Time Series Analysis for Major US City Temperatures

Stat 248, Spring 2016

Sici Huang

May 9, 2016

1 Introduction

The scientific question that will be considered in this study is whether time series can capture annual and year-to-year trends of ground-level air temperature for major U.S. cities. Additionally, effects of climate cycles on air temperature over land and their significance levels will be discussed. The study on climate change has become increasingly crucial since it is closely linked to crop production and conservation efforts of endangered species. One of the major culprits of global warming is the rising percentage of greenhouse gases in the atmosphere. Cities, responsible for more than 70% of greenhouse gas emissions[1], also suffer from its consequences such as the heat island effect.

In this paper, I will examine monthly average air temperatures of the three most populated cities in the United States—New York City, Los Angeles, and Chicago—in order to reveal underlying temperature movement patterns and to predict future temperatures. To achieve this objective, I will first perform exploratory data analysis and produce summary statistics plots of monthly average temperatures of the three cities. Then, by fitting linear regression, ARIMA, and harmonic regression models, I will predict monthly temperatures for future years. Finally, I will conduct spectral analysis to study frequency content of the temperature data.

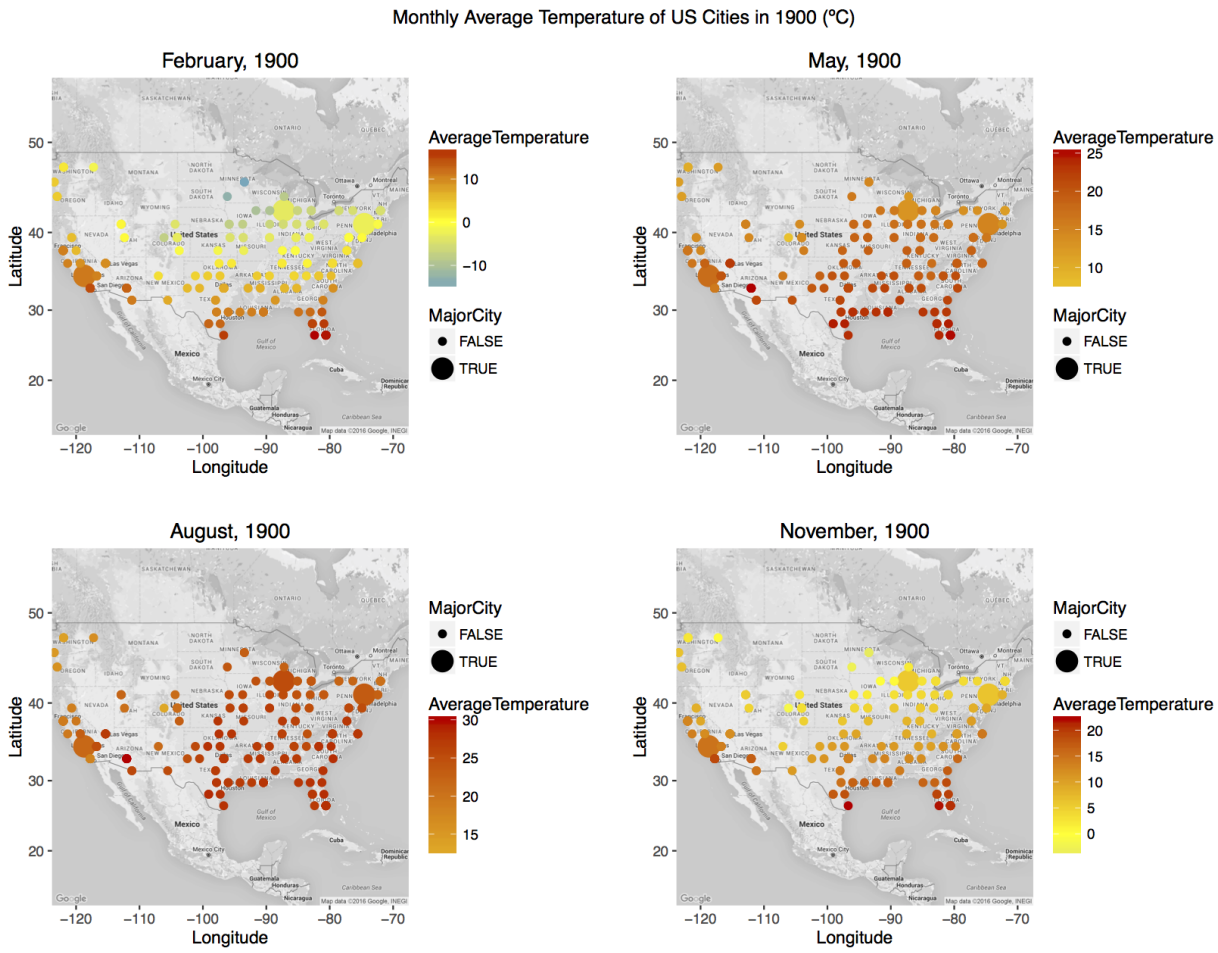


Figure 1: US City Mid-season Temperatures (year 1900)

2 Data

2.1 Data Description

The dataset of global temperatures was collected over 263 years (1750–2012) and downloaded from Berkeley Earth, a non-profit environmental monitoring organization. The dataset contains monthly average temperatures of cities from 243 countries and geographic coordinates of the cities.

For the purpose of this study, I subsetted the dataset to include only U.S. city temperatures.

As shown in Figure 1 and Figure 2, the mid-season temperatures increased by a noticeable amount between 1900 and 2000. However, this observation alone does not imply the overall rise of U.S. ground-level air temperature nor the presence of a linear trend. Further investigation is needed.

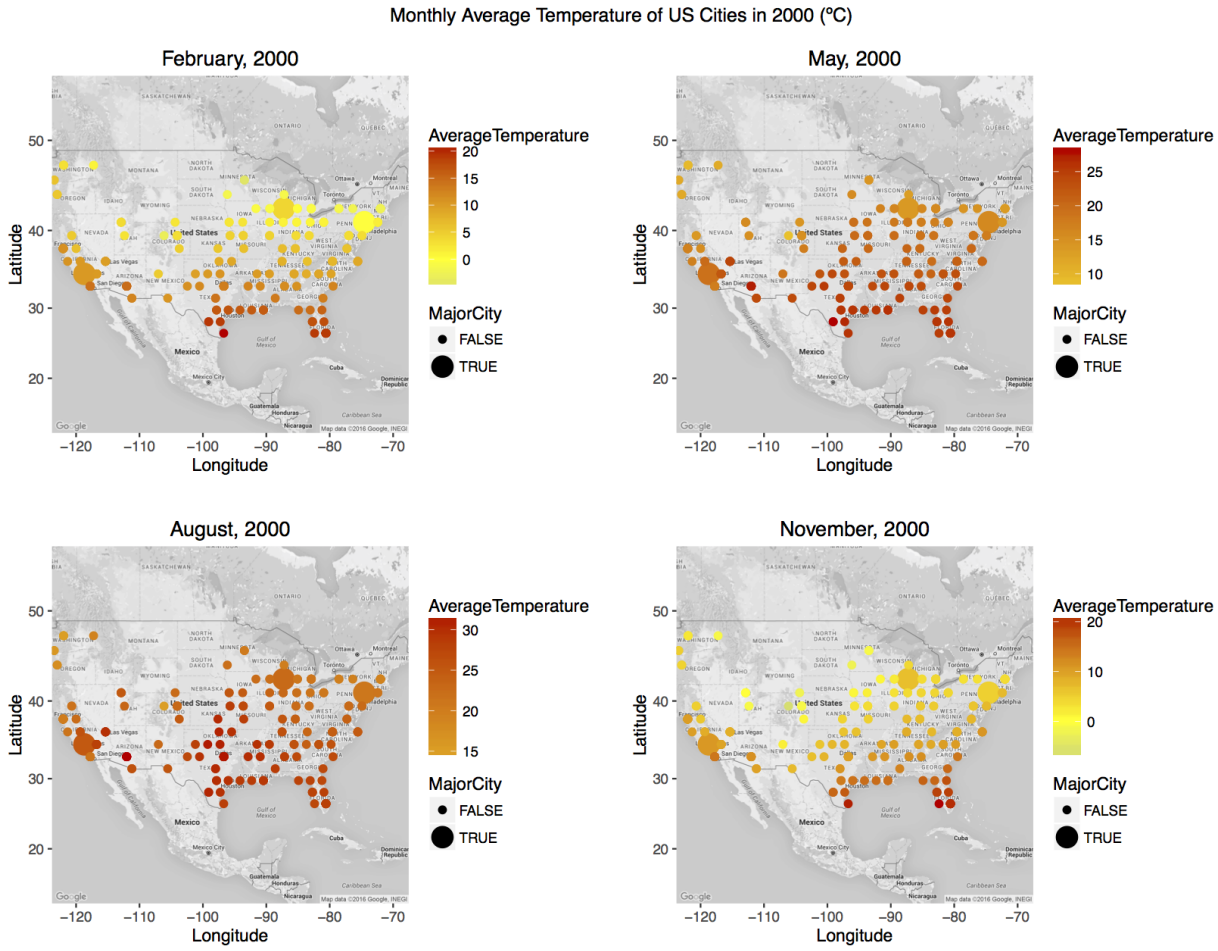


Figure 2: US City Mid-season Temperatures (year 2000)

2.2 Why NYC, LA, and Chicago?

Due to the high concentration of human activities, global climate change is prominently observed in major cities. Being the three most populated cities in the United States with populations of 8175133, 3792621, and 2695598 respectively [2], New York City, Los Angeles, and Chicago (NYC, LA, and CHI below) might be representative of the other cities in their

respective regions in terms of temperature shifts. Furthermore, as shown in Appendix 6.1, temperatures of the three cities indeed exhibit the same overall patterns as that of their states. By studying temperature changes in these cities, we might be able to gain certain insights on U.S. regional temperature fluctuations.

3 Exploratory Data Analysis

We start by plotting the raw temperature data of NYC, LA, and CHI (see Appendix 6.1 and right panel of Figure 3). The line plots reveal that monthly average temperature follows an annual cycle, and the annual cycle oscillates in a not yet clear pattern. Although further testing is required, temperature data for all three cities seems to be stationary (further analysis can be found in the ARIMA section). The histograms in Figure 3 resemble the shape of a saddle; notice that the distributions of NYC and CHI temperatures are more spread out while the distribution of LA temperature is more concentrated. Moreover, LA temperature is higher overall, which is to be expected due to its lower latitude. The aforementioned plots indicate the absence of obvious outliers.

