Forecasting Hourly Air Temperature

STAT443

Kanika Chopra, Nidhi Kumar, Divya Chandrashekar, Samka Marfua, Amanda Du
Group 1

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

AGENDA

- Introduction
- Data
- Regression
- Smoothing
- Box-Jenkins
- Conclusions

INTRODUCTION AND MOTIVATION

- **Goal**: Future prediction of air temperature for upto 1 year past the time series time frame
  - Project Plan:
    - Clean
    - Explore
    - Test prediction methods
    - Compare findings
    - Recommend the best predictor

DATA - Description

- It’s a weather time series Data recorded by the Max Planck Institute for Biogeochemistry

- **Features:** *Day, Month, Year, Time, Hour, Pascal SI (mbar), Temperature (Celcius), Temperature (Kelvin), Dew Point, Relative Humidity, Saturation Vapor Pressure, Vapor Pressure, Specific Humidity, Water Vapor Concentration, Airtight, Wind Speed, Maximum Wind Speed, and Wind Direction (degrees).*

- Dataset contains 70091 observations and it’s hourly data starting from 2009 to 2016.

DATA - Cleaning

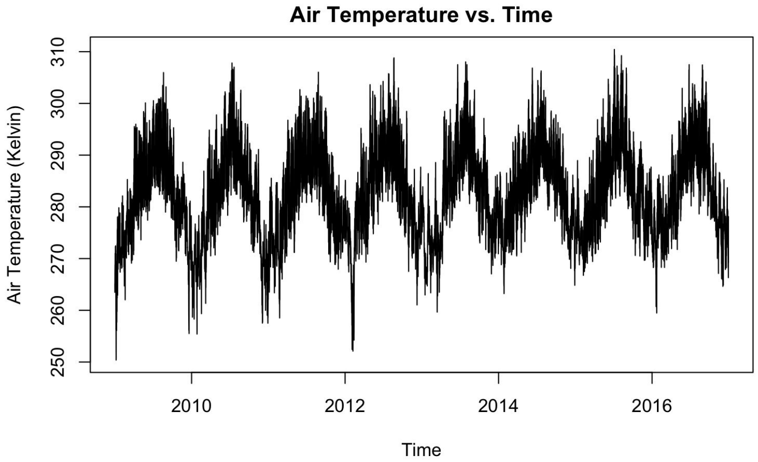
Holt-winters: Temperature (Kelvin) is preferred

Regression: All variables used for variable selection

Smoothing: Only time and Temperature

DATA - Exploration

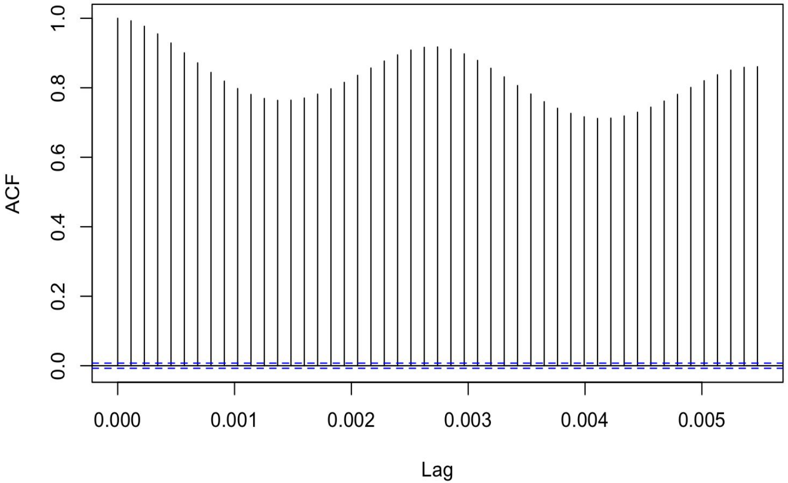
Time-series plot (Air temperature vs. time)



