DTU02417, Time Series Analysis, Assignment 3
ARIMAX model for building data

Edward J. Xu (Jie Xu), s181238, DTU Management, edxu96@outlook.com

April 19th, 2019

Question 3.1: Data Visualization

The data about heating, external temperature and solar irradiation are plot as a function of time in the following three figures correspondingly. The training data and testing data are distinguished by different line style. Overall, there are seasonal fluctuations in all figures. External temperature and heating have obvious trends, while the trends of solar irradiation is not obvious. There are more detailed descriptions about every figure next to the figure.

Figure 1. There are fluctuations and increasing trend. Because it's from a stochastic process depending on other processes, it's essential to establish the model with other processes and predict it's value on their predictions.

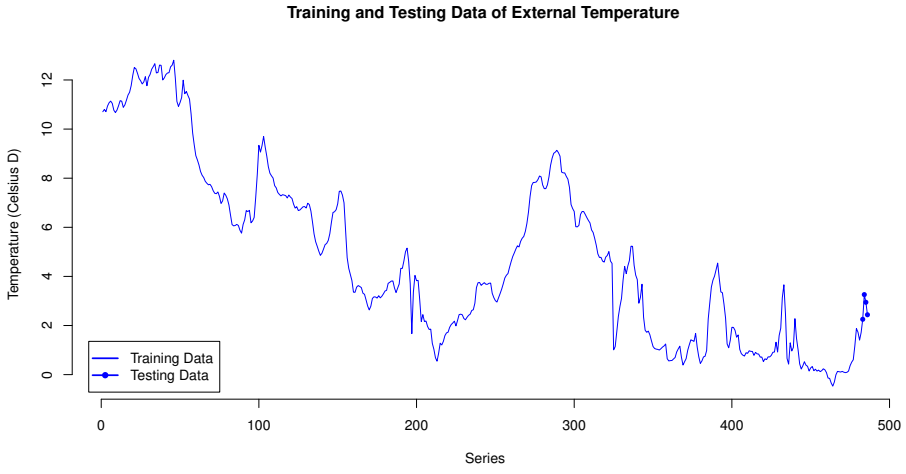

Figure 2. The downward trend is obvious, which can be described by linear regression. Although there is no visible curve in the trend, a quadratic trend model can be established to be compared to usual linear trend model. The fluctuations are unstable, there will be large variations in the validation and prediction.

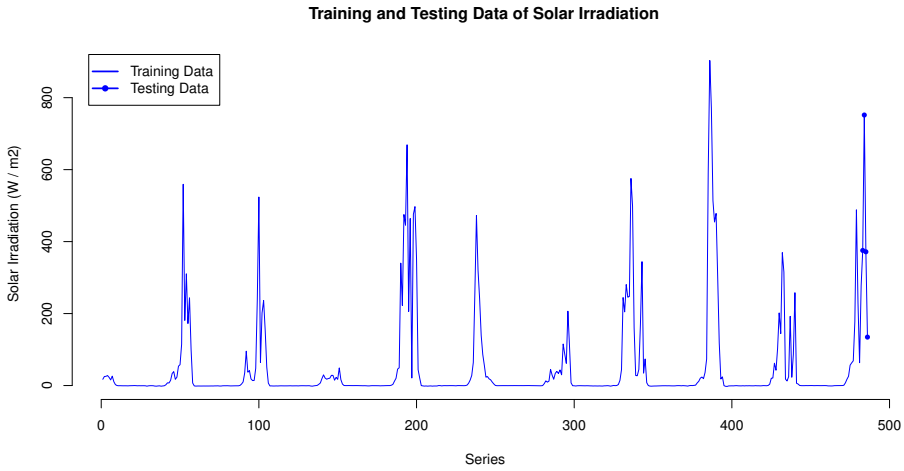

Figure 3. There are regular fluctuations, which can be explained by the regular solar movement. So it can be modeled by seasonal local trend model. Also, there is a bit increasing trend, which can be explained by the time when the data are obtained. The height angle of the sun is increasing, so the solar irradiation in the corresponding time during the day is increasing, with the exception of that during cloudy days.

Question 3.2: Correlation Structure

Correlation structures regarding auto-correlation and cross-correlation of the data of heating power are plotted.

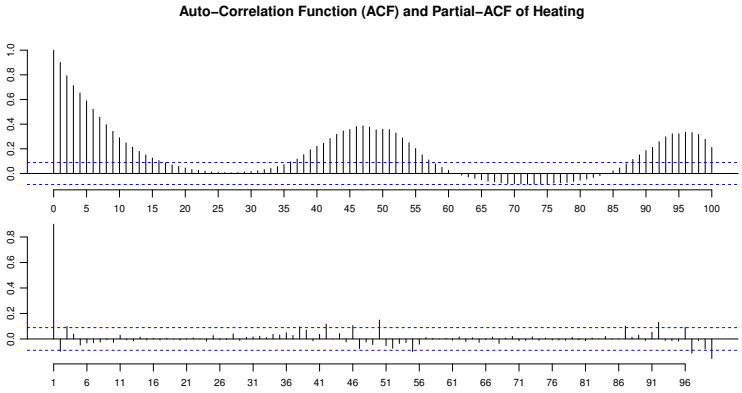

Figure 4. TS, ACF, PACF of Heating. The ACF does decreases to 0, but slowly. There are significant values around lag 36 - 57, with most significant ones around lag 48. Also, there are values around 96. There is nothing significant after lag 3 in PCF. The significance appears again around 48 and 96.