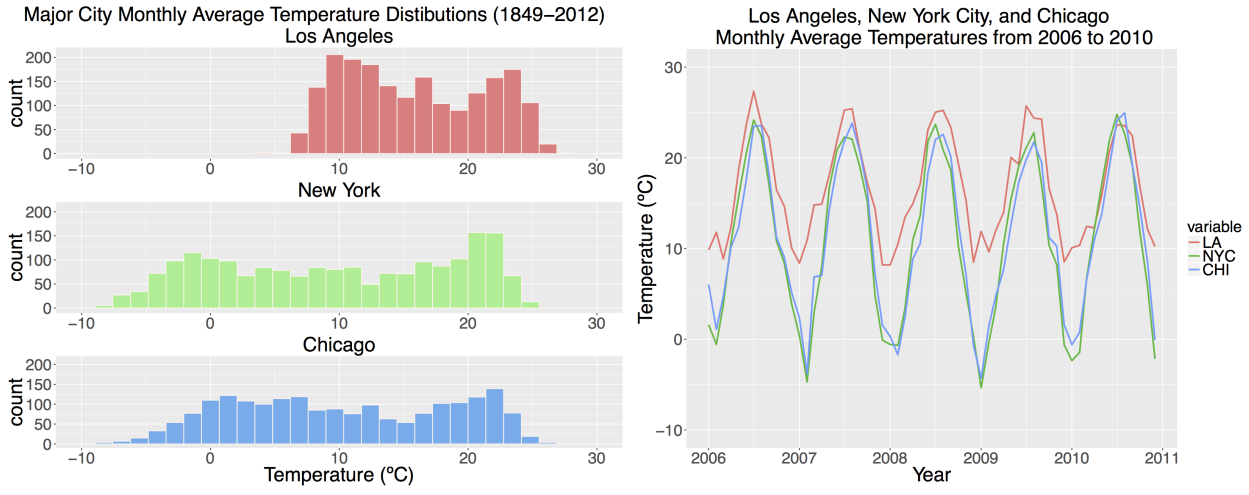


Figure 3: EDA Summary Plots

3.1 Seasonal Decomposition

To better understand trend and seasonality of the data, we carry out seasonal decomposition on the monthly average temperatures following the model:

$Y_t = T_t + S_t + R_t$, with $t = 1, 2, ..., (263yrs * 12mos)$

where T_t is the trend component determined using loess regression, S_t is the seasonal component, and R_t is the remainder component. As shown in Figure 4, unsurprisingly, LA temperature data contains a clear annual cycle. The trend component displays a slight upward tendency. Although judging by the sizes of the vertical reference bars on the right side of the plots, the magnitude of the trend was unlikely to be significant. The remainder component resembles white noise which signifies that the majority of the seasonality and trend were successfully captured in the seasonal and trend components. Following the same analysis, similar conclusions were reached for NYC and CHI. Their seasonal decomposition plots can be found in Appendix 6.2.

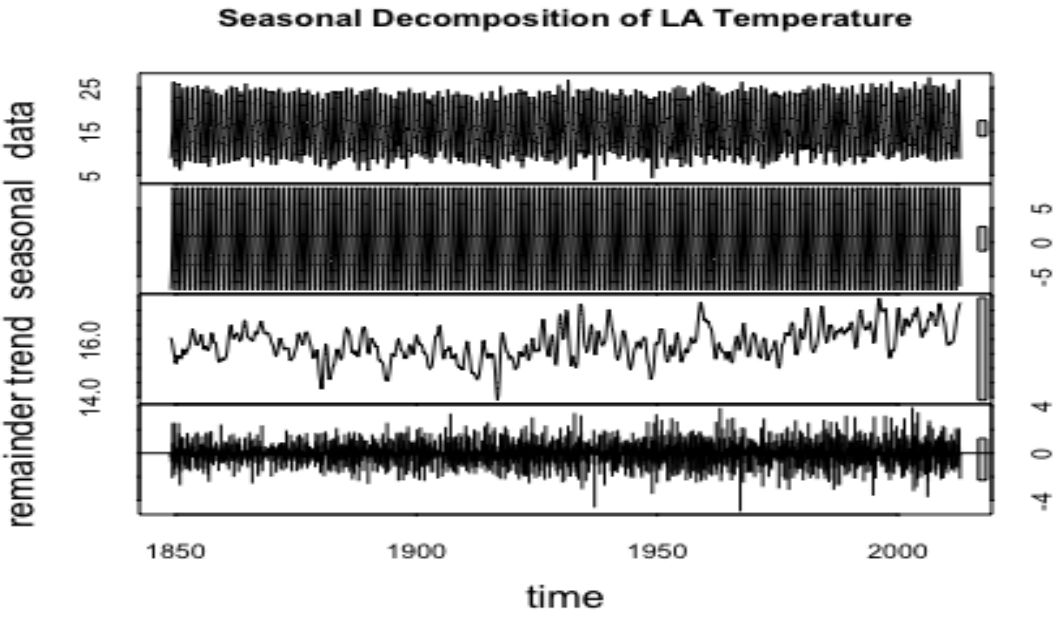


Figure 4: LA Temperature Seasonal Decomposition

4 Model Fitting

4.1 Linear Regression

We start by fitting a simple linear regression:

$Temp = b_1 * Mo_1 + b_2 * Mo_2 + b_3 * Mo_3 + b_4 * Mo_4 + b_5 * Mo_5 + b_6 * Mo_6 + b_7 * Mo_7 + b_8 * Mo_8 + b_9 * Mo_9 + b_{10} * Mo_{10} + b_{11} * Mo_{11} + b_{12} * Mo_{12} + b_{13} * Yr$, with $t = 1, 2, ..., (263yrs * 12mos)$

where $Temp = LA, NYC, CHI$ is the monthly average temperature of the city, $Mo_1, ..., Mo_{12}$ are dummy variables corresponding to the 12 months, and $Yr = 1849, ..., 2012$ is the year.

As shown in the table below, p-values of Los Angeles, New York City, and Chicago reveal that regressor Yr is significant at less than the 0.001 level. Since the coefficient of Yr was positive for all three cities, we can deduce a slight upward trend from year to year in average temperature from this model. Upon further inspection, I found that the distribution of the regression residuals is slightly light-tailed (see Appendix 6.3 Q-Q plot) but overall close to the normal distribution (see Appendix 6.3 histogram). However, the regression residuals displayed a non-random pattern which indicates a poor fit for this linear model (see Appendix 6.3 scatter plot).

Temp	Coeff. of Yr	p-value of Yr	Adjusted R^2 of model
LA	0.0047105	4.21e-15	0.9944 ***
NYC	1.047e-02	< 2e-16	0.9848 ***
CHI	9.690e-03	< 2e-16	0.9766 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.2 ARIMA

To improve our linear regression model, we consider adding autoregressive and moving average terms, hence the ARIMA model.

To fit an ARIMA model, we first decide whether or not our time series is stationary by inspecting ACF and PACF plots of the series. As shown in Figure 5, LA temperature seems to be stationary. We further test its stationarity by conducting the Augmented Dickey-Fuller test. Since the p-value was less than 0.01, we reject the null hypothesis and conclude that LA temperature data is stationary. Following the same approach, we found that both NYC and CHI temperatures were stationary. Therefore, a differencing term is unnecessary in the ARIMA models (see Appendix 6.4.1 for ACF and PACF plots of NYC and CHI).

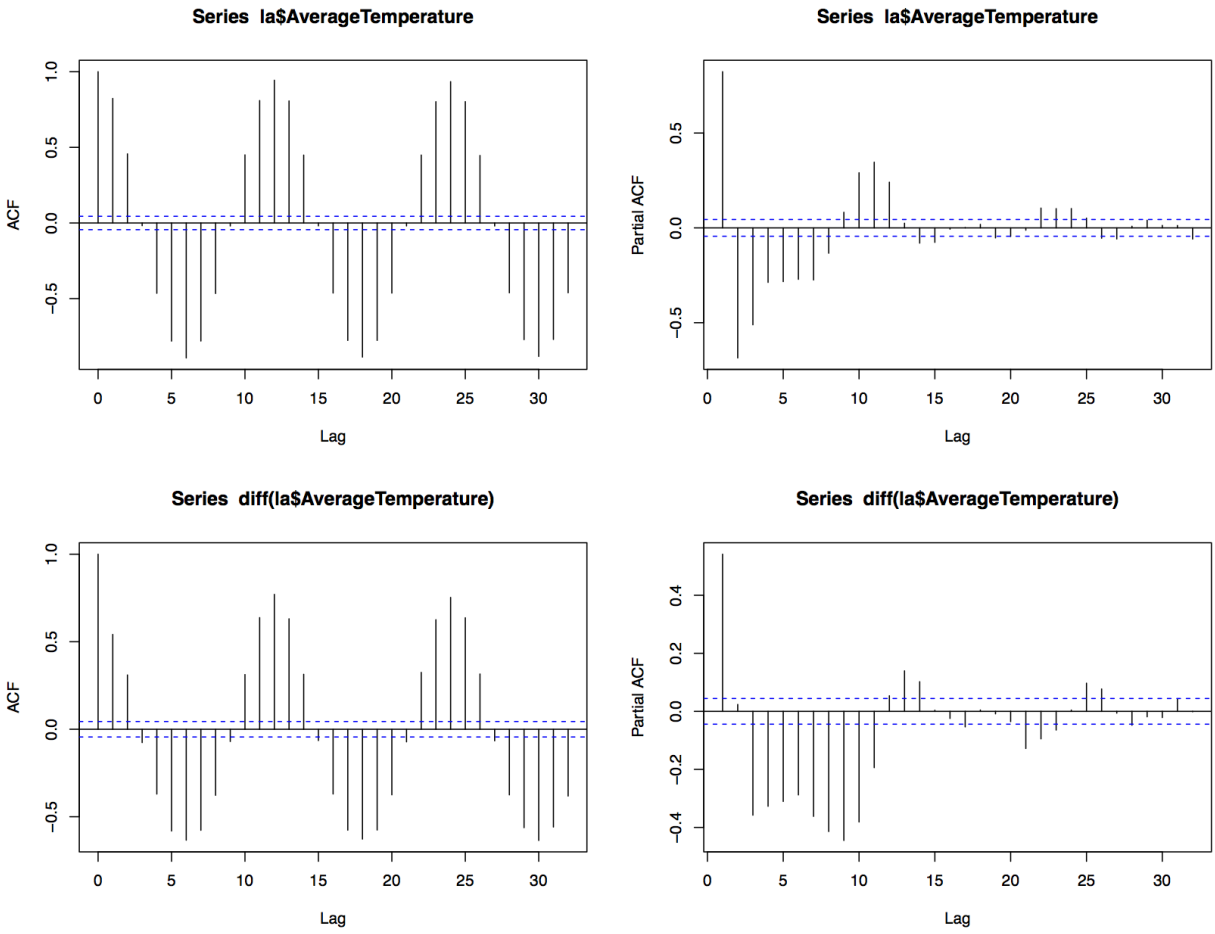


Figure 5: ACF and PACF of LA Temperature with and without Differencing

After deciding the order of differencing, we proceed to select the number of AR and MA terms. Since the monthly average temperature data contains a clear annual trend, a seasonal

ARIMA model is called upon. Moreover, because the effects of AR and MA terms to some extent cancel each other out, we follow a forward selection process: after fitting each model, we inspect the ACF and PACF plots of the residuals and decide if additional AR or MA terms are needed. We then take the top 5 possible models and choose the best model based on BIC. Here I used BIC instead of AIC because BIC penalizes overly complex models, which helps prevent overfitting. The selected models for each city along with their BIC's can be found in the table below. To quantify model accuracy, I divided the dataset into a training set (1849-2007) and a test set (2008-2012), using only the training set to fit the model. Next, we analyze the residuals of the selected models by looking at their ACF and PACF plots (see Appendix 6.5) and comparing their distributions to the normal distribution. The Q-Q plot of LA temperature residuals demonstrates that the overall distribution of the residuals is close to the normal distribution though the former is slightly light-tailed. Such a distribution indicates proper selection of the model because the normality assumption was met (see left panel of Figure 7). Q-Q plots of NYC and CHI displayed similar distributions (see Appendix 6.6). Therefore, the selected models were adequate for our data.

