STAT443 Final Project

Weather Forecasting in Waterloo Region

Winter 2024

Group 10

Reported To

Professor Reza Ramezan

Department of Statistics and Actuarial Science

University of Waterloo

Written By

Ruiyun Chao (Code generation, Box Jenkins method)

Kendra Fan (Code generation, Regression, Smoothing method)

Silin Wang (Code generation, Box Jenkins method)

Danqing Zhang (Code generation, Summary, Introduction, Methodology and Discussion)

Xinyin Zhang (Code generation, Summary, Introduction, Methodology and Discussion)

08/April/2024

Table of Contents

1. Introduction

This report provides a weather forecast for the first week of 2024 for the Waterloo region, using 2023 temperature data from the University of Waterloo Weather Station. We will use different forecast models: unregularized regression, regularized regression, smoothing, and Box Jenkins, to make predictions and select the best model by comparing their average prediction squared error (APSE).

We have chosen this topic for many reasons. Firstly, relatively accurate temperature forecasts allow people to make better decisions. Based on our experience studying here for four years, Waterloo's temperatures could change quickly, increasing the risk of illnesses. Furthermore, we have found trustworthy data from the University of Waterloo Weather Station. Additionally, we will gain valuable learning experience through working on the project. Incorporating forecasting models, such as regression, smoothing, and Box-Jenkins, into our project could help us apply the concepts we learned in class in a real-world context. After improving the models we built, we could generate more weather forecasts that could contribute to the future planning of the city of Waterloo.

2. Methodology

2.1 Preprocessing

The original data contains 35040 observations, which measure the temperature every 15 minutes. Since our project is about analyzing data and forecasting the temperature using the data for a year (2023), we reduced the data set to 1460 observations by taking the mean value of the original data with a 6-hour block using Python. The present data takes weather information in Waterloo collected from January 1st 00:00 to December 31st 18:00 with a block of 6 hours,

which provides significant meteorological data including temperature, dew point, precipitation, etc. To analyze temperature data during 2023, we split the temperature column of the data during the implementation. However, the observations contain a few missing data as the weather information is not recorded, to solve the issue, we apply na.omit() statement in R Studio to omit the NA data.

2.2 Brief Data Analysis

Firstly, we divide the data into two groups: the training set, from January 1st 00:00 to November 30th 18:00, and the testing set from December 1st 00:00 to December 31st 18:00. Figure 2.2.1 reveals an apparent decreasing trend and seasonality with high frequency.

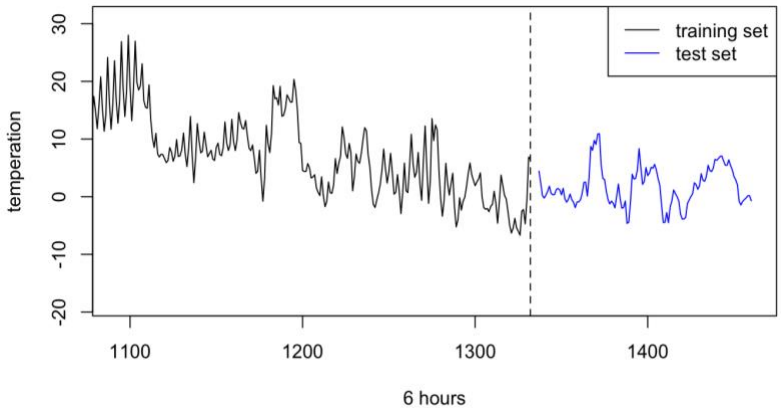


Figure 2.2.1

Figure 2.2.2 shows the training data's ACF, which is a slow linear decay with seasonal behaviour. To verify whether the seasonal component exists, we apply regular differencing to the data.

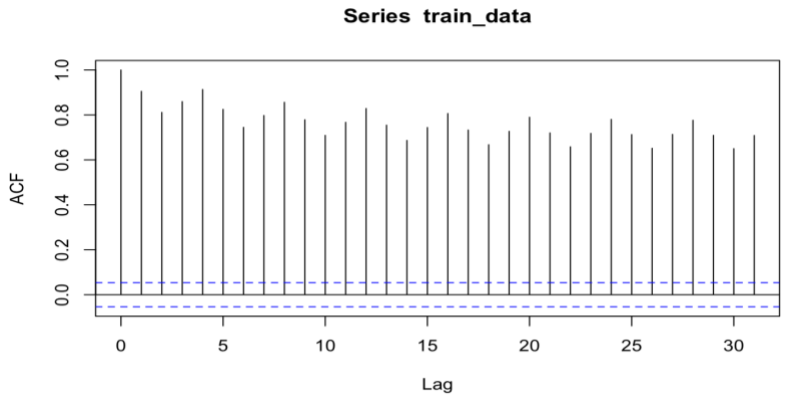


Figure 2.2.2

After moving the trend by differencing, the ACF plot shows a daily jump (4 time lags). Looking at Figure 2.2.3, the period is 4.