The lag k value returned by ccf(x, y) estimates the correlation between x[t+k] and y[t].

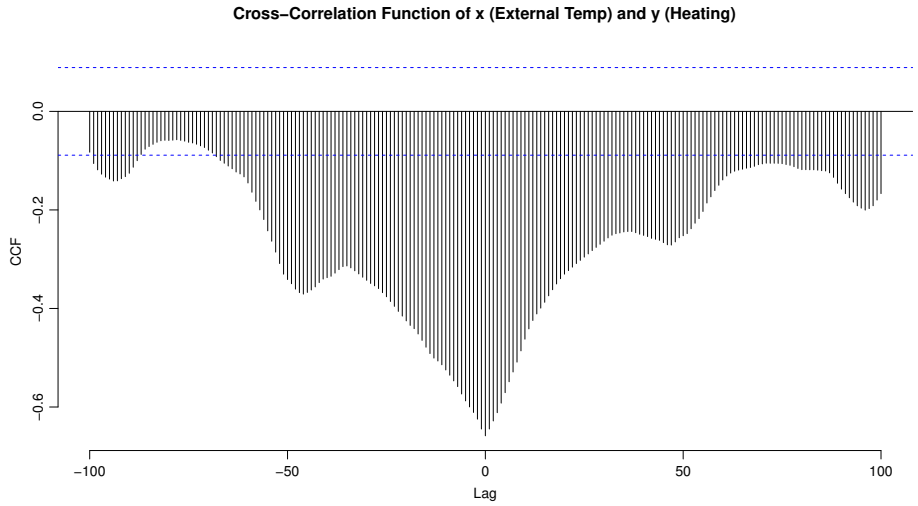

Figure 5. Cross-Correlation Function of External Temperature and Heating. CCF Being negative means there is a negative correlation between external temperature and heating, which proves the fact more heating is needed when the temperature is low. Overall, the absolute value of the function decreases exponentially, but there are increasing around lag 48 and 96. The increasing is soon offset by decreasing.

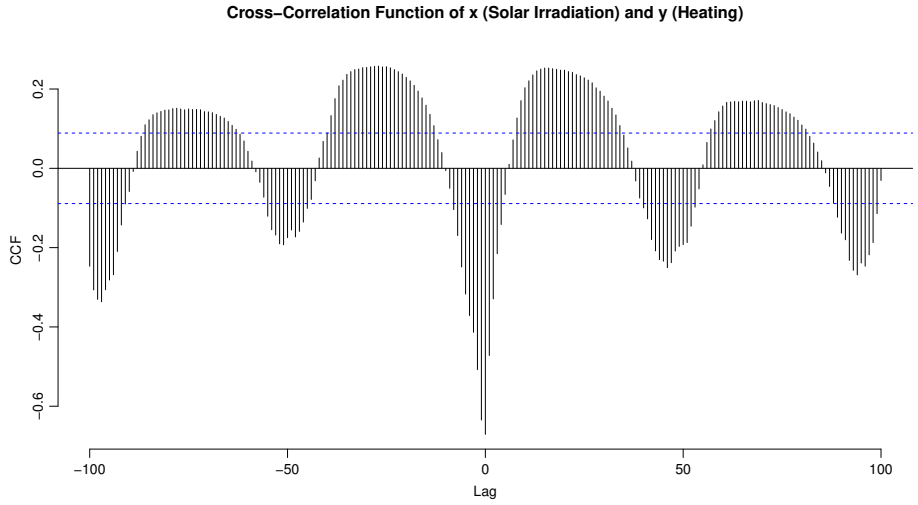

Figure 6. Cross-Correlation Function of Solar Irradiation and Heating. The function shows sine-fluctuation, being negative around lag 0, 48 and 96. The number of lags with positive values is higher than that with negative values. Compared to the plot of solar irradiation, the value being positive can be explained by the solar irradiation being zero during the night.