City	Model	BIC	MSE on test set
LA	ARIMA(1,0,1)x(1,0,1)	6238.159	1.568664
NYC	ARIMA(1,0,1)x(1,0,1)	7421.719	2.963375
CHI	ARIMA(1,0,0)x(1,0,1)	8191.665	4.028274

Figure 6: Selected Models

Using the selected models, we predict temperature from 2008-2017 and evaluate model performance by calculating the MSEs (see Table 18). As shown in the right panel of Figure 7, the ARIMA model quite accurately predicted the yearly trend in LA temperature. However, although the MSEs were relatively low, the model did not pick up much of the year-to-year trend.

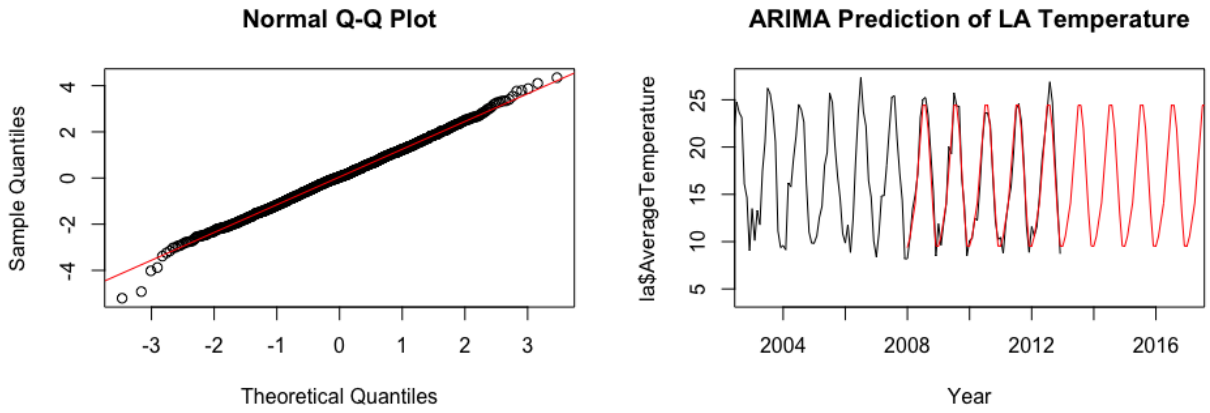


Figure 7: Predictions and Residual Q-Q Plot of LA ARIMA(1,0,1)x(1,0,1) (Right panel: black line–real data, red line–prediction)

4.3 Harmonic Regression

Climate cycles such as *El Niño* Southern Oscillation (every 2-7 years) and sunspot cycles (around every 11 years) have been proven to have various levels of influence on global climate [3]. To study the effects of the climate cycles, we attempt to perform a harmonic regression which regresses the temperature data on a combination of harmonic (sin and cos) waves. Due to the nature of monthly average temperature data, we should expect the sin and cos waves with one year periods to be important. We also include waves with periods of 1/2-year, 3.5-years, 6-years, and 11-years in order to take into account the effects of *El Niño* and sunspot cycles.

p-values of the period coefficients indicate that 1/2-year, 1-year, 3.5-year, 6-year, and 11-year periods were significant for NYC and CHI while 1/2 year and 1-year periods were significant for LA (see Appendix 6.7). Similar to the ARIMA model, we regress on the training set (1849-2007) and predict temperature from 2008-2017. As shown in Figure 8, the harmonic regression produced decent annual predictions of LA temperature, but unfortunately, did not pick up the year-to-year trend. In terms of MSE, harmonic regression performed slightly

worse than ARIMA for all three cities. More careful selection of the harmonic waves might help improve the MSEs (see Appendix 6.7).

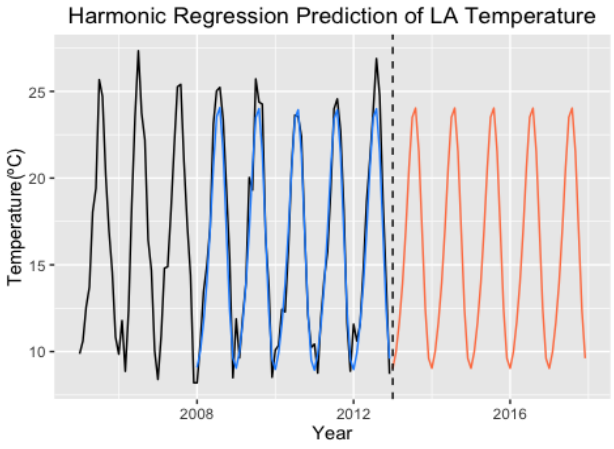


Figure 8: Harmonic Regression Prediction of LA Temperature (black line–real data, blue line–prediction on test set, orange line–prediction of the future

5 Spectral Analysis

To expand on harmonic regression, I conducted spectral analysis on the temperature data by analyzing its periodograms and spectrograms. The theoretical reasoning behind the periodogram is that by fitting N sin waves and N cos waves to data of size N, the regression will perfectly predict the data. Although the regression will be overfitted, it is possible to gain insights on the frequency content of the data by studying which coefficients are significantly different from zero. From the raw log-scaled periodograms, it seems that 1-year, 1/2-year, and 1/3-year periods were significantly different from zero for LA and CHI while only the 1-year period was significant for NYC (see Appendix 6.8.2). However, these raw periodograms are hard to interpret due to high levels of noise. To address this problem, I performed kernel smoothing using Daniell kernels with weighted short moving average ($m = 9$). The smoothed periodograms of LA and CHI displayed strong significance for 1-year and 1/2-year periods and mild significance for 1/3-year period while that of NYC displayed strong significance for

only 1-year period (see Figure 9 and Appendix 6.8.3); this result is consistent with what the raw periodograms revealed. Additionally, by tapering the smoothed periodograms by 10% and 30%, the magnitude of significance for the above periods was made even more prominent (see Appendix 6.8.4). However, upon inspecting the spectrogram of the continuous Morlet wavelet transform, the majority of LA temperature data only matched the yearly cycle (see Figure 10). Similar results were produced for NY and CHI temperature data. Their spectrograms can be found in Appendix 6.8.5.

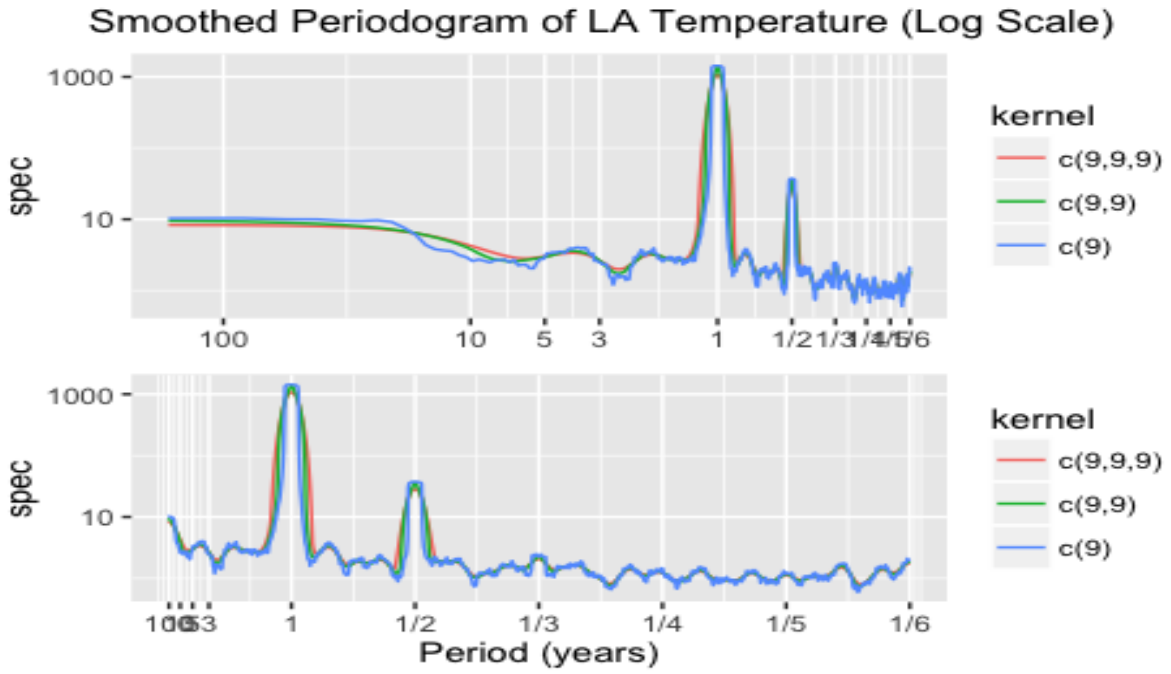


Figure 9: Smoothed Periodogram of LA

6 Conclusion

The results from the above analyses demonstrate the predictive ability of time series models—mainly the ARIMA model and the harmonic regression model—on annual ground-level air temperature trends. However, they both failed to capture year-to-year trends, which is more essential for studying global climate change. Results of harmonic regression reinforced the

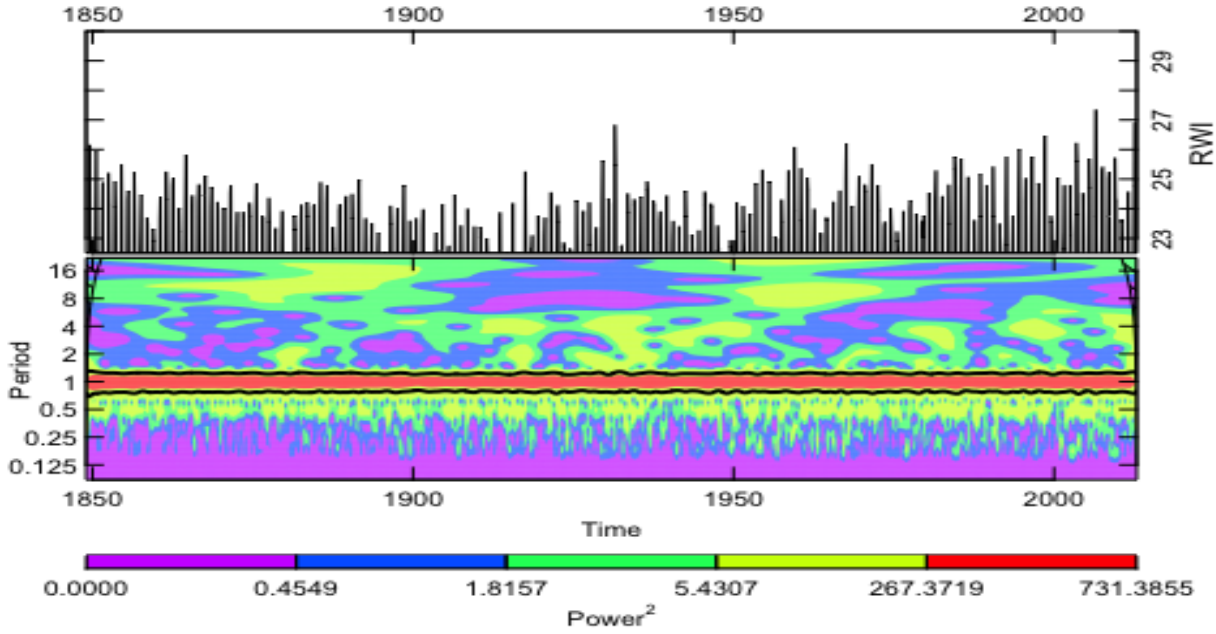


Figure 10: Spectral Spectrum of LA

fact that *El Niño* and sunspot cycles have significant effects on air temperature over land. Although the linear regression model indicated the increasing trend of temperature from year to year, the results were not conclusive. In future studies, one can improve the harmonic regression model by applying the conclusions from the frequency analysis. In addition, one can examine the relationship between temperature changes and historical events and separate human and nature factors contributing to climate change.

