number of trees)

The most important features were the lagged engineered feature and the Humidity, which was to be expected however, to my surprise Wind did not play a larger part (in hindsight for Spain, the wind-chill index might not affect the temperature that much).

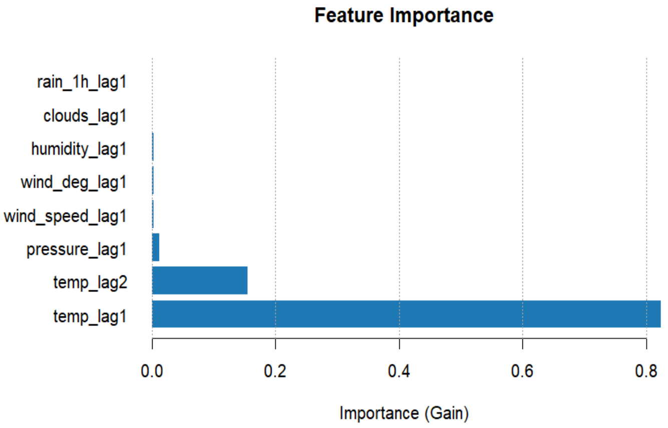


Figure 8. Relative importance of each feature

The XGBoost model, trained on 90% of data and forecasting the rest 10% looks like this:

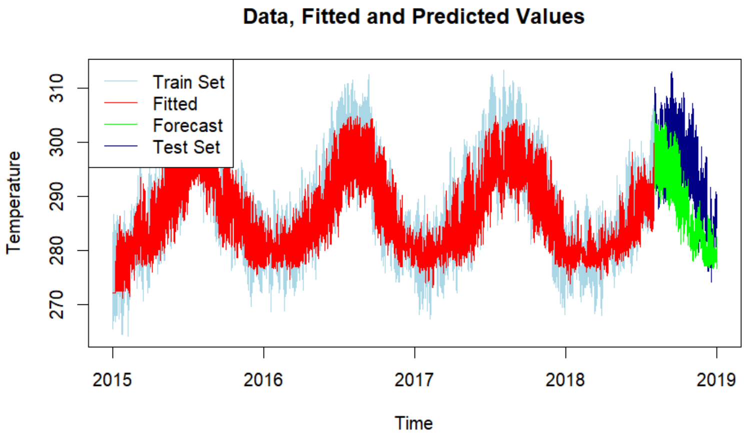


Figure 9. XGBoost fitted model

The overall fit visually looks pretty good and the test set APSE is 21.93661

Conclusion

Comparing all APSE values, we conclude that the overall best model is XGBoost.

Table 5. APSE of best models in each category

Best Model	APSE
Linear Model (Lasso)	26.19445
Holt-Winters	151.3149
Box-Jenkins	40.61618
XGBoost	21.93661

Below is a one-week forecast of the data using the optimal XGBoost model.