Question 3.3: ARIMA Model of Heating

0.1 Identification and Estimation

Name	Type	Coefficients	AIC	LogLikRatio Test	F-Dist Test
mod_3.1.4	ARIMA(0,1,2)(0,0,2)[48]	0, -0.1261, 0.0696, 0.1217	3316.468	1.0000000	NaN
mod_3.1.1	ARIMA(2,1,0)(0,0,2)[48]	0.0059, -0.1162, 0.0667, 0.12	3318.960	1.0000000	NaN
mod_3.1.2	ARIMA(2,1,0)(0,0,2)[48]	0, -0.116, 0.0665, 0.1195	3316.976	1.0000000	NaN
mod_3.1.3	ARIMA(0,1,2)(0,0,2)[48]	-0.0083, -0.1271, 0.0695, 0.1214	3318.435	0.8557635	NaN

Figure 7. Comparison of Different ARIMA Models of Heating

$$(1-\phi_1 B-\phi_2 B^2)(1-B)(1-\Phi_1 B^{48}-\Phi_2 B^{96})X_t=\varepsilon_t \tag{1}$$

$$\phi=[0,-0.2034] \tag{2}$$

$$\Phi=[0.0853,0.1465] \tag{3}$$

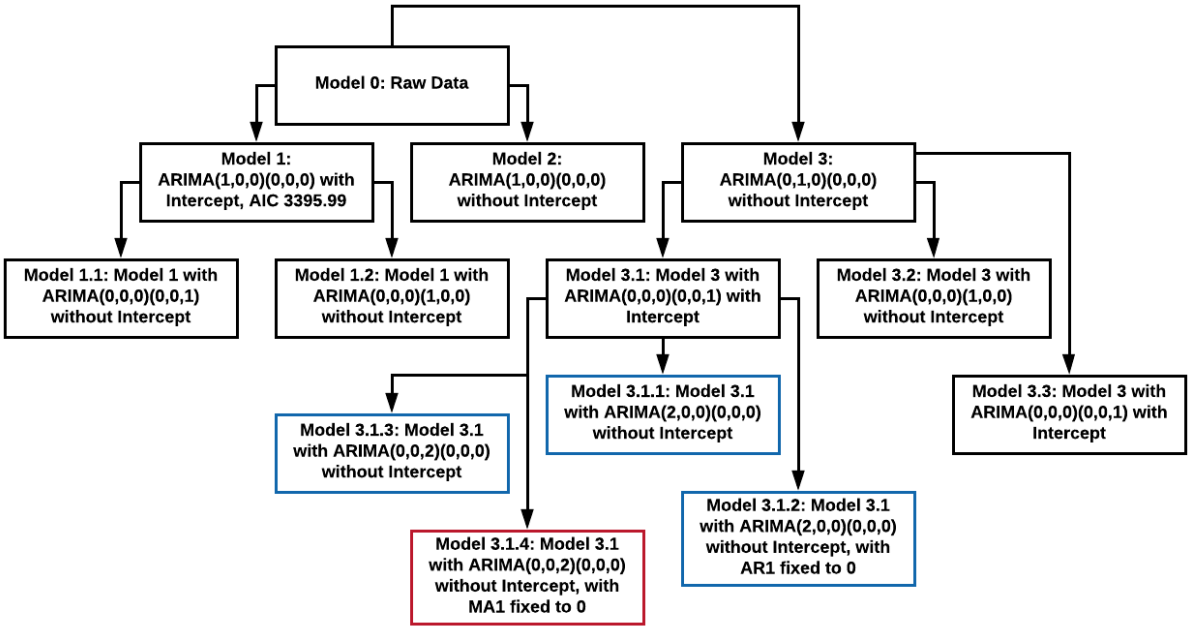

Figure 8. ARIMA Models of Heating.