References

[1] *Cities and Climate Change*. (2016). Sustainable Urban Futures. Retrieved from http://urban.ias.unu.edu/index.php/cities-and-climate-change

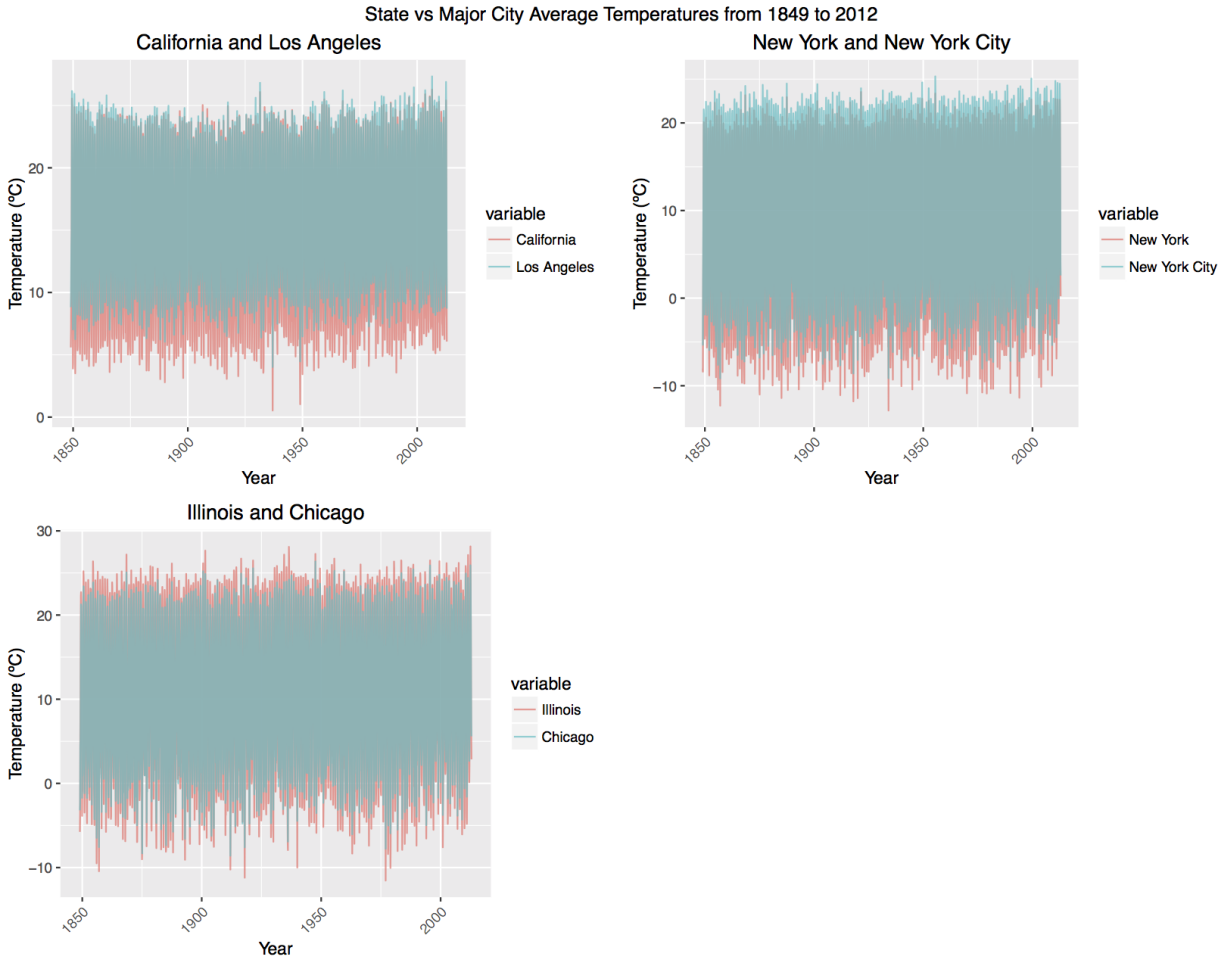
[2] *American FactFinder–Results*. (2014). United States Census Bureau, Population Division. Retrieved from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk

[3] *El Nino, La Nina, and the Southern Oscillation*. (2012). Met Office. Retrieved from http://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/el-nino-la-nina/enso-description

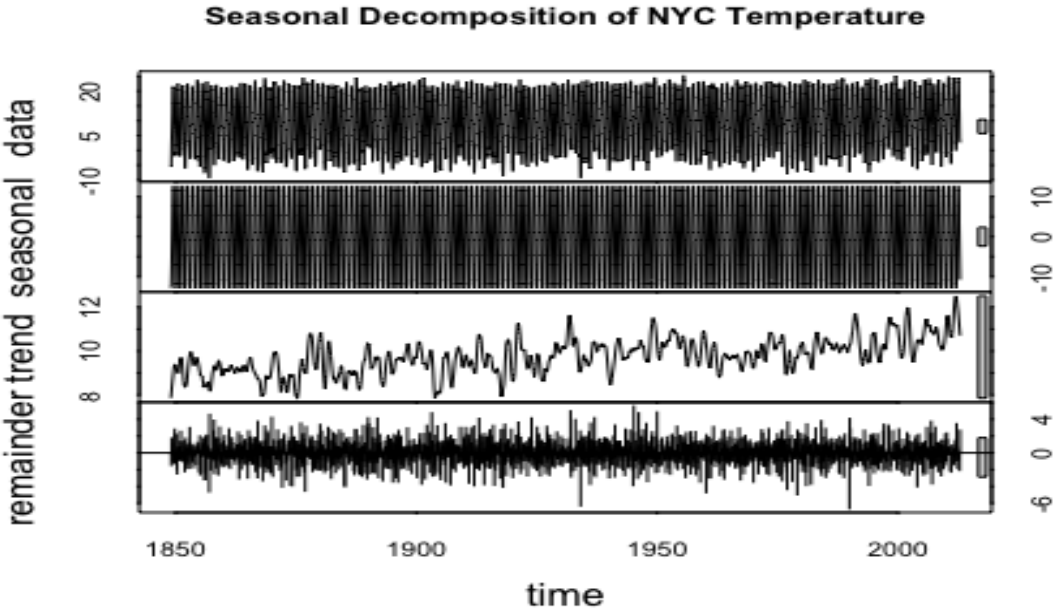
[4] *Spectral Analysis of Time Series*. Rstudio Pubs. Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/9428_1197bd003ebd43c49b429f22ea4f36e5.html