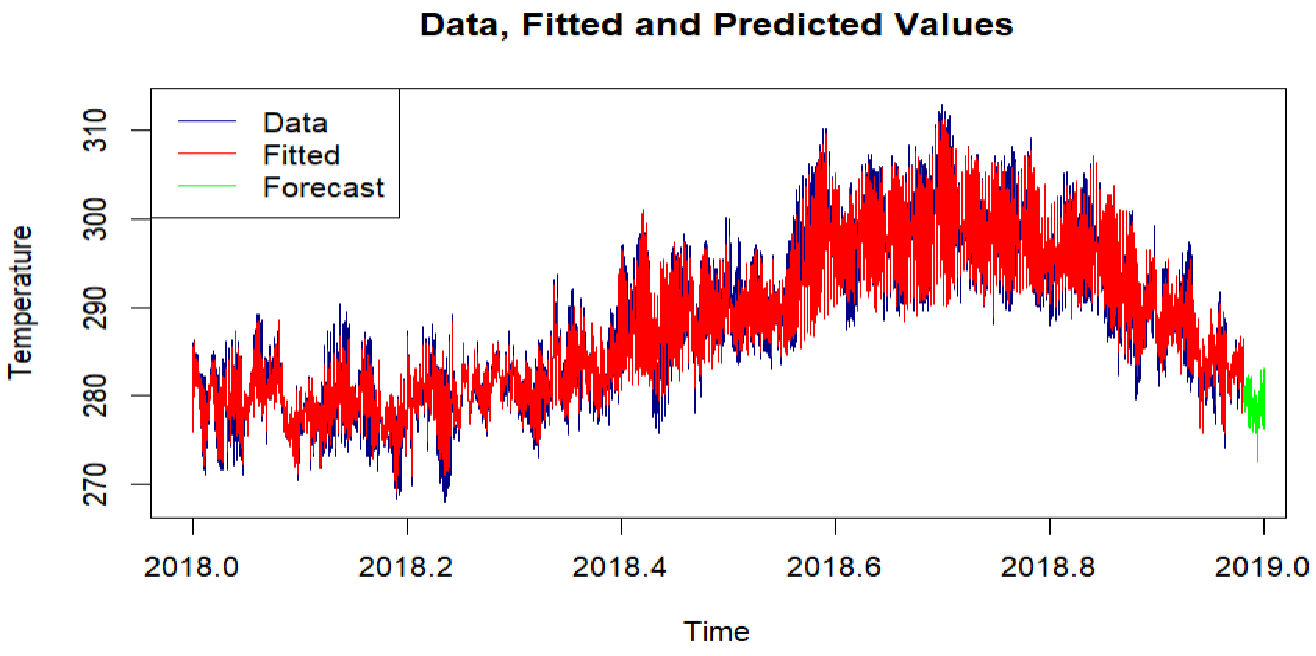


Figure 10. Forecast with the best model

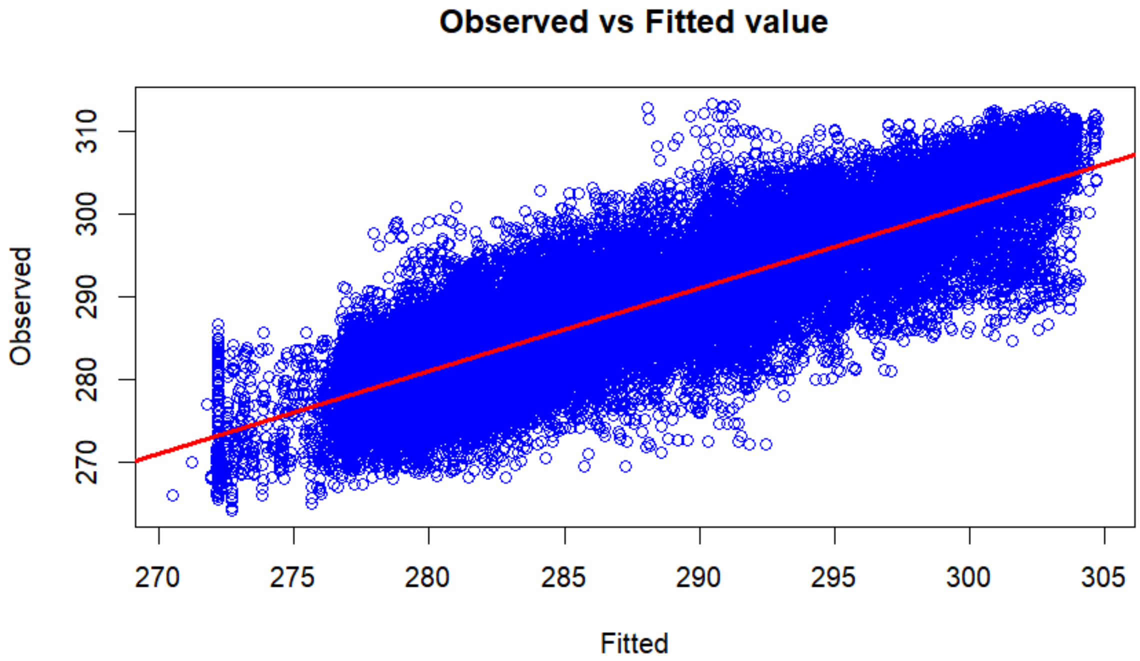


Figure 11. Observed-Fitted graph of XGB Residuals

Aside from the clustering around 272.5 all of the observations fall along the line of inference, implying an overall good fit.