- Constant mean
- Seasonality

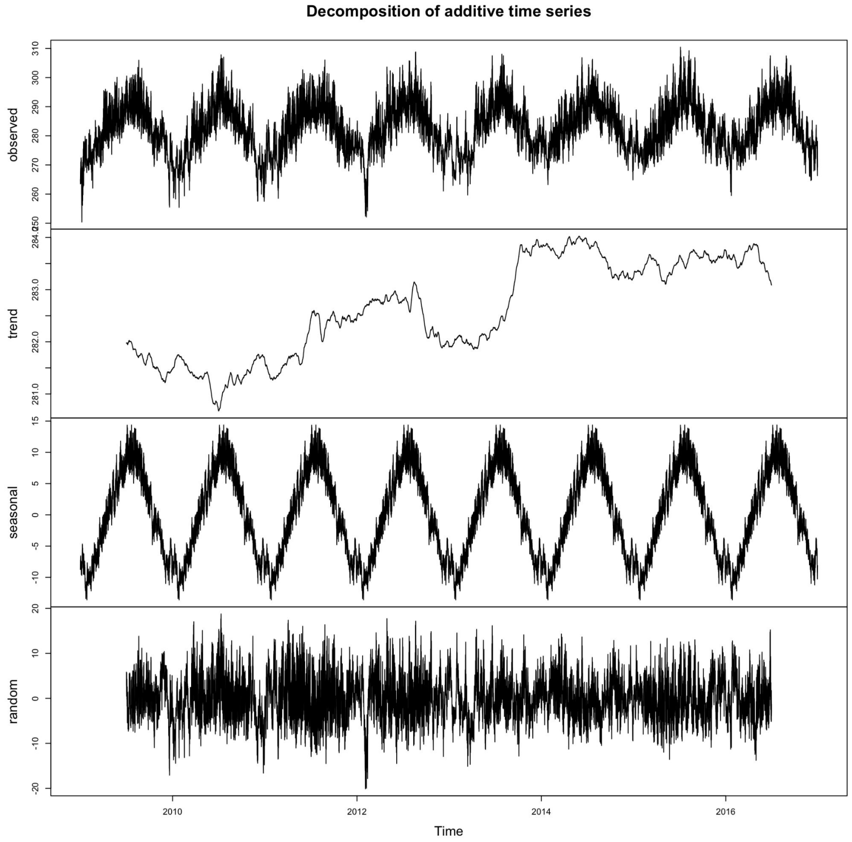
→ non-stationarity

ACF on Air temperature → seasonality



UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

DATA - Exploration (Continued)



- Increasing trend
- Seasonality

DATA-Training/Test Set Split

- **Prior to regression and smoothing**
- **No changing points**
- **Reserve seasonality**
- **Goal: [1 year] forecasting**

❖ **Training set**: January 1, 2009 - December 31, 2015 (61325 observations)

❖ **Test set**: January 1, 2016 to December 31, 2016 (8766 observations)

REGRESSION

- Response as a linear combination of the time and non-time variates
- Two-step process:
  - Variable selection of the non-time variates
  - Combinations of different time increments

Which method of variable selection?

3 options:

- Classic Selection
- LASSO
- Forward

VARIABLE SELECTION - CONT.

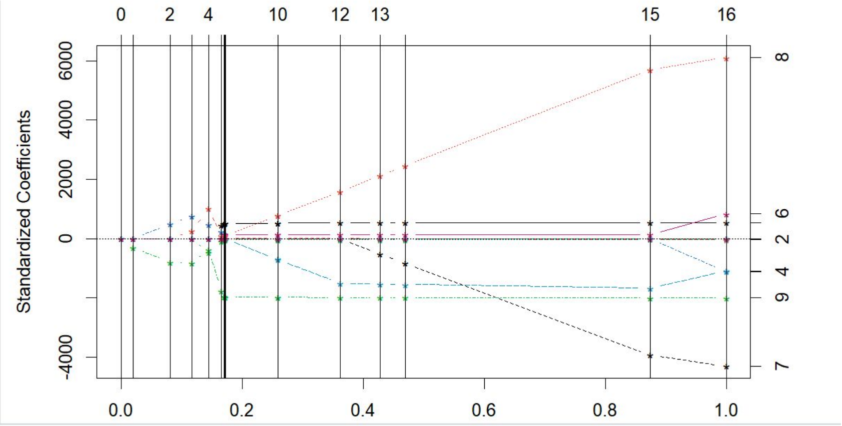
Classic:

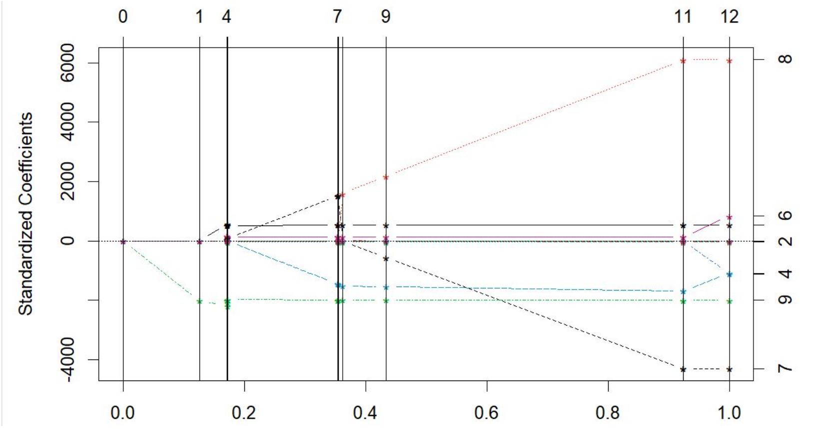
1. Least squares fitting, removing/adding variable at each step
2. Unconstrained
3. 9 variables