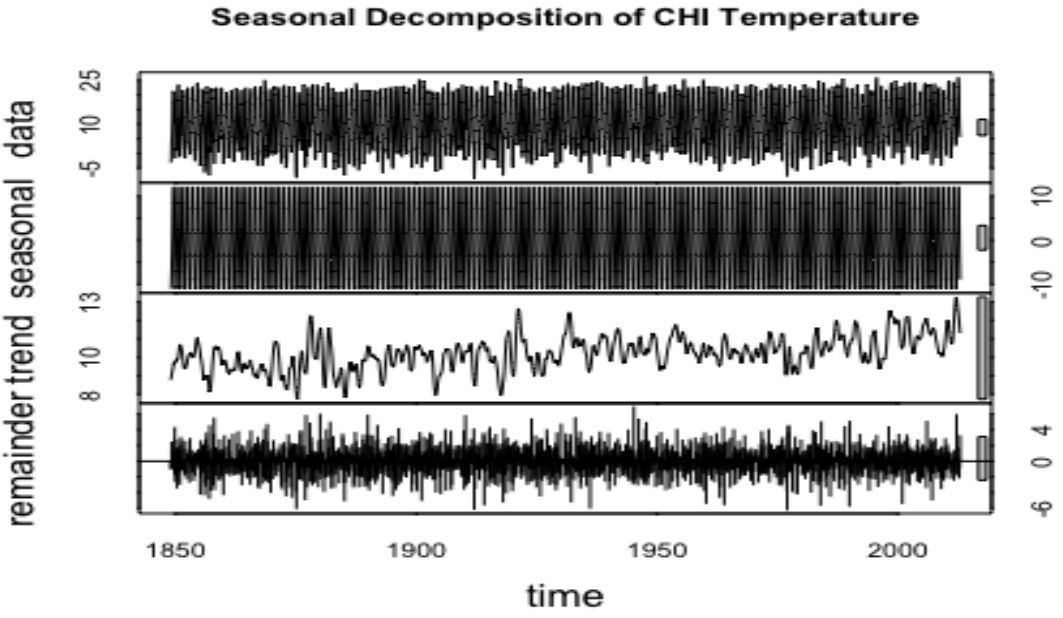
Appendices

6.1 State vs Major City Average Temperatures

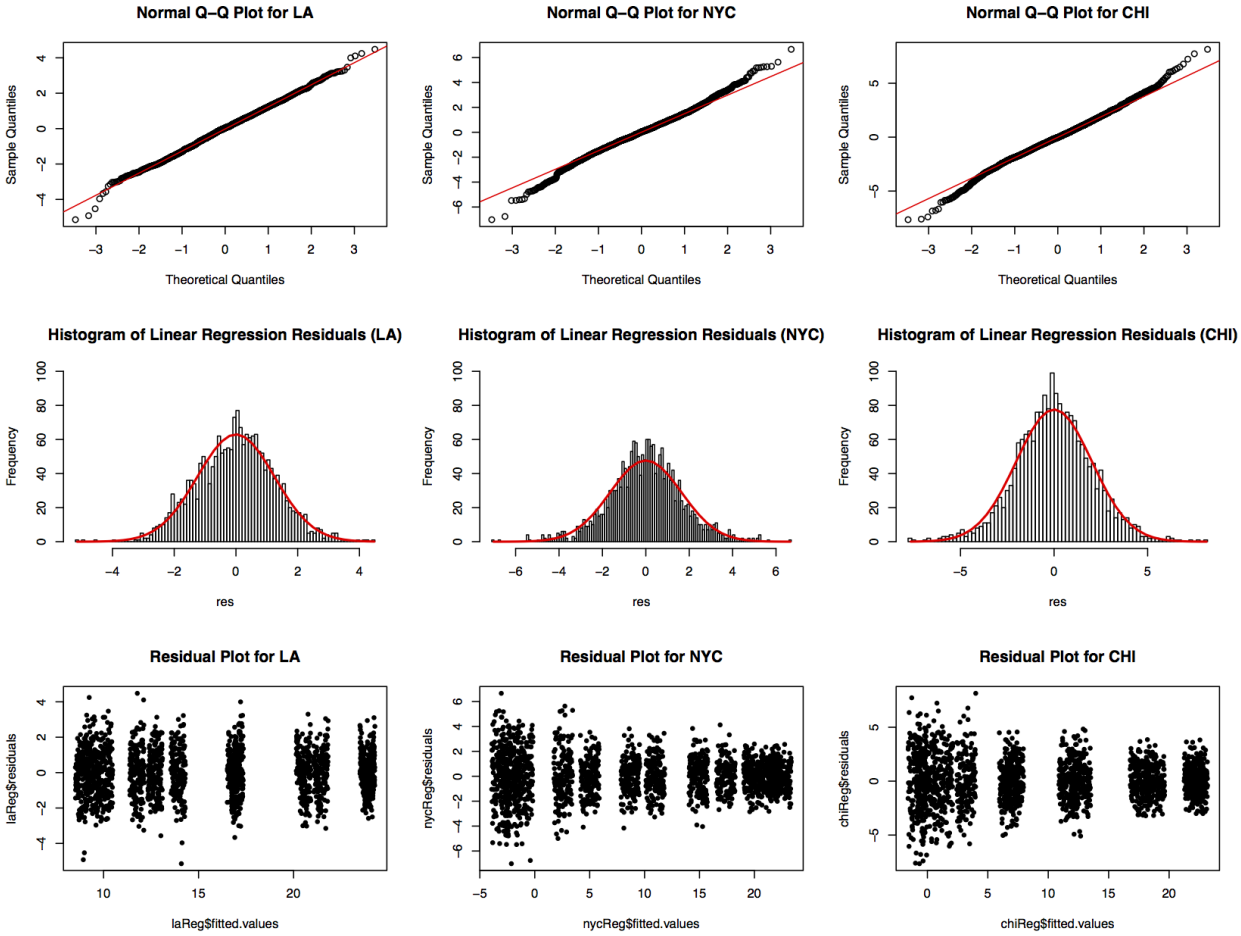


6.2 Seasonal Decomposition





6.3 Linear Regression Residual Analysis



6.4 ARIMA

6.4.1 ACF/PACF

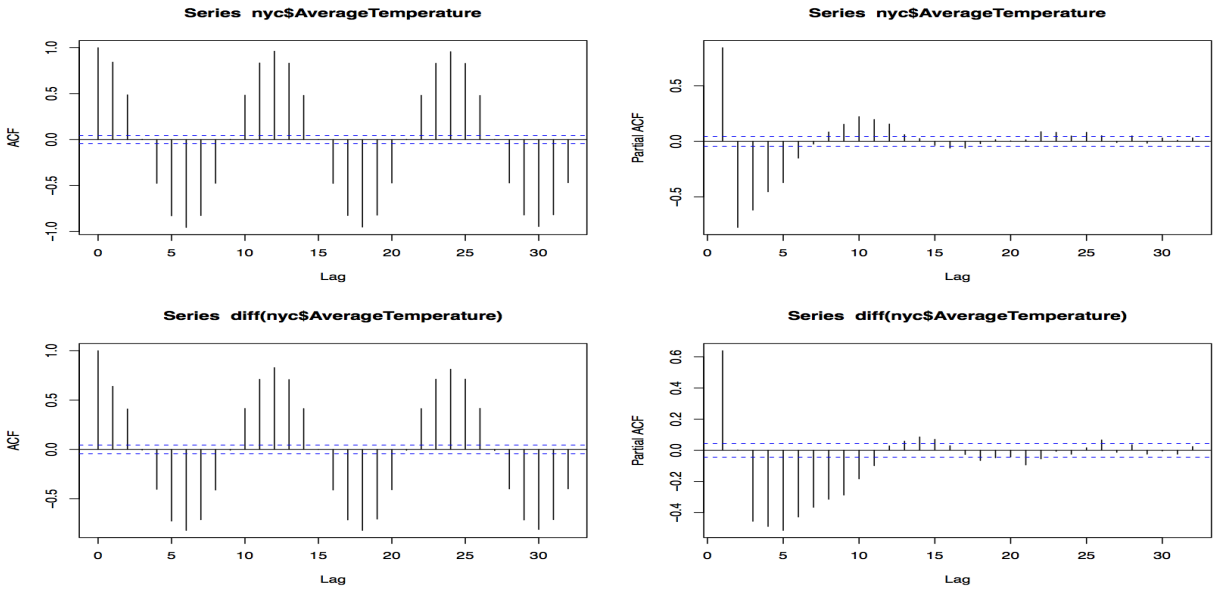


Figure 11: ACF and PACF of NYC Temperature with and without Differencing

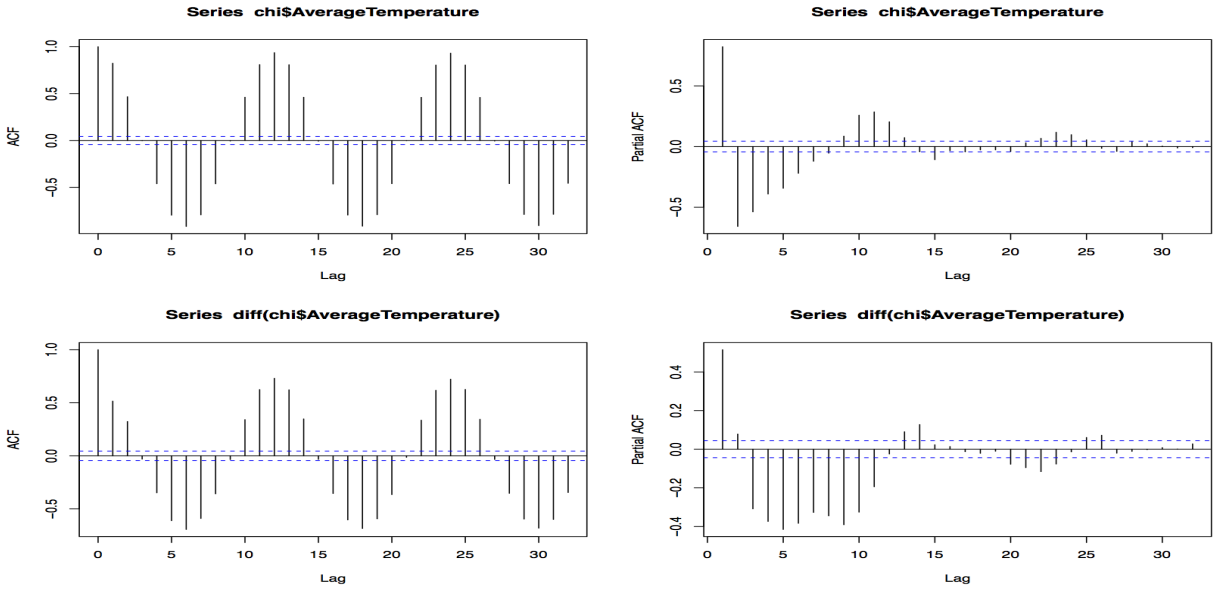


Figure 12: ACF and PACF of CHI Temperature with and without Differencing

6.5 Residual Analysis

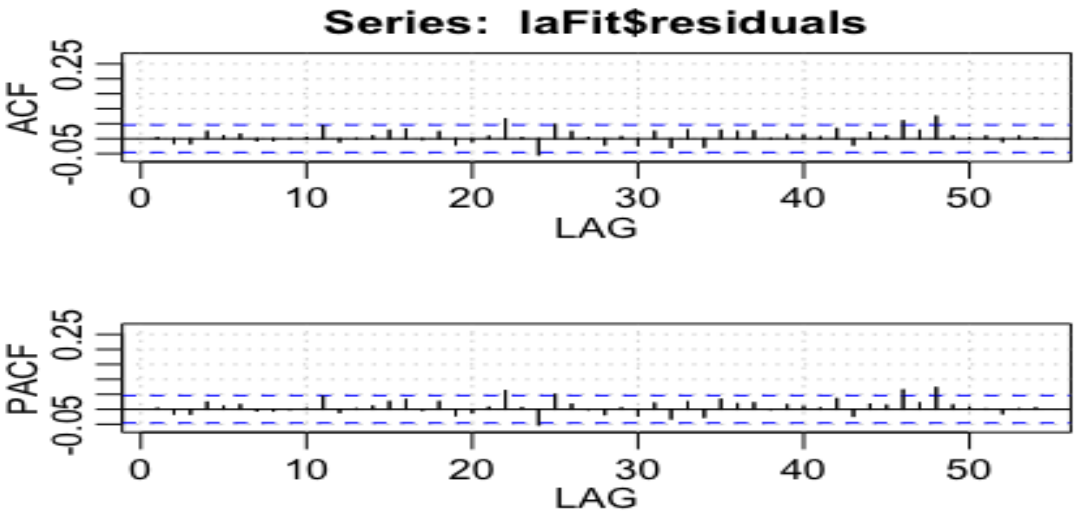


Figure 13: LA ARIMA Residual ACF/PACF

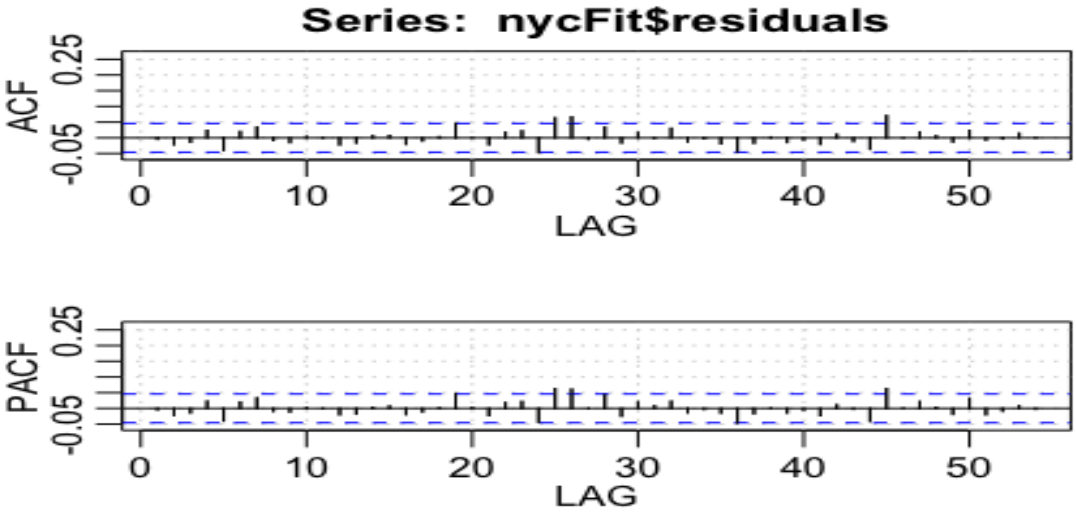


Figure 14: NYC ARIMA Residual ACF/PACF

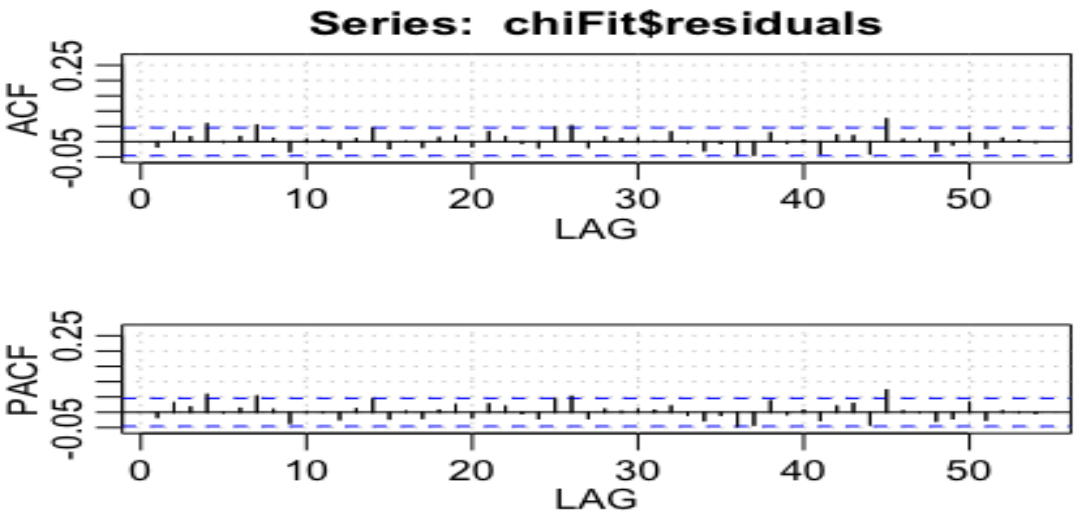