Our analysis shows that it is possible to accurately predict the weather one week ahead. Even though we only investigated a singular city, the final takeaways of this analysis can be applied to any location. Having the means to do this may unlock numerous other possibilities such as modelling extreme weather phenomena as well as improving our understanding of much larger issues, such as climate change.

One thing to consider is how much data we actually need to do short-term predictions just as well. We used 4 years of hourly data. Despite the fact that a greater number of data narrows the prediction interval, how much of that is actually relevant to the point predictions? If we could use less data, it would be less expensive to run our models computationally as well as increase their scope of usefulness. Thinking about the answer to this question can greatly enhance the efficiency of the whole process. Although we tried some rudimentary analysis to explore this question we ultimately found that it was outside the scope of this report.

Overall it was an interesting learning experience as we were able to apply the concepts and techniques learned in STAT 443 and it challenged some of our preconceived notions about modeling, statistics and the weather.

Appendices

Appendix 1: Linear Model Residual Diagnostics

Linear model performance assessed based on residual diagnostics to check the assumptions of linear regression.

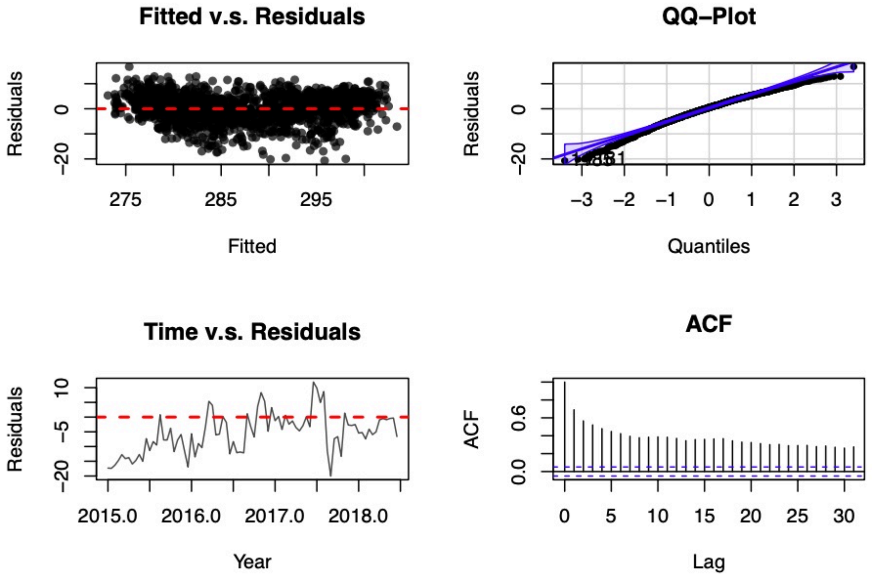


Figure 12. Residual diagnostic plots for original regression model

Updated linear model residual diagnostics does not show much improvements, thus no evidence of significant improvements. In both, the trend is not visible in fitted v.s. residuals and time v.s. residual plots. However, more points are observed below the line of 0 and thus an average of negative, indicating systematic bias or underprediction in the model. The QQ-Plot shows a somewhat concave down trend, suggesting the data has heavier tails than the normal distribution and potential extreme values. Finally, the ACF displays the correlated residuals meaning that the model has not captured the structure of the data well.