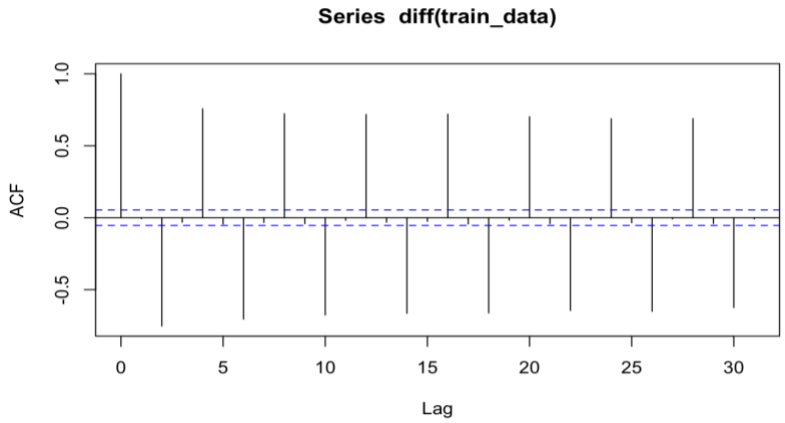


Figure 2.2.3

3. Statistical Analysis

3.1 Unregularized Regression

We calculate the APSE with degrees from 1 to 10. According to Table 3.1.1, Degree 2 has the smallest APSE. Polynomial regression with the seasonal component has a higher APSE. However, based on the observation in 2.2, and the context of weather forecasting, we still choose to include the seasonal component.

Polynomial Degree	APSE	APSE with Seasonal
1	269.01840	271.64148
2	21.99922	24.69160
3	309.31214	312.98696
4	58.20487	61.41930
5	100.64489	105.77534
6	199.21466	205.18666
7	127.44515	139.50160
8	51.59736	59.64027
9	88.88465	117.59807
10	457.59477	519.56802

Table 3.1.1

Graphical diagnostics, Figure 3.1.2, are required for model checking. The two residual plots have some patterns. Residuals show a correlation and slow decay patterns exist based on the ACF plot. Therefore, the selected model is not good.

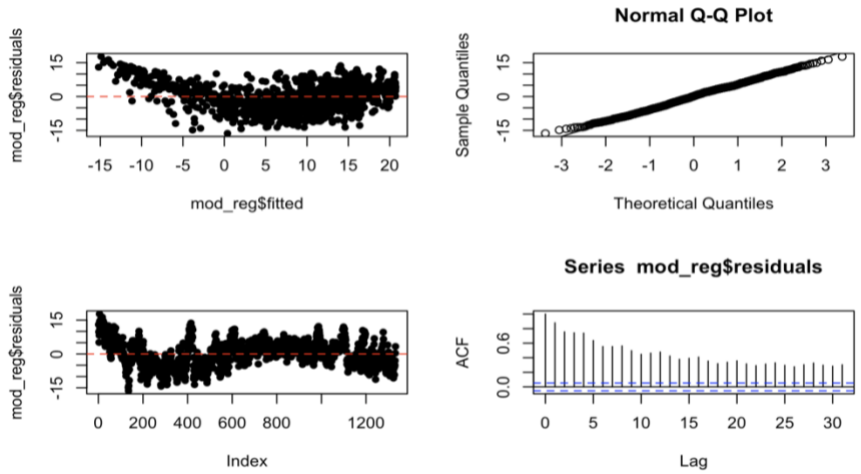


Figure 3.1.2

Figure 3.1.3 shows the predicted temperature for the first week of 2024.

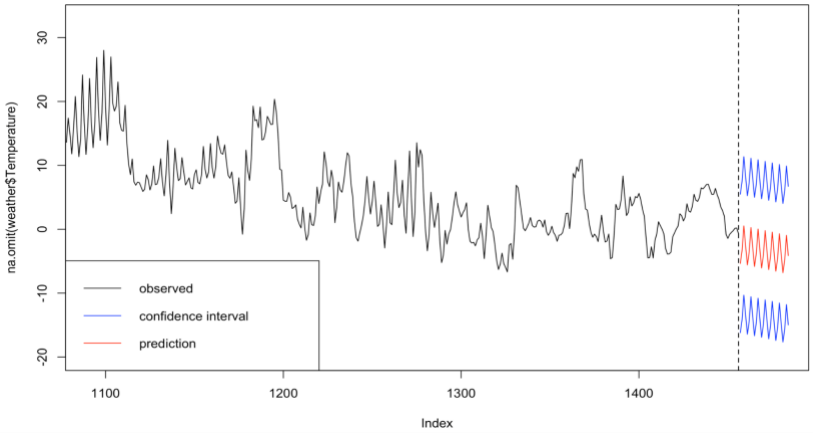


Figure 3.1.3

3.2 Regularized Regression

Regularization methods can penalize the complexity of the model by applying an additional penalty term, which can handle multicollinearity and avoid overfitting. We use elastic net regression with alpha = 0 (ridge), 0.5 and 1(LASSO). By comparing their predicted performance, the optimal model is LASSO regression, which has the smallest APSE.

Method	**APSE**
Ridge Regression (alpha=0)	136.9263
Elastic Net Regression (alpha=0.5)	133.3762
Lasso Regression (alpha=1)	130.0699

Table 3.2.1

3.3 Smoothing

Double exponential smoothing has the smallest APSE. Double exponential smoothing is the best model among all smoothing methods.

Method	**APSE**
Simple Exponential Smoothing	13.13694
Double Exponential Smoothing	12.7599
Additive Holt-Winters Smoothing	47.41026
Multiplicative Holt-Winters Smoothing	2201524

Table 3.3.1.