Figure 15: CHI ARIMA Residual ACF/PACF

6.6 Residual Q-Q Plots

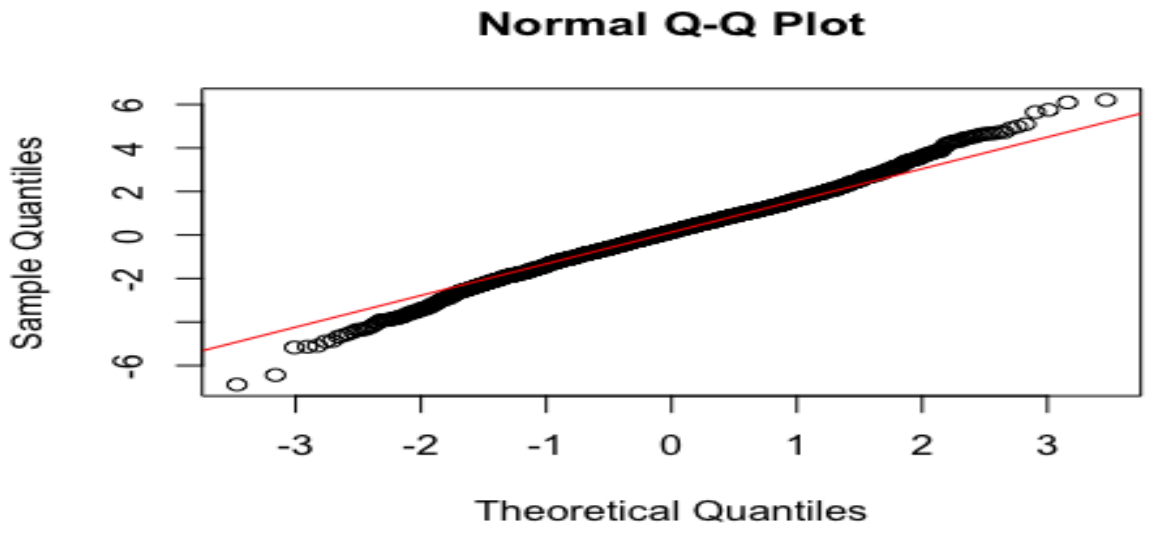


Figure 16: NYC ARIMA Residual Q-Q Plot

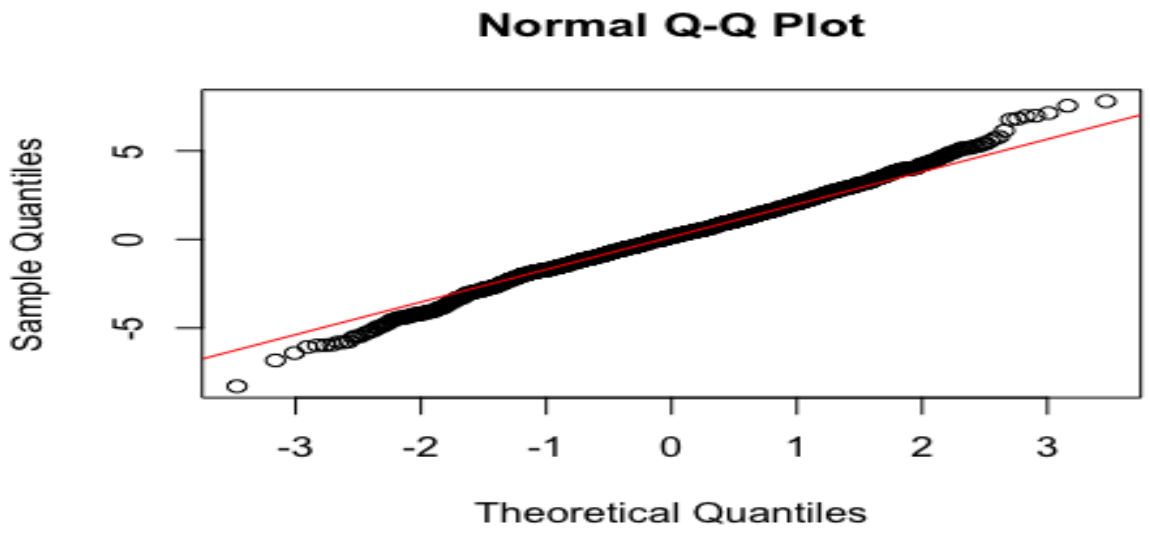


Figure 17: CHI ARIMA Residual Q-Q Plot

6.7 Harmonic Regression

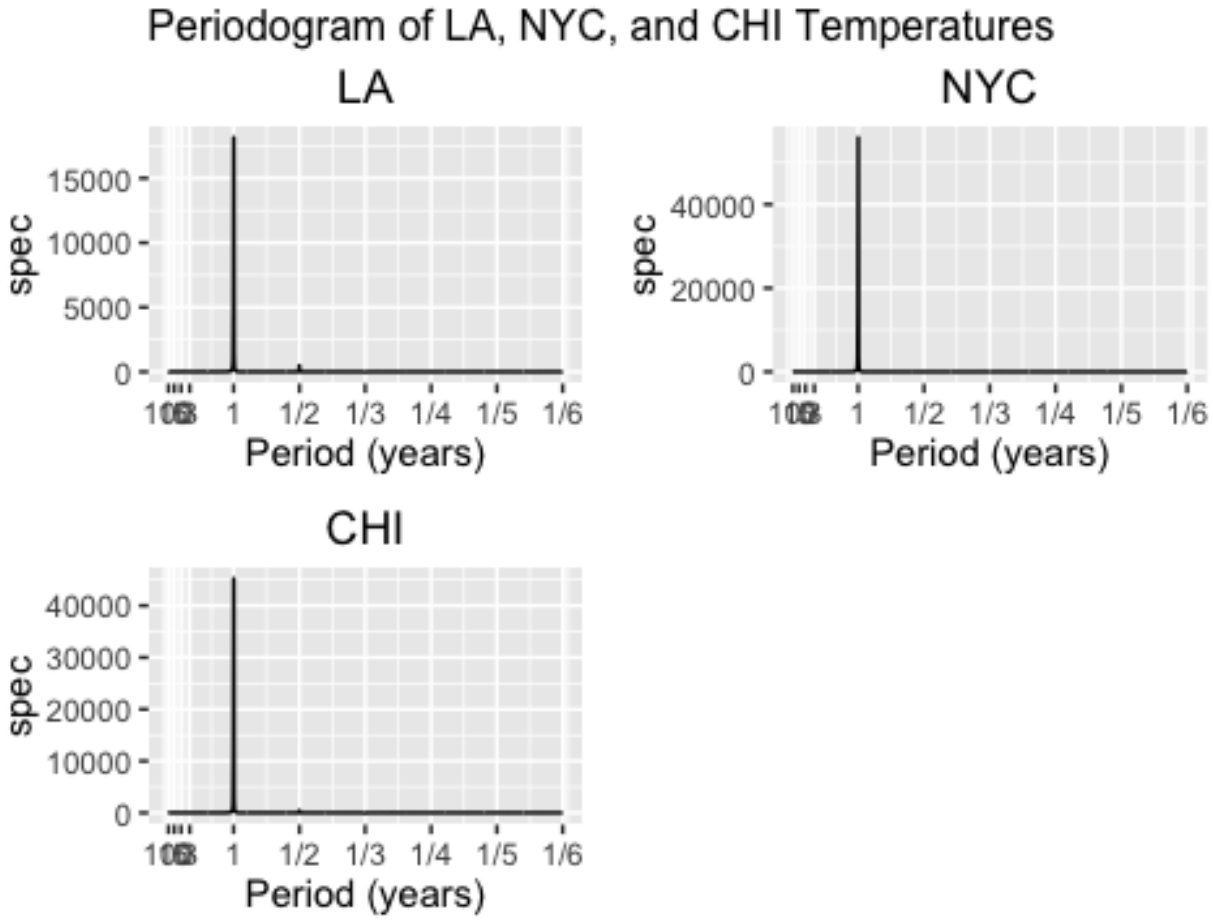
City	1/2 year	1 year	3.5 year	6 year	11 year	MSE on test set
LA	sin: <2e-16 *** cos: <2e-16 ***	<2e-16 *** <2e-16 ***	0.910 0.844	0.481 0.508	0.467 0.374	2.07559
NYC	sin: 0.320264 cos: 0.000711 ***	<2e-16 *** <2e-16 ***	0.538281 0.066742 .	0.157414 0.005573 **	0.024862 * 0.280236	4.29451
CHI	sin: <2e-16 *** cos: 2.13e-05 ***	<2e-16 *** <2e-16 ***	0.130227 0.020551 *	0.798048 0.030661 *	0.000785 *** 0.199943	5.162126