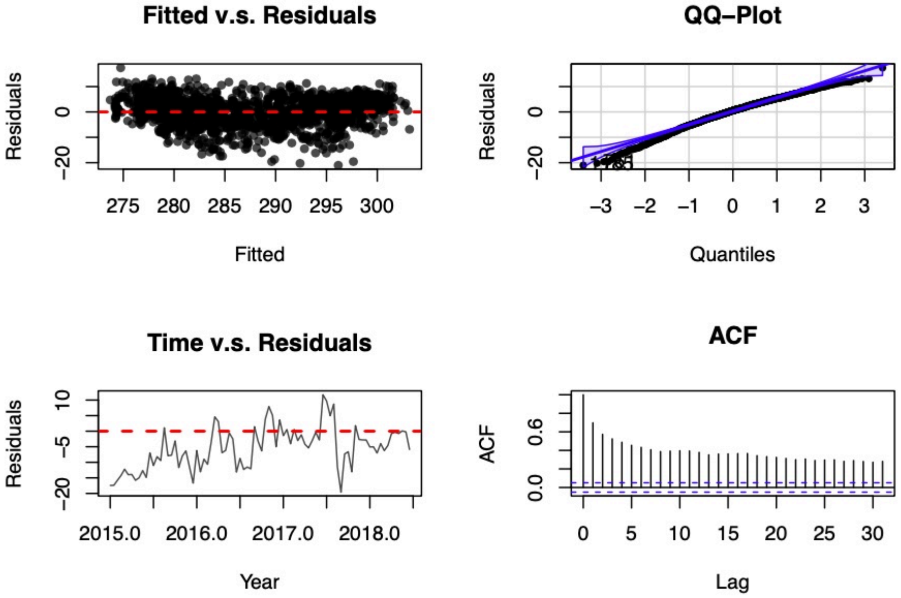


Figure 13. Residual diagnostics plots for updated regression model using BIC

Appendix 2: Residual Diagnostics for Additive Holt-Winters

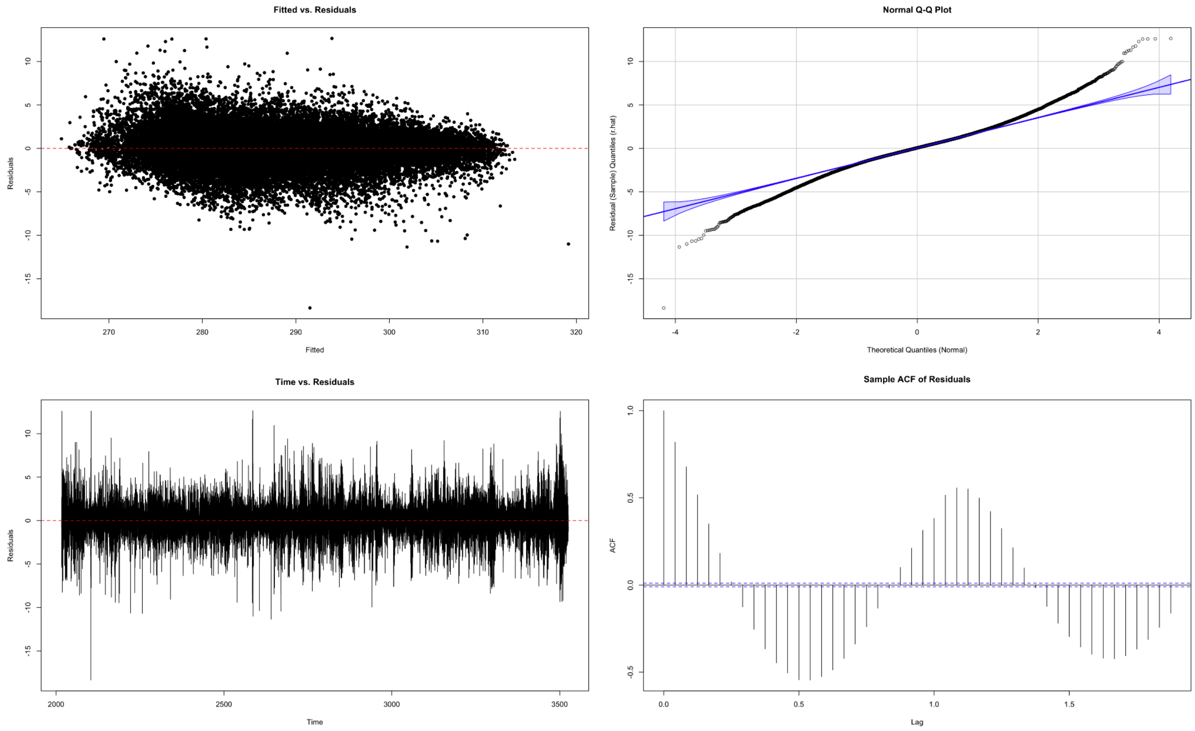