Figure 3.3.2 shows the predictions by double exponential smoothing model:

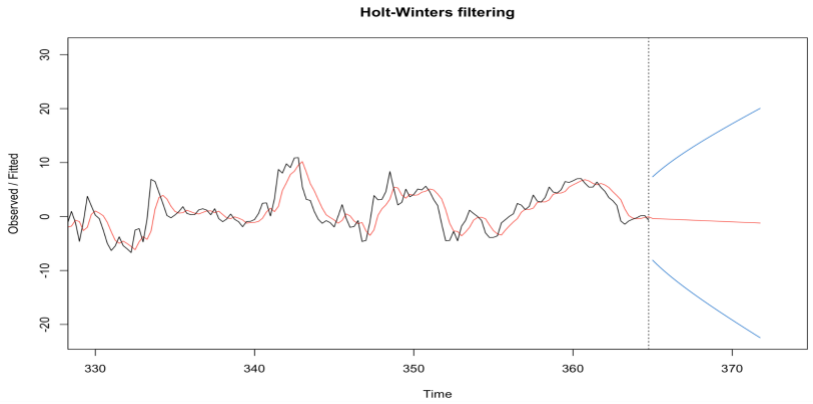


Figure 3.3.2

3.4 Box-Jenkins

After the smoothing method, we would like to discuss the data using Box-Jenkins models. Box-Jenkins is one of the most popular methods for modeling time series. It applies ARMA and ARIMA models to forecast the time series, relying on ACF and Partial ACF.

Firstly, the data plot indicates an obvious quadratic trend. In Figure 3.4.1, the original ACF plot shows a seasonal pattern and an exponential decay with a period of 4 from ACF. Moreover, all the lags stay outside of the confidence interval. Therefore, the process is not stationary and cannot be white noise.

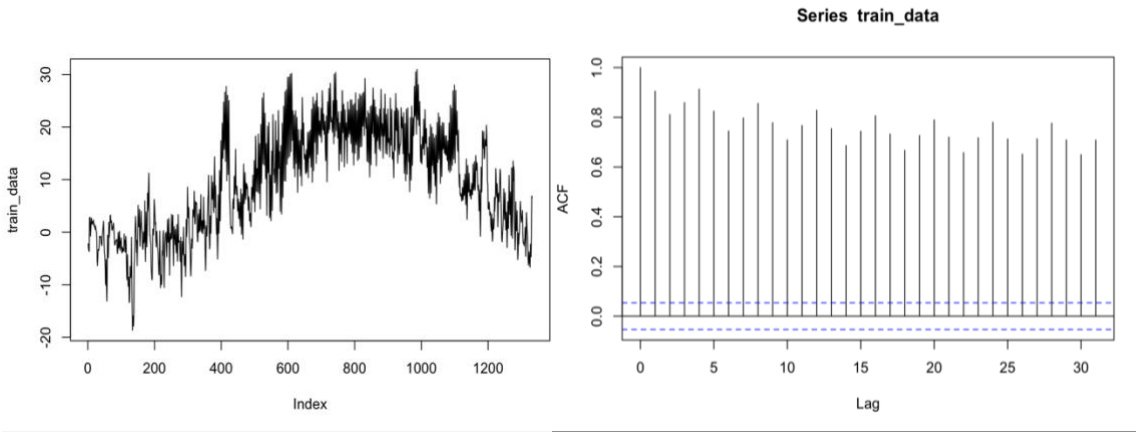


Figure 3.4.1

Upon examination of the time series plot, it achieves stationarity, displaying neither clear trends nor seasonal patterns. Subsequent application of seasonal differencing with a lag of 4 reinforces this observation, as evidenced by the sample autocorrelation function (ACF). The ACF does not present any seasonality, where there is no periodic pattern or a slow decay on seasonal lags or across lags, further supporting the conclusion of stationarity within the data.