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

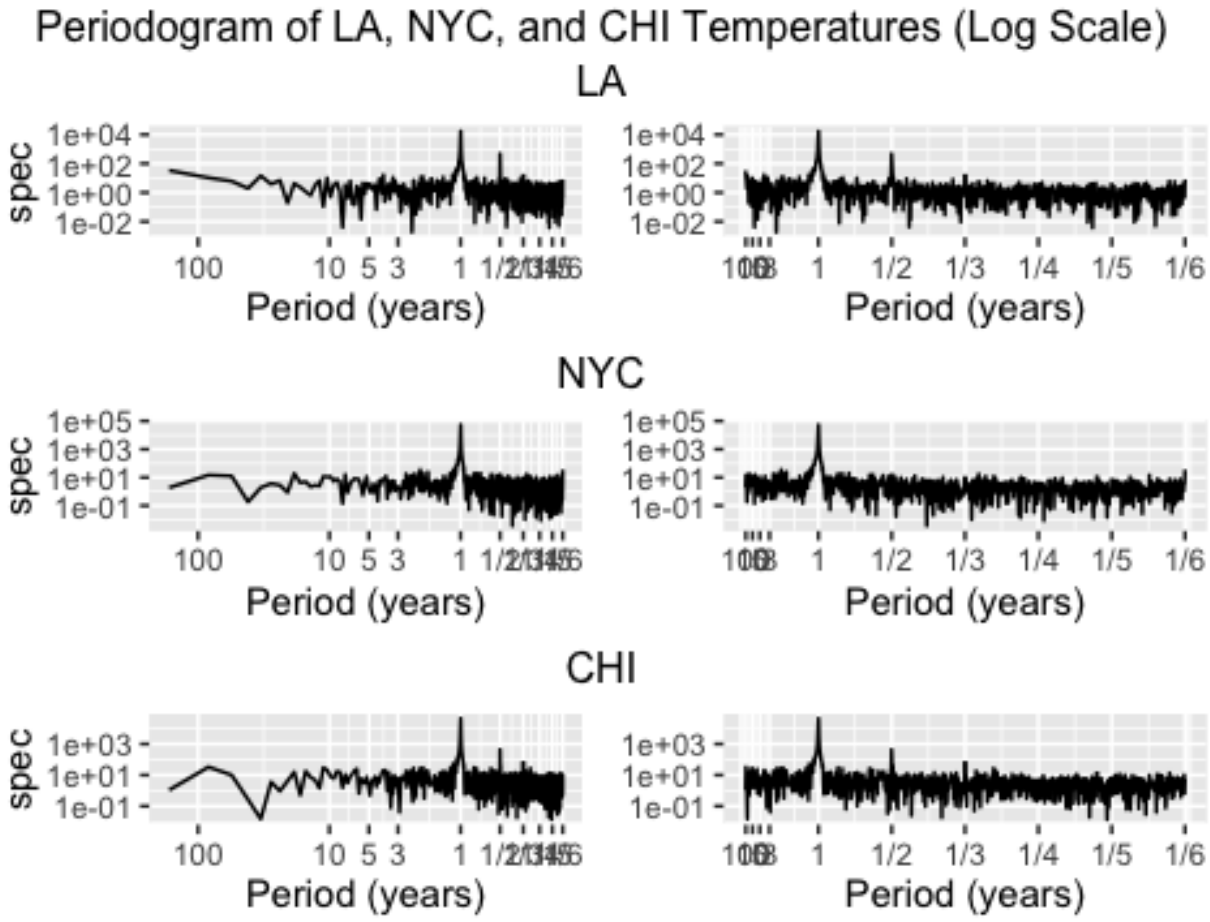
Figure 18: Harmonic Regression p-values and MSEs