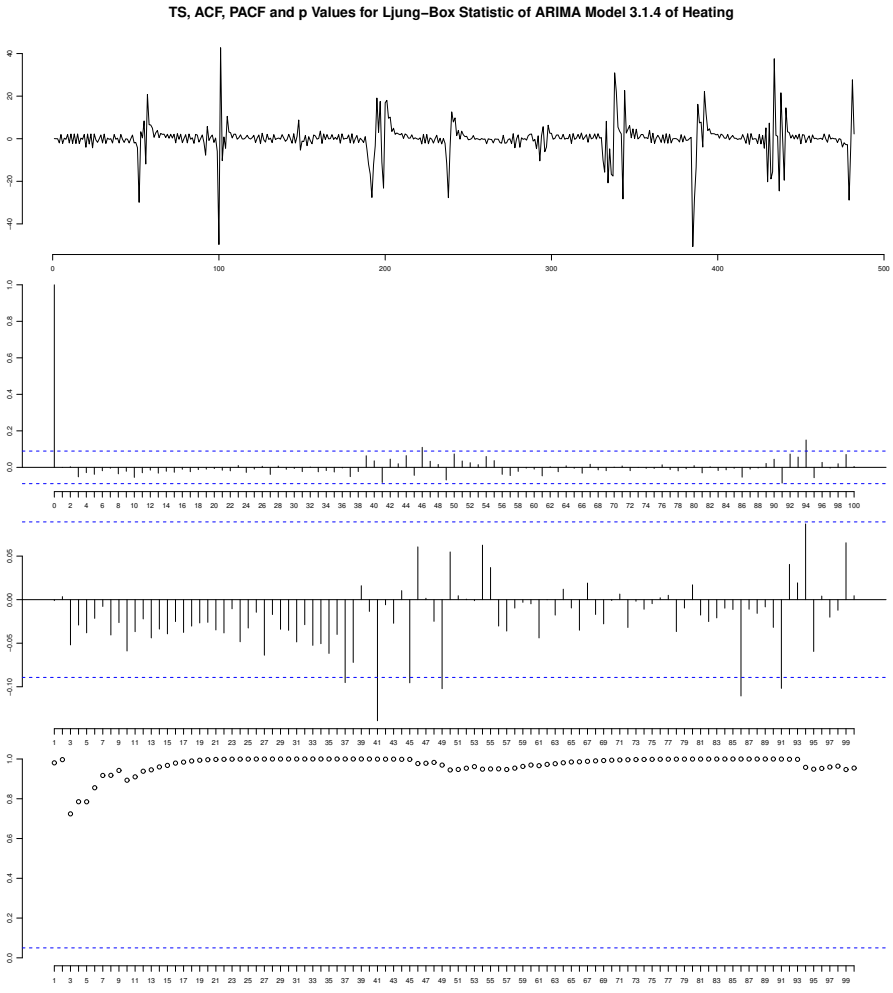

Figure 9. ARIMA (2, 1, 0)(0, 0, 2)48 Model 3.1.4 of Heating. There are still some correlation around lag 48 and lag 96.

0.2 Prediction

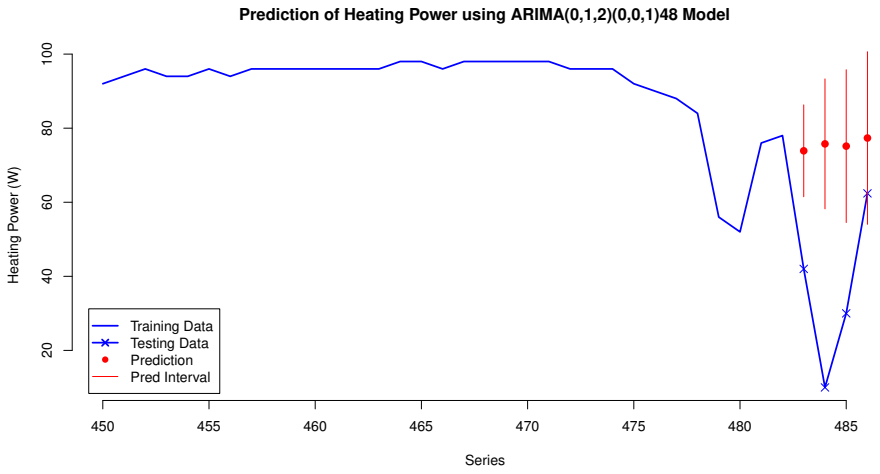

Figure 10. Prediction of Heating using ARIMA(0,1,2)(0,0,1)48 Model. The prediction result is not very satisfying.

Question 3.4: ARIMAX Model of Heating

0.3 ARIMA Model of External Temperature

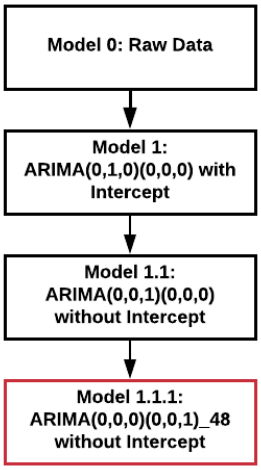

Figure 11. ARIMA Models of External Temperature. There is not too much difficult to find the ARIMA model for external temperature

$$(1-B)(X_{1,t}-\alpha) = (1+\theta_1 B)(1+\Theta_1 B^{48})\varepsilon_t \tag{4}$$