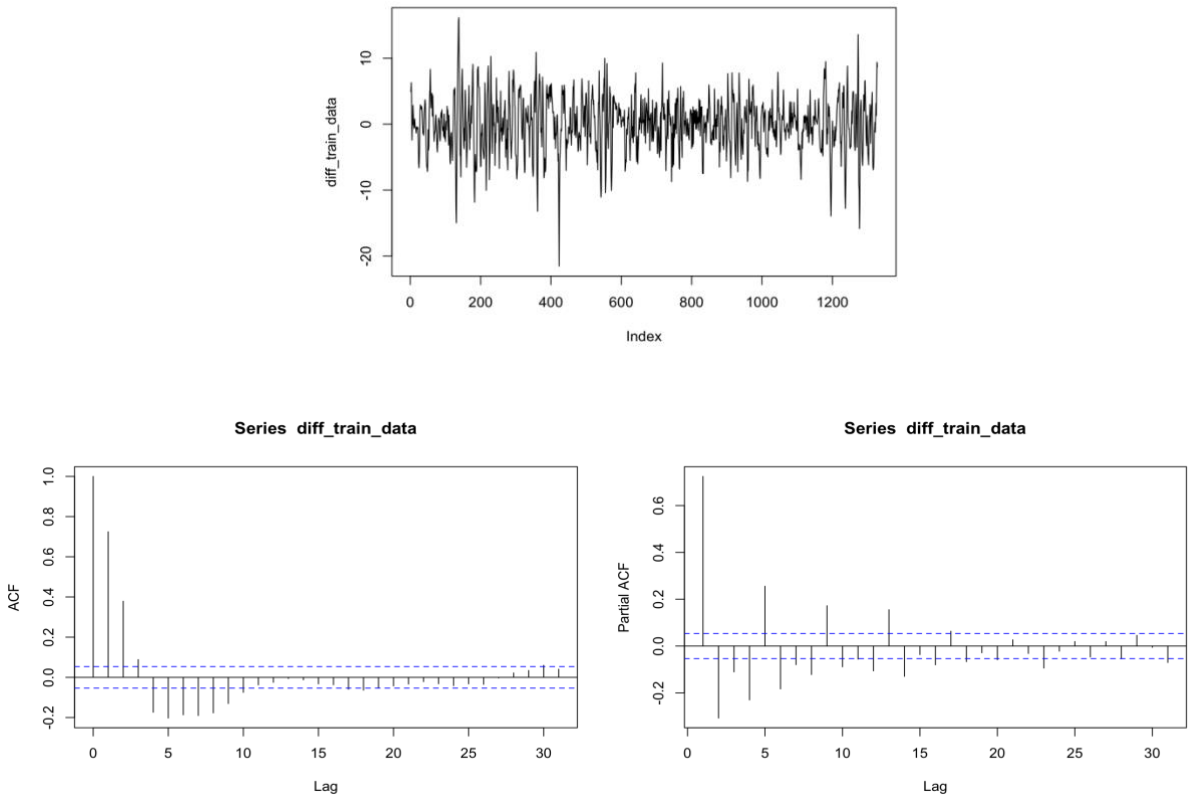


Figure 3.4.2

Observing the Sample ACF and PACF shows no slow decay or clear pattern. Since we perform one seasonal differencing in lag 4 (the period is 4) and no regular differencing, we obtain D = 1, d = 0, s = 4.

Ignoring ACF and PACF at seasonal lags s, 2s, 3s, ..., where s = 4, we find both have an obvious tails-off behaviour with a damped sinusoid. In PACF, after lag 2, there is a damp

sinusoidal (p=2). Whether after lag 1 or lag 2 (q=1 or q=2), there is a damp sinusoidal in ACF. Therefore, p = 2, q = 1. Additionally, we consider the proposed model with p = 3, q = 2 since it is possible.

If we only look at integers lags s, 3s, 3s, ..., where s = 4, we observe both ACF and PACF tails off at lag 1. It is obvious that there is no slow decay or seasonality in ACF after lag 4(Q=1) or lag 8 (Q=2) and PACF after lag 4 that P=1. Therefore, we suppose P = 1, and Q = 1 or 2. Therefore, we have the following proposed SARIMA model:

Model 1: SARAMI(2,0,1)(1,1,1)[4]	Model 2: SARAMI(3,0,1)(1,1,1)[4]
Model 3: SARAMI(2,0,1)(1,1,2)[4]	Model 4 : SARAMI(3,0,1)(1,1,2)[4]

Table 3.4.1

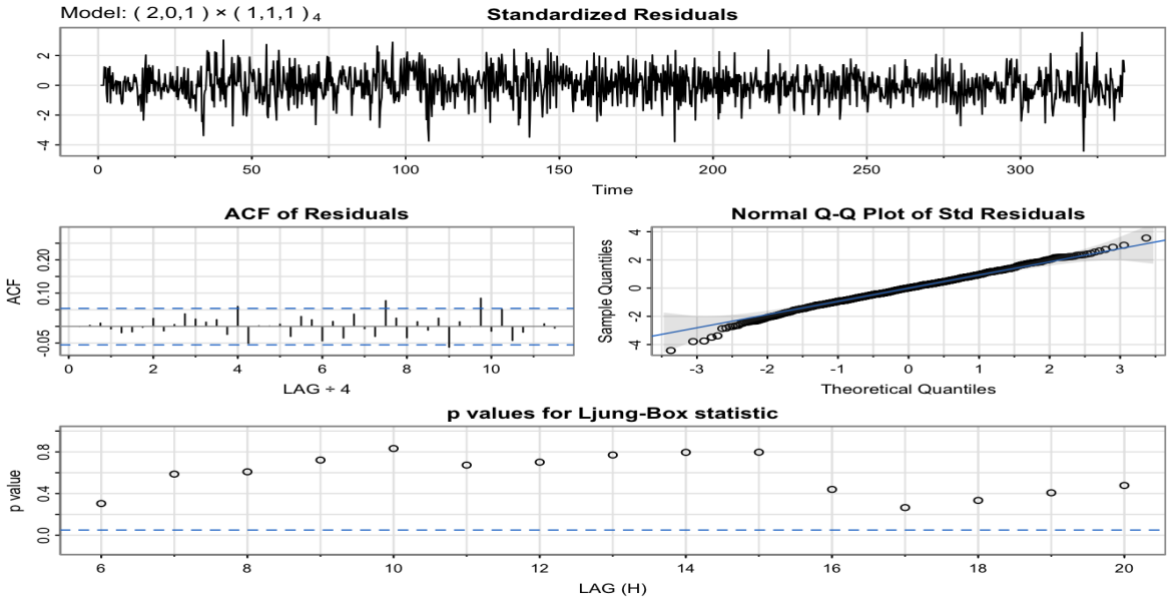


Figure 3.4.3

By R output (see Appendix), under Model 1, 2, 3, and 4, the time series plot looks random, without non-constant variance, trend, or seasonality. Through diagnostic checking, these models all have a zero mean from the Standardized Residuals plot.

For the ACF of residuals of all models, almost all the lags are inside the confidence intervals, although a few of them have exceeded the upper bound of CI; However, no more than 5% of the spikes show correlation, and these are at random places. Therefore, the model has uncorrelated residuals.