6.8 Spectral Analysis

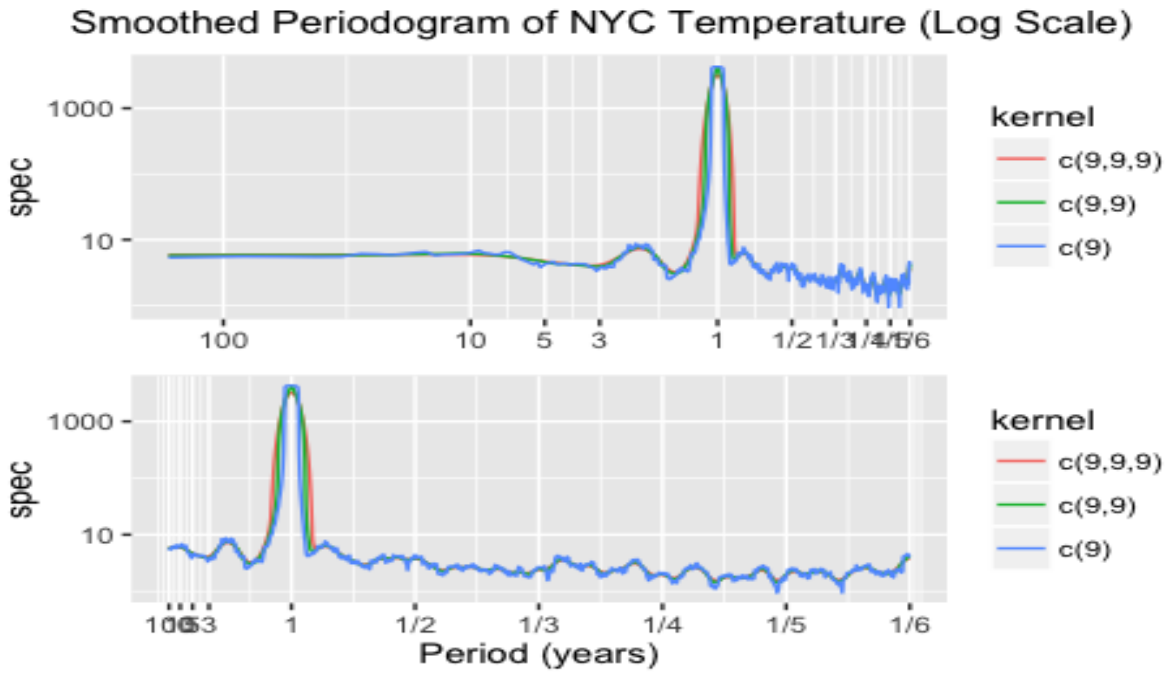
6.8.1 Periodograms

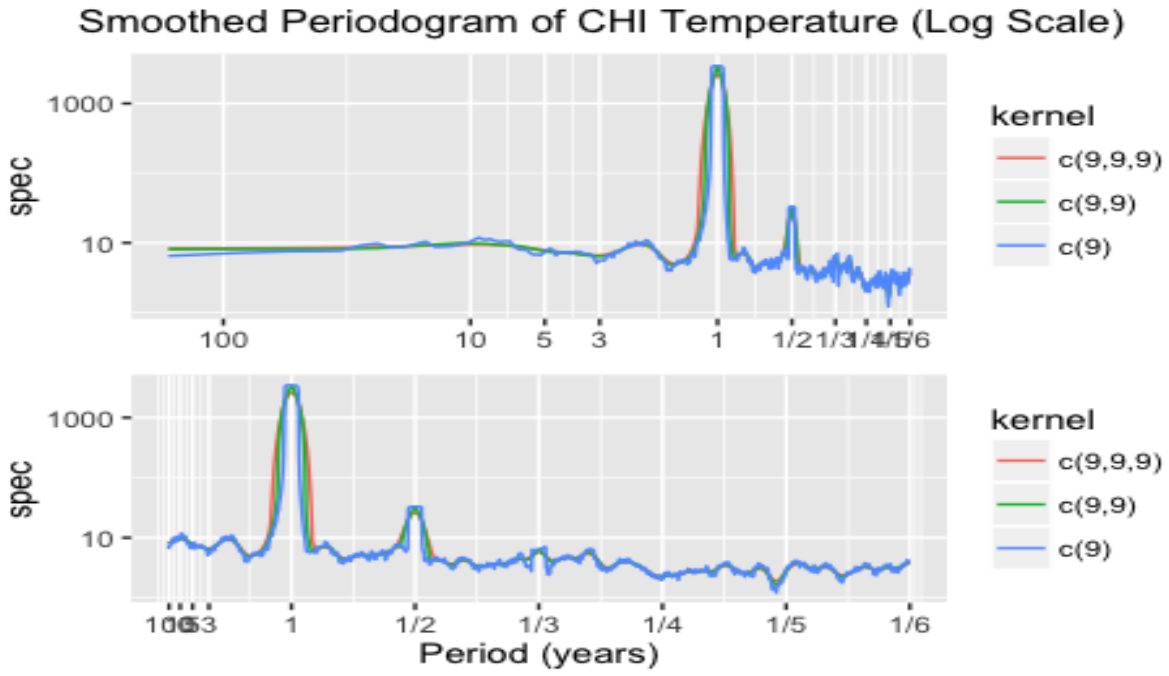


6.8.2 Periodograms (Log)



6.8.3 Smoothed Periodograms





6.8.4 Smoothed and Tapered Periodograms

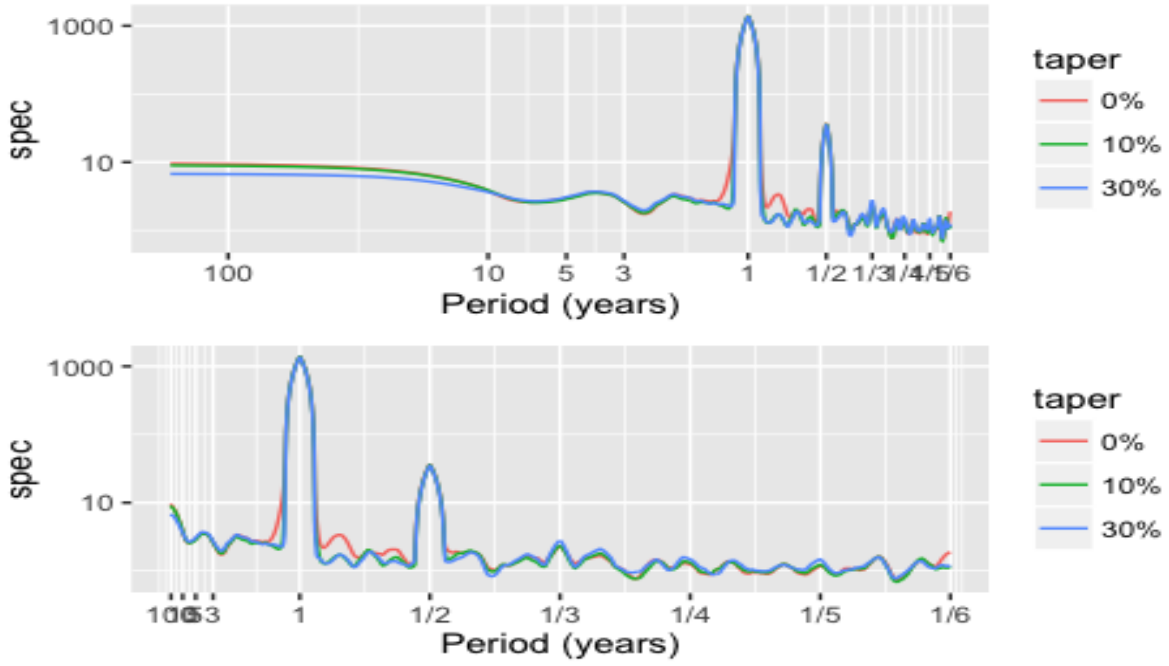


Figure 19: Smoothed and Tapered Periodogram of LA

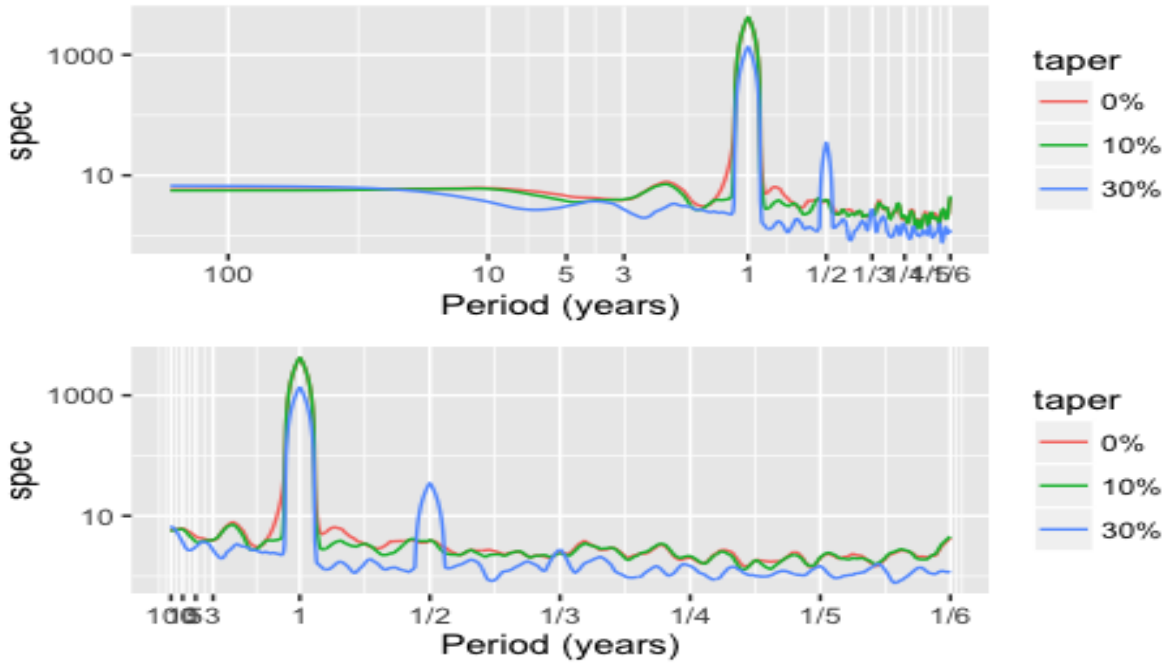


Figure 20: Smoothed and Tapered Periodogram of NYC

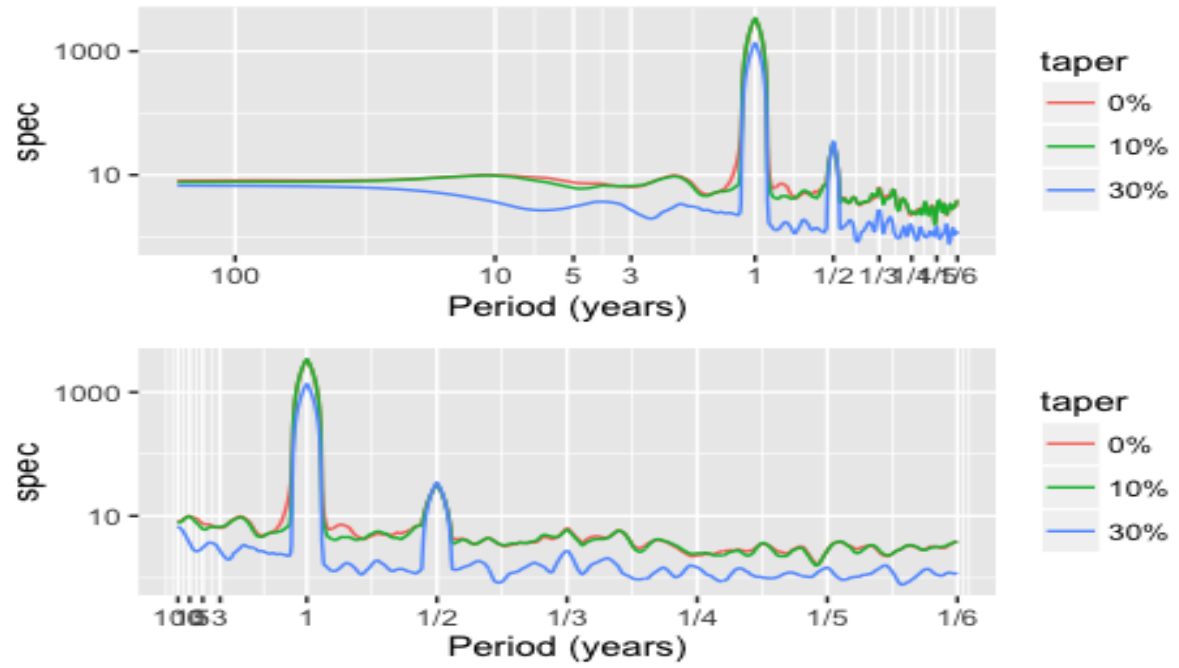


Figure 21: Smoothed and Tapered Periodogram of CHI

6.8.5 Spectral Spectrums

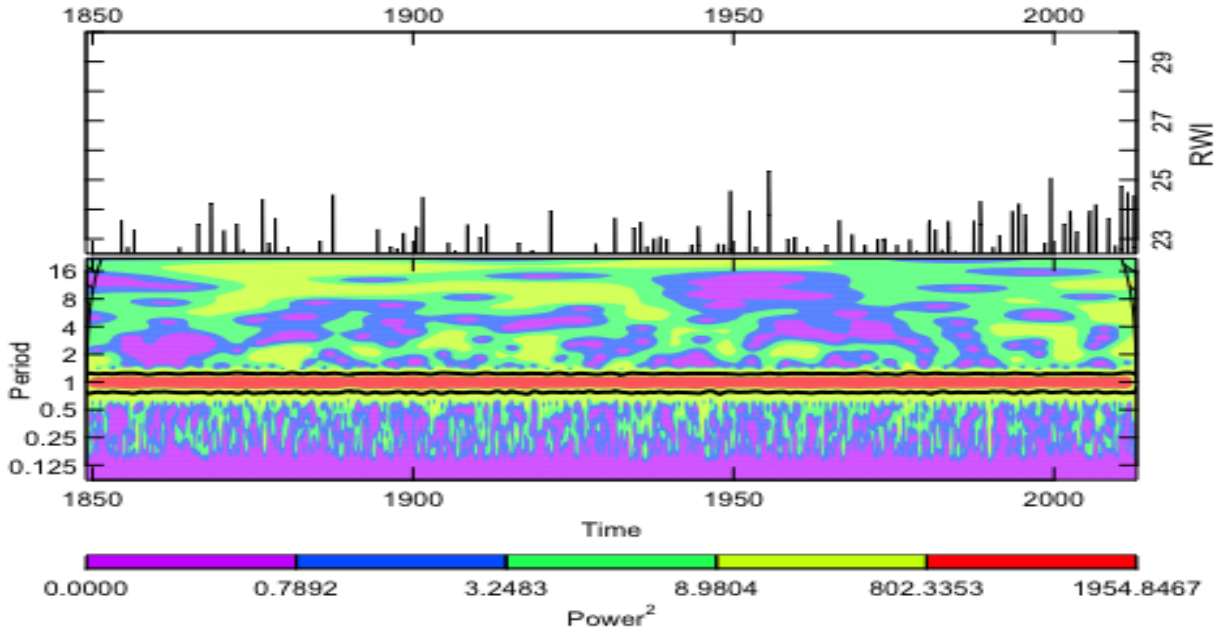


Figure 22: Spectral Spectrum of NYC

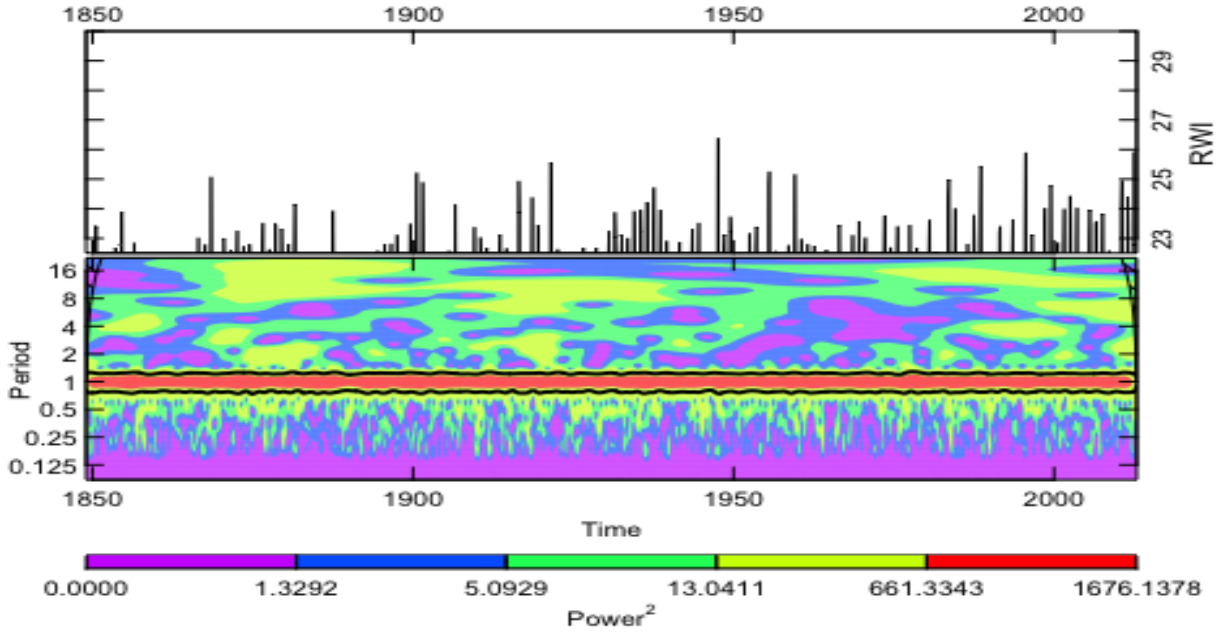


Figure 23: Spectral Spectrum of CHI