$$\alpha = 0 \tag{5}$$

$$\theta = 0.2820 \quad (\text{se} = 0.0419) \tag{6}$$

$$\Theta = 0.0588 \quad (\text{se} = 0.0451) \tag{7}$$

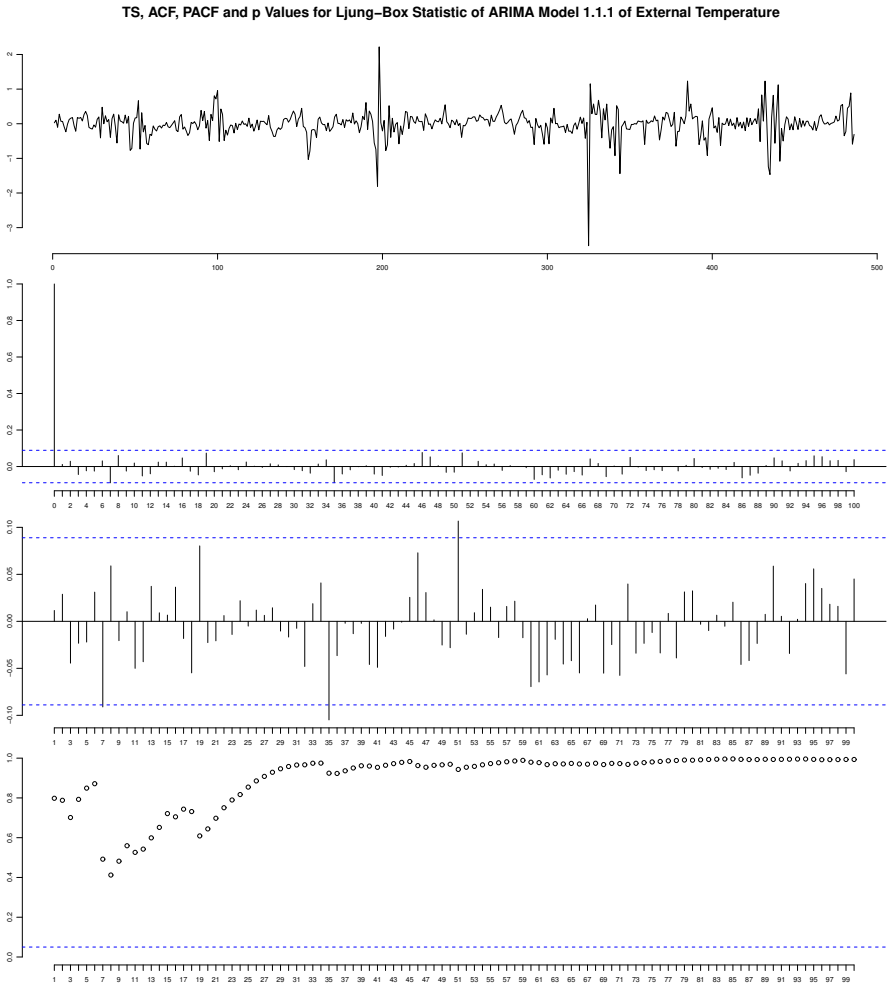

Figure 12. ARIMA Model (0, 1, 1)(0, 0, 1)48 1.1.1 of External Temperature

0.4 ARIMA Model of Solar Irradiation

Name	Type	Coefficients	AIC	LogLikRatio Test	F-Dist Test
od_1.1.2	ARIMA(4,1,0)(0,0,2)[48]	-0.2081, -0.0825, -0.2036, 0.035, -0.0255, 0.1262	5613.469	1	NaN
mod_1.1.1	ARIMA(4,1,0)(0,0,1)[48]	-0.2127, -0.0677, -0.1831, 0.0404, -4e-04	5618.287	1	NaN

Figure 13. Comparison of Different ARIMA Models of Solar Irradiation

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3-\phi_4 B^4)(X_{2,t}-\alpha)=(1+\Theta_1 B^{48}+\Theta_2 B^{96})\varepsilon_t \tag{8}$$

$$\alpha = 46.8750 \quad (\text{se} = 17.1655) \tag{9}$$

$$\phi = [0.6910, 0.1118, -0.1447, 0.1155] \quad (\text{se} = [0.0457, 0.0561, 0.0555, 0.0471]) \tag{10}$$