By observing the Normal QQ plot of all models, most of the observations align with the fitted line and inside the prediction interval, while only a few observations at the beginning are outside the interval. Since they are located around the lower bound of the prediction interval, the QQ plot indicates that the normality assumption holds for the model.

Lastly, the plot of tests for serial correlation shows most p-values > 0.05. By comparison, model 1 performs the best, with the highest overall p-value. In Table 3.4.2, by comparing their AIC, AICc, and BIC values, all four models show slight differences in fitted performance. Because of that, we take their APSE values to identify their prediction performance, and we find model 1 has the smallest APSE value, 12.50606.

	AIC	**AICc**	**BIC**	**APSE**
Model 1	4.405332	4.405380	4.432696	12.50606
Model 2	4.406822	4.406886	4.438096	12.51587
Model 3	4.404611	4.404675	4.435885	13.02631
Model 4	4.406073	4.406155	4.441256	13.00245

Table 3.4.2

Therefore, we conclude that model 1, **SARIMA(2,0,1)(1,1,1)[4]** is our best model. Additionally, by using auto.arima() in the forecast package in R studio, we obtain the same result, where it outputs SARIMA(2,0,1)(1,1,1)[4], aligning with our expectations. In the short term, the forecasting shows that it follows the seasonal behaviour of the original data, which has

a slightly decreasing trend; in the long term, the result turns out to be more stable, with a mean temperature degree equal to 2.

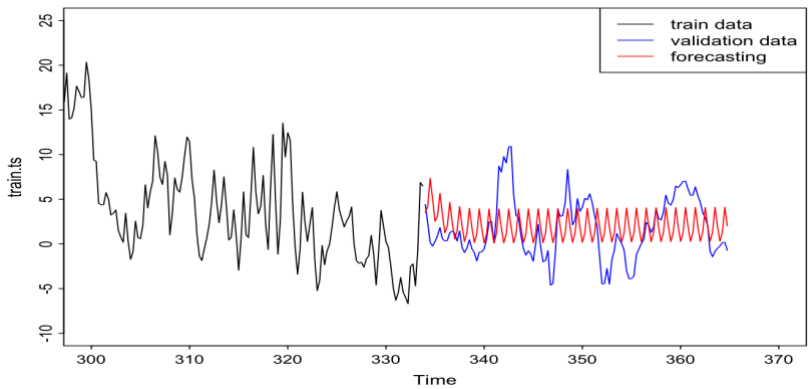


Figure 3.4.4

As we perform future forecasting for the first week in Jan 2024. The forecasting would follow the period as original data and, in the short term, shows a slightly increasing trend with temperatures above 0.

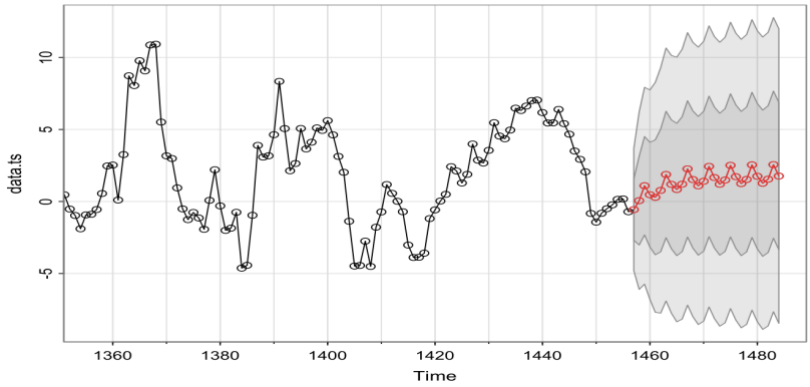


Figure 3.4.5

4. Discussion

We organized a table containing all the APSE values for model comparison through regression, smoothing, and the Box-Jenkins method. Based on the results, we conclude that the Box-Jenkins model (SARAMI(2,0,1)(1,1,1)[4]) has the least APSE value. Therefore, we select this model as our best model for predicting future temperatures.

Method	APSE
Unregularzied Regression	24.69160
Ridge	136.9263
Elastic Net Regression	133.3762
Lasso	130.0699
Simple Exponential Smoothing	13.13694
Double Exponential Smoothing	12.7599
Additive Holt-Winters Smoothing	47.41026
Multiplicative Holt-Winters Smoothing	1064.606
Box-Jenkins Model: $(2,0,1) \times (1,1,1)_4$	12.50606
Box-Jenkins Model: $(3,0,1) \times (1,1,1)_4$	12.51587
Box-Jenkins Model: $(2,0,1) \times (1,1,2)_4$	13.02631
Box-Jenkins Model: $(3,0,1) \times (1,1,2)_4$	13.00245

Table 4.1

Moreover, we conclude that both limitations exist in the regression and smoothing method. Previous prediction graphs indicate that the trend component predominantly influences the overall prediction. On the other hand, the Box-Jenkins method can capture slight seasonal movements while also considering the overall trend component for temperature prediction. Table 4.2 is the output of the prediction value using Box-Jenkins model. Based on our best model, we have shown that temperatures for the first week of 2024 are fluctuating between 0-2 degrees Celsius.