Figure 14. Residual diagnostics plots for Holt-Winters

Appendix 3: Residual diagnostics for $SARIMA(0,1,0) \times (3,1,1)_{24}$

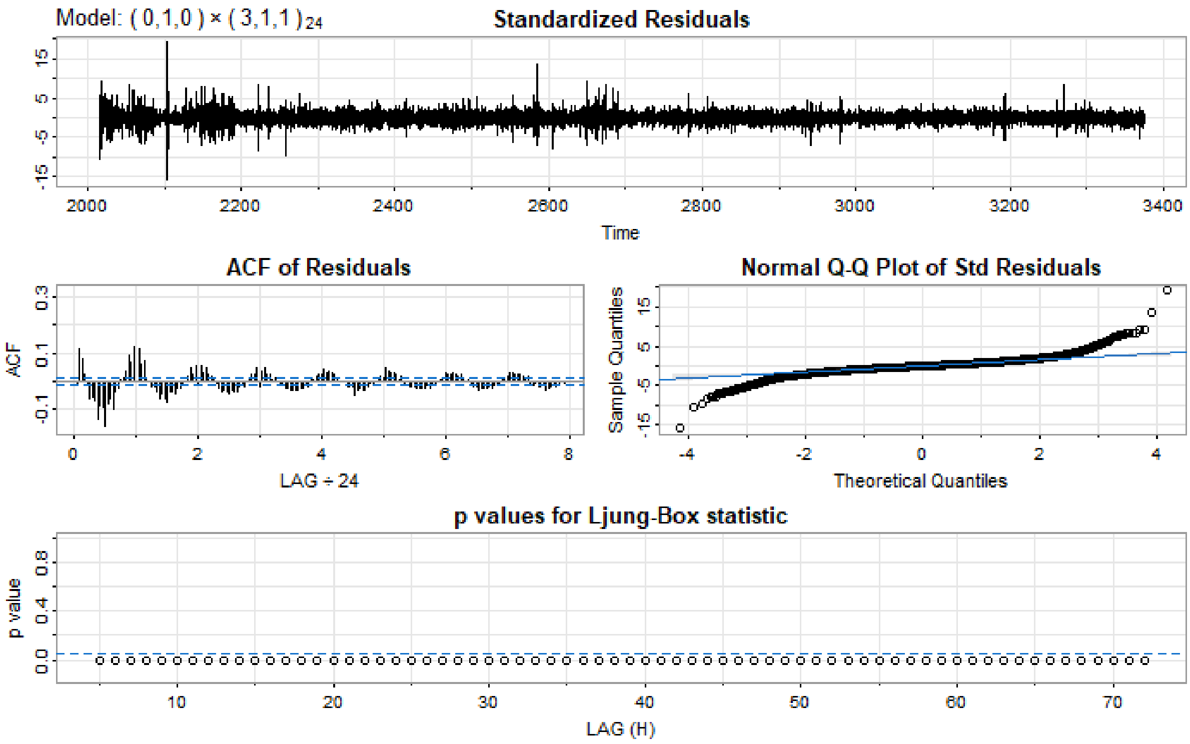


Figure 15. Residual diagnostics plot for Box-Jenkins model