$$\Theta = [-0.0125, 0.1343] \quad (\text{se} = [0.04860.0476]) \tag{11}$$

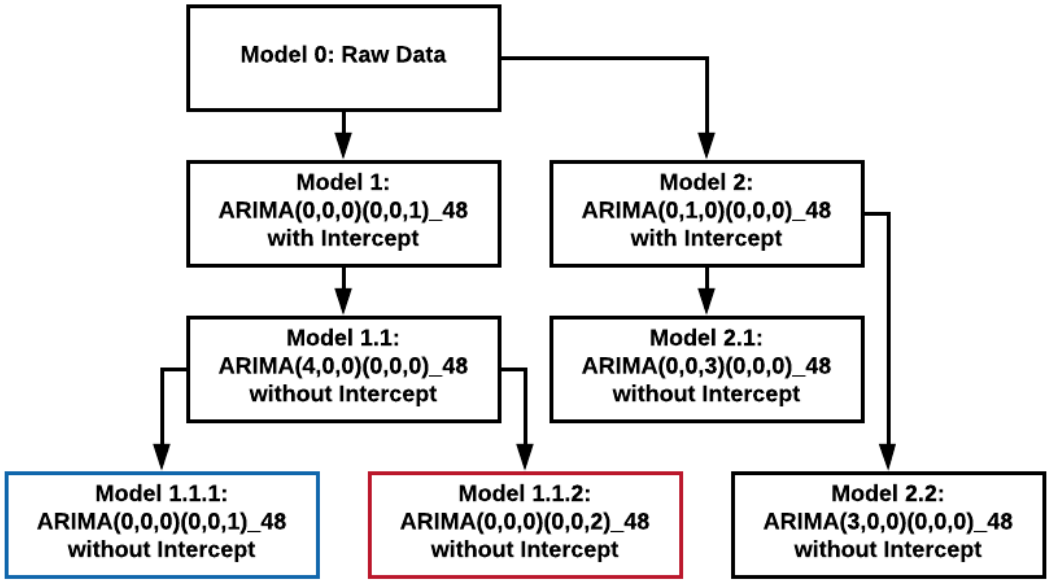

Figure 14. ARIMA Models of Solar Irradiation

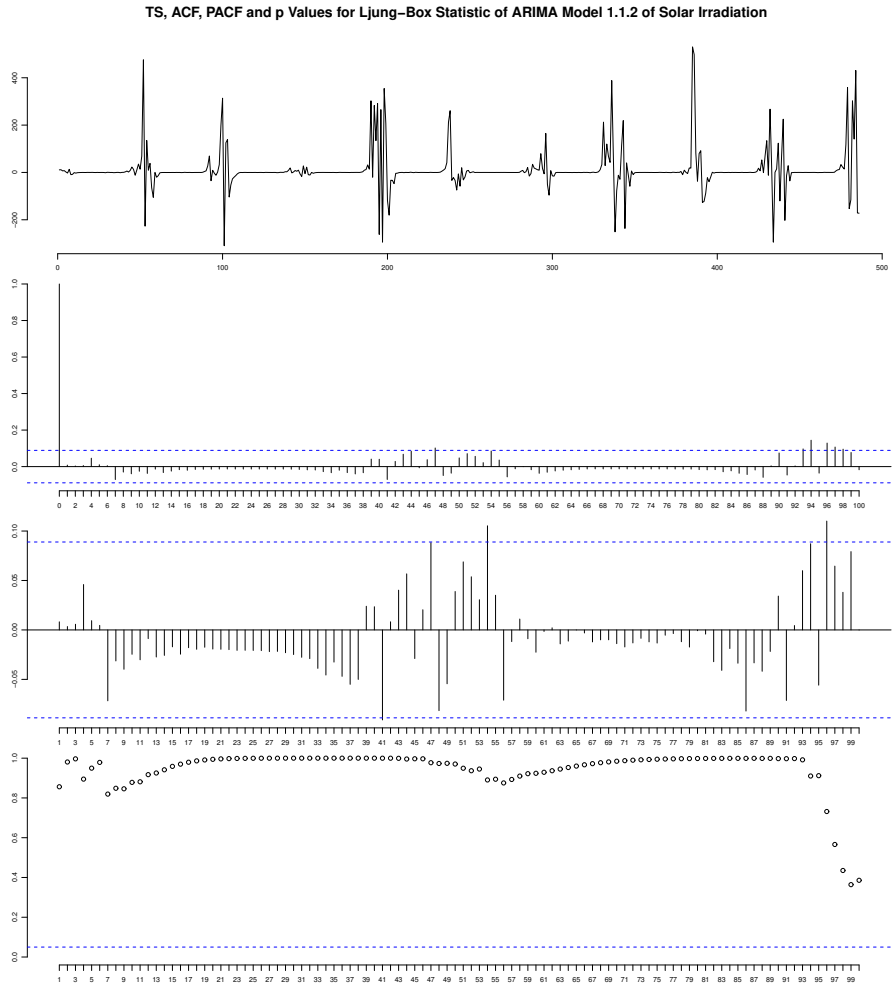

Figure 15. ARIMA Model (4, 0, 0)(0, 0, 2)48 1.1.2 of Solar Irradiation

0.5 Pre-whitening of External Temperature

It's difficult to to multi-variate pre-whitening. We can do uni-variate pre-whitening instead to provide indirect reference for choosing the right lag of eXogenous input.

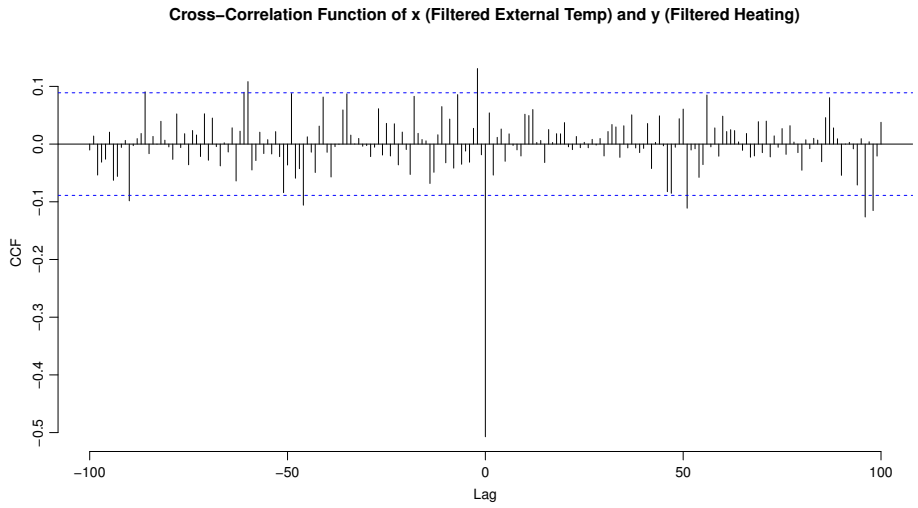

Figure 16. CrossCorrelation Function of x (Filtered External Temp) and y (Filtered Heating). Filtered by ARIMA Model (0, 1, 1)(0, 0, 1)48 1.1.1 of External Temperature. There is just some correlation in lag 0. So the furthest lag for ARIMAX modelling is lag -1.