Date & time	00:00:00	06:00:00	12:00:00	18:00:00
Jan 1st 2024	-0.56453621	0.05852009	1.09488492	0.47161701
Jan 2nd 2024	0.29013466	0.76911056	1.87536754	1.19114817
Jan 3rd 2024	0.83282686	1.20219310	2.26637790	1.52250850
Jan 4th 2024	1.08720216	1.40352770	2.43683431	1.66266137
Jan 5th 2024	1.19659735	1.49068635	2.50920923	1.72208144
Jan 6th 2024	1.24426245	1.52963109	2.54198486	1.74972606
Jan 7th 2024	1.26752763	1.54959818	2.55953849	1.76532817

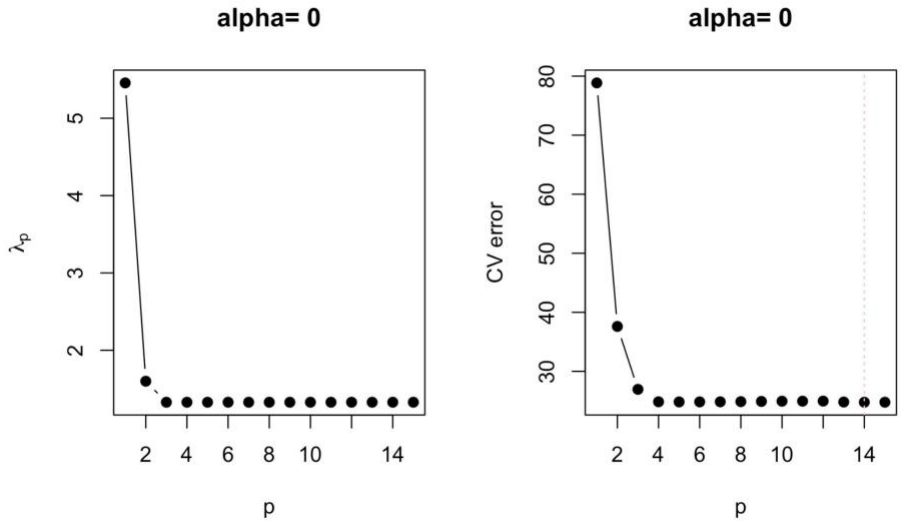
Table 4.2

However, our study has limitations. We only incorporated the temperature data from 2023 to predict temperatures for the first week of 2024, potentially requiring more accuracy in our predictions. Besides, we did not analyze other factors that could influence temperature, such as ocean currents, wind patterns, and precipitation.

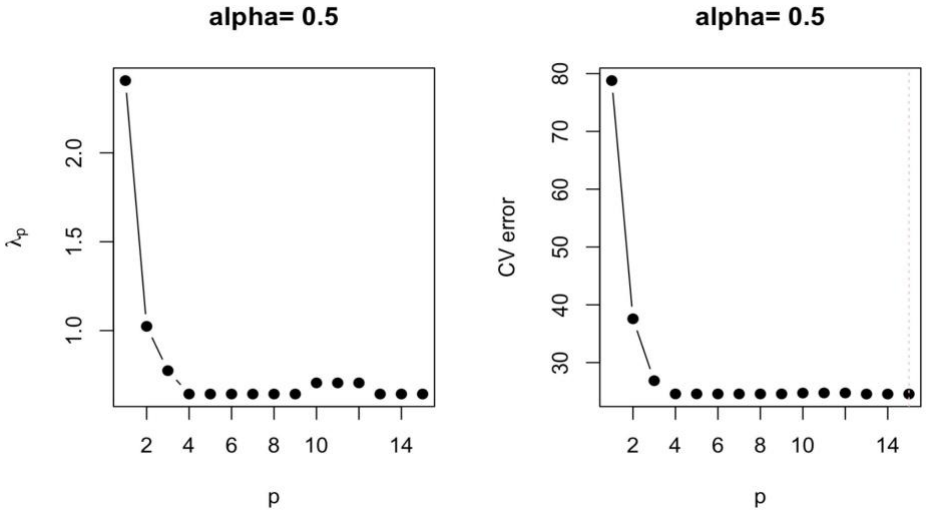
Overall, the prediction results highlight a potential slow-declining trend in temperature between 0 and 2 degrees Celsius. With these results, all residents of the Waterloo region can be prepared for any sudden change in temperature and prevent the negative impacts caused by weather fluctuations.

Appendix

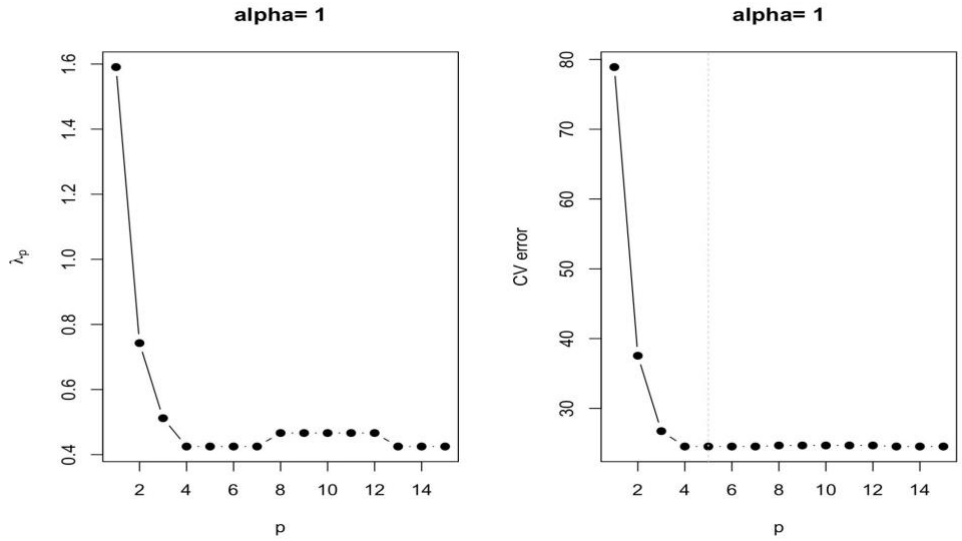
Regularized Regression



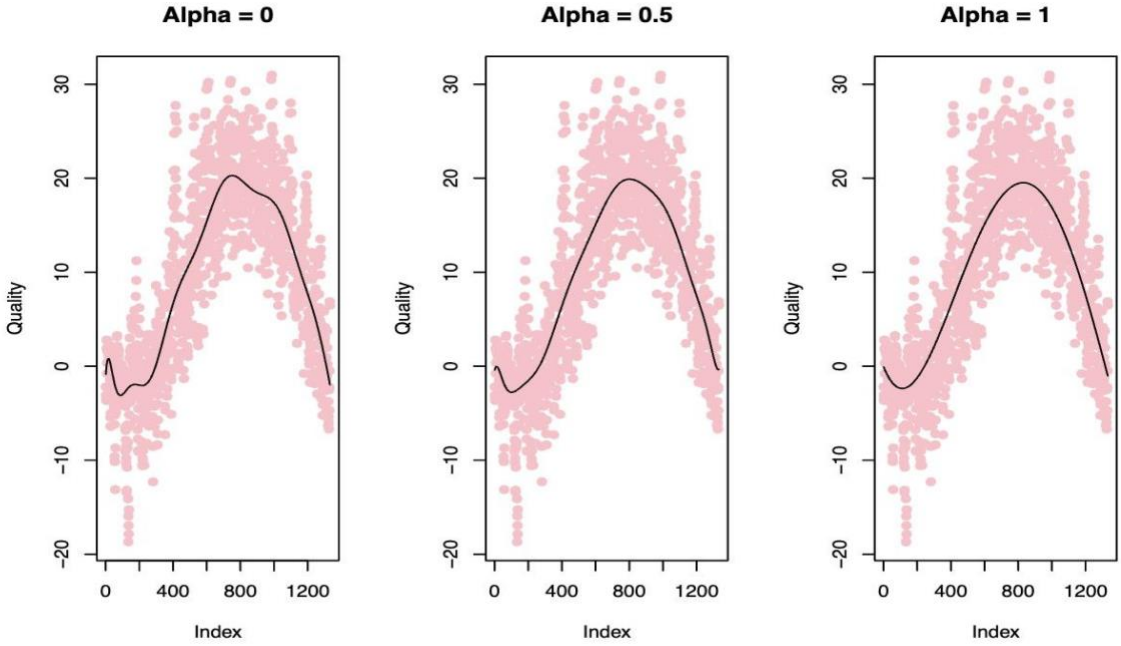
The optimal p for Ridge Regression is 14