0.6 Pre-whitening of Solar Irradiation

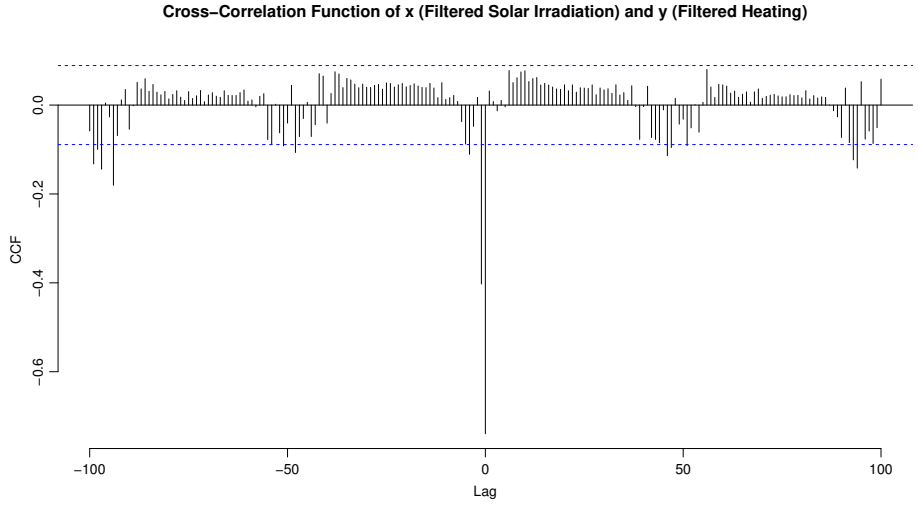

Figure 17. CrossCorrelation Function of x (Filtered Solar Irradiation) and y (Filtered Heating). Filtered by ARIMA Model (4, 0, 0)(0, 0, 2)48 1.1.2 of Solar Irradiation. There is some correlations in lag 0 and lag - 1. So the furthest lag for ARIMAX modelling is lag -2.

0.7 Prediction using Multiple-Input ARIMAX Model

When there are multiple-input in the system, the uni-variate pre-whitening cannot provide sufficient information about which lags should be modelled in the ARIMAX model. We have to try different combinations according to the pre-whitening result.

$$(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(Y_t-\alpha)=(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3+\theta_4 B^4)(1+\Phi_1 B^{48}+\Phi_1 B^{96})\varepsilon^t+\beta_1 X_{1,t}+(\beta_2+\beta_3 B)X_{2,t} \tag{12}$$

$$\phi=[-0.3030,0.2717,0.9089] \quad \text{se}=[0.1117,0.1127,0.0751] \tag{13}$$

$$\theta=[0.6855,0.3925,-0.2983,0.1985] \quad \text{se}=[0.1198,0.0998,0.0728,0.0500] \tag{14}$$

$$\Phi=[0.1568,0.2190] \quad \text{se}=[0.0477,0.0511] \tag{15}$$

$$\alpha=86.8396 \quad \text{se}=3.5250 \tag{16}$$

$$\beta=[-1.5537,-0.0677,-0.0373] \quad \text{se}=[0.4309,0.0019,0.0018] \tag{17}$$