The optimal p for Elastic Net Regression is 15



The optimal p for LASSO regression is 5



Predicted values for three regularized regression methods using the optimal p

Box-Jenkins

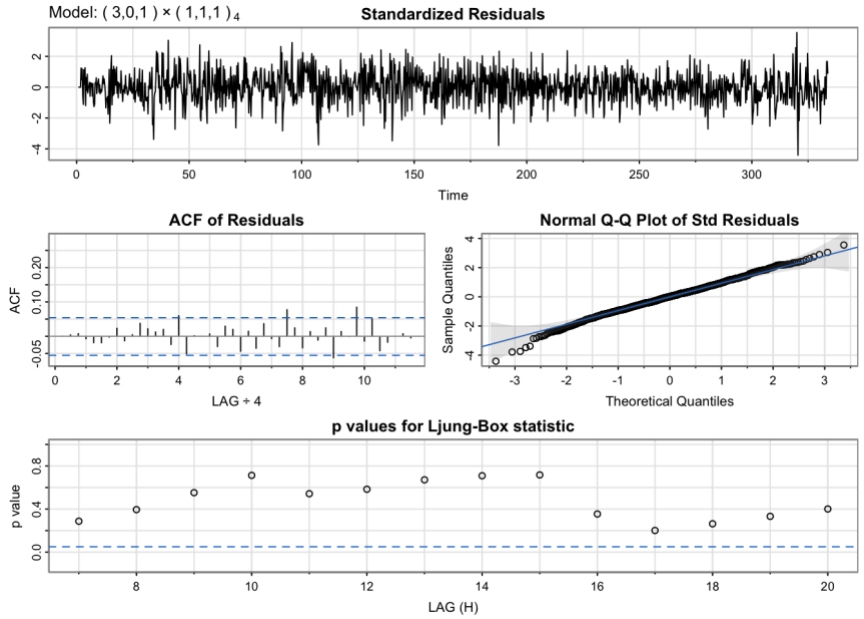


Figure 1

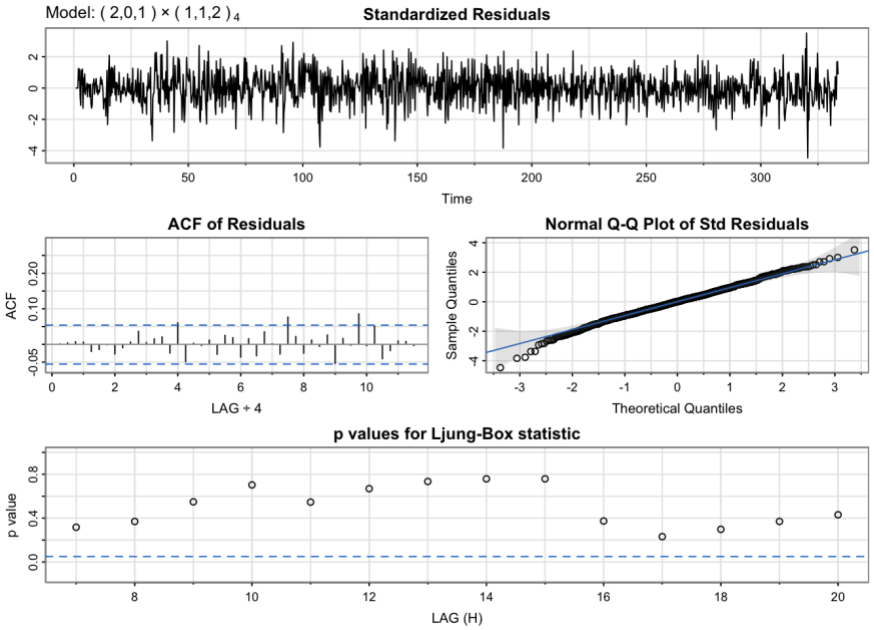


Figure 2

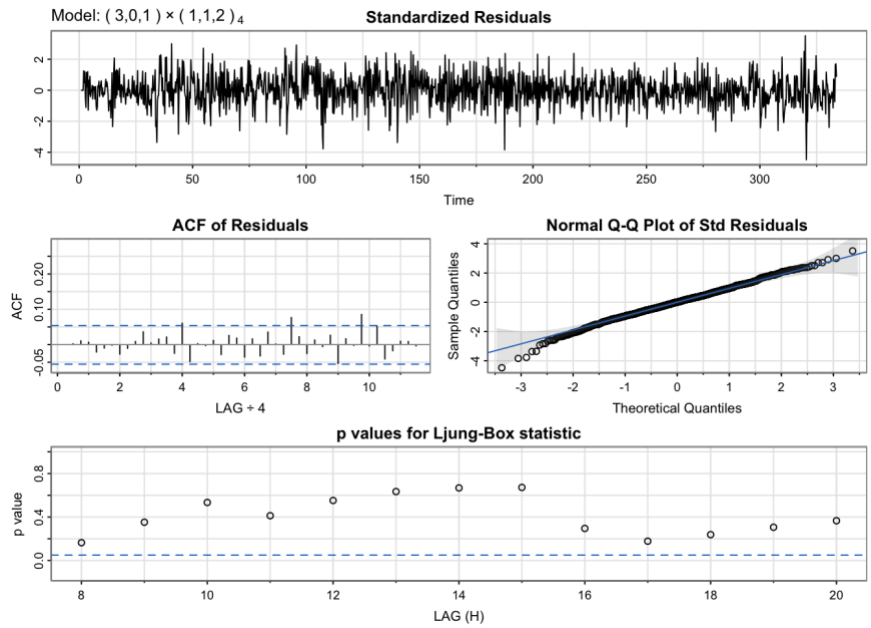


Figure 3

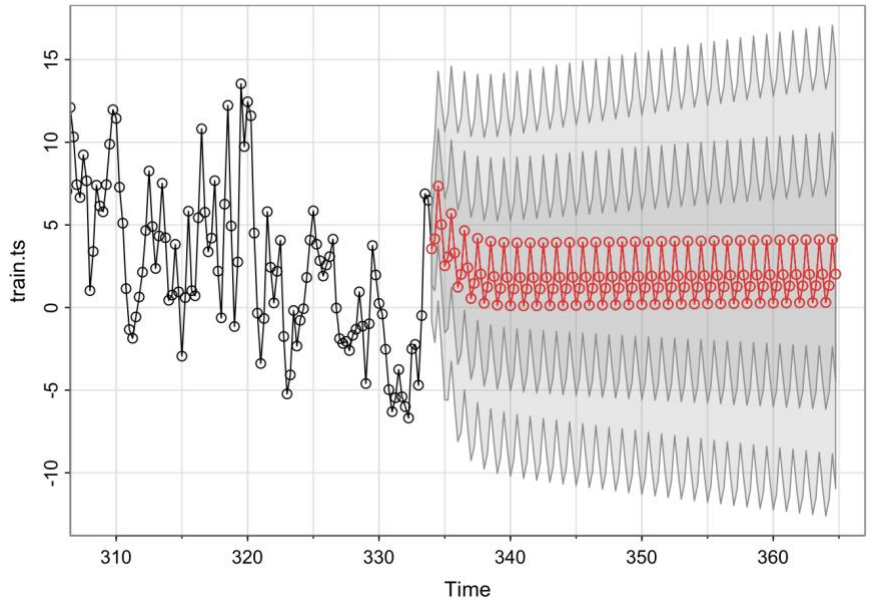


Figure 4

Figure 1, 2, 3 shows the sarima output for other 3 proposed models SARAMI(3,0,1)(1,1,1)[4], SARAMI(2,0,1)(1,1,2)[4], SARAMI(3,0,1)(1,1,2)[4] respectively.

Figure 4 indicates the output of sarima.for() for the best model SARIMA(2,0,1)(1,1,1)[4], which indicates in the short term, there is a slight decrease in temperature, while remaining stable afterwards.