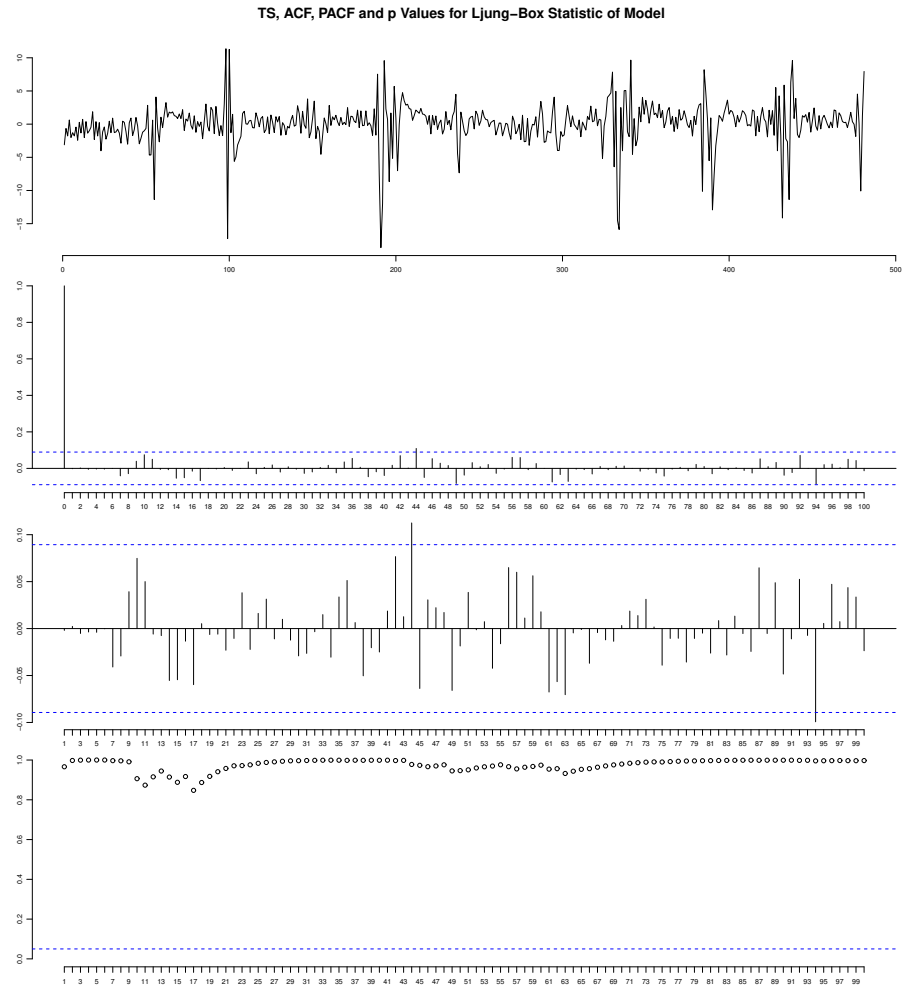

Figure 18. Residuals, ACF, PACF and p Values for LjungBox Statistic of ARIMAX Model 4 [(3, 0, 4) (0, 0, 2)48] with known external temperature and solar irradiation as eXogenous inputs.

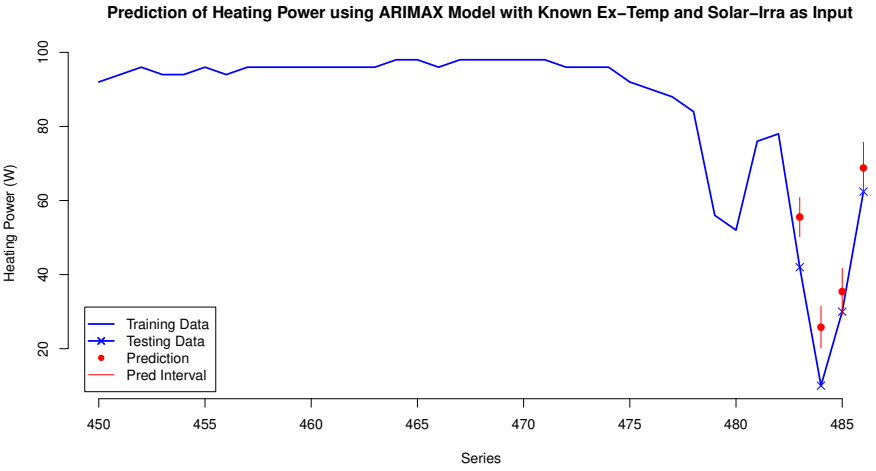

Figure 19. Prediction of Heating using ARIMAX Model with Known External Temperature and Solar Irradiation as Input.

Question 3.5: Comparison

It is harder to find a direct ARIMA model for heating. Besides, it takes more time to fit a model, because it needs lots of guesses and trial-and-error. Instead, to construct the ARIMA models for inputs first, then fit an ARIMAX model can save much time. The residuals from ARIMA models of inputs are better behaved. Furthermore, the residual from the ARIMAX model and the prediction results are quite satisfying. Also, the prediction intervals stay narrow all the time, while those from the ARIMA model widen significantly. All in all, it is not worthy of modeling a time series directly when there are eXogenous inputs.

Appendix: Code

```r
testLogLikRatio <- function(mod_ref, mod){
    # mod_ref$loglik - mod$loglik
    # don't use "1 - pchisq(-2 * ( fit1$loglik - fit2$loglik ), df = 1)", which will return 0 most of the time because
    # of numerical approximation
    value <- pchisq(- 2 * (mod_ref$loglik - mod$loglik), df = 1, lower.tail = FALSE)
    return(value)
}
```

Code 1. Function to Perform Likelihood Ratio Test of the Model Against Reference Model

```r
testFDist <- function(mod_ref, mod){
    sum_residual_ref <- sum(mod_ref$residuals^2)
    sum_residual <- sum(mod$residuals^2)
    num_para_ref <- length(mod_ref$coef)
    num_para <- length(mod$coef)
    stat_f <- (sum_residual_ref - sum_residual) / (num_para - num_para_ref) /
        (sum_residual / (length(mod_ref$residuals) - num_para)) # If num_para = num_para_ref, it will be Inf
    value <- pf(stat_f, df1 = num_para - num_para_ref, df2 = (length(mod_ref$residuals) - num_para),
        lower.tail = FALSE)
    return(value)
}
```

Code 2. Function to Perform F-Test of the Model Against Reference Model

```r
calIntervalPred <- function(n, p, y.hat, se, prob = 0.95){
    quantileStudentDist <- qt(p = 0.95, df = n - p)
    boundUp <- y.hat + quantileStudentDist * se
    boundLow <- y.hat - quantileStudentDist * se
    return(list(boundUp = drop(boundUp), boundLow = drop(boundLow)))
}
```

Code 3. Function to Calculate Prediction Interval from Standard Error