LASSO:

1. Least squares with lambda constraint
2. Minimizes parameters with some hitting zero
3. 6 variables

LASSO VS FORWARD - LAR





VARIABLE SELECTION - CONCLUSION

Priorities: (tradeoff)

- Fewer parameters - LASSO/Forward
- Lower MSE - Classic

Final pick: Forward

- Lower MSE than LASSO
- Fewer parameters than Classic

TIME INCLUSION

Questions:

1. Which level to use: Hour, Day, Month, Year
2. Linear relationship or higher order e.g. time^2, time^3, etc.

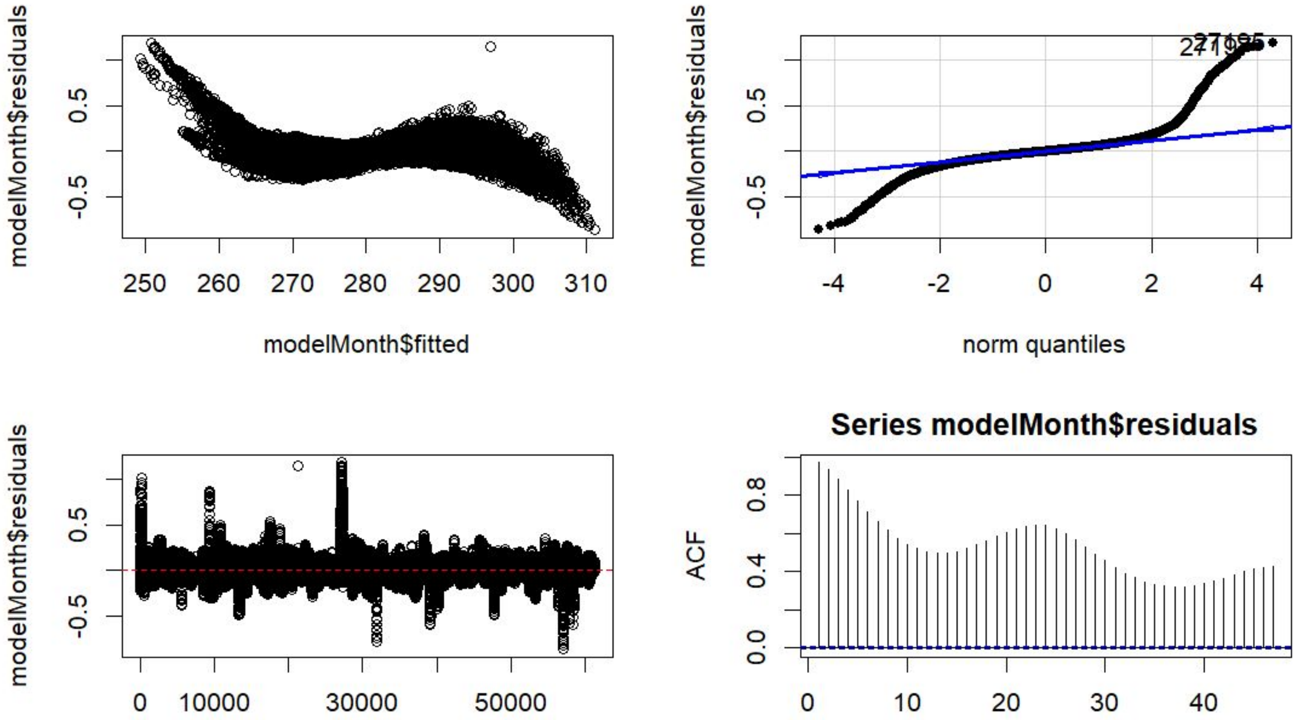
Testing methodology:

- Fit several combinations of LM models with our variates from previous selection on training set
- Compare prediction MSEs from testing set

TIME INCLUSION - RESULTS

Model	MSE
Hour-only	0.005885588
Day-only	0.005916561
Month-only	0.005508379
Year-only	1.647803
All at degree 1	0.005853096
All at best degree	1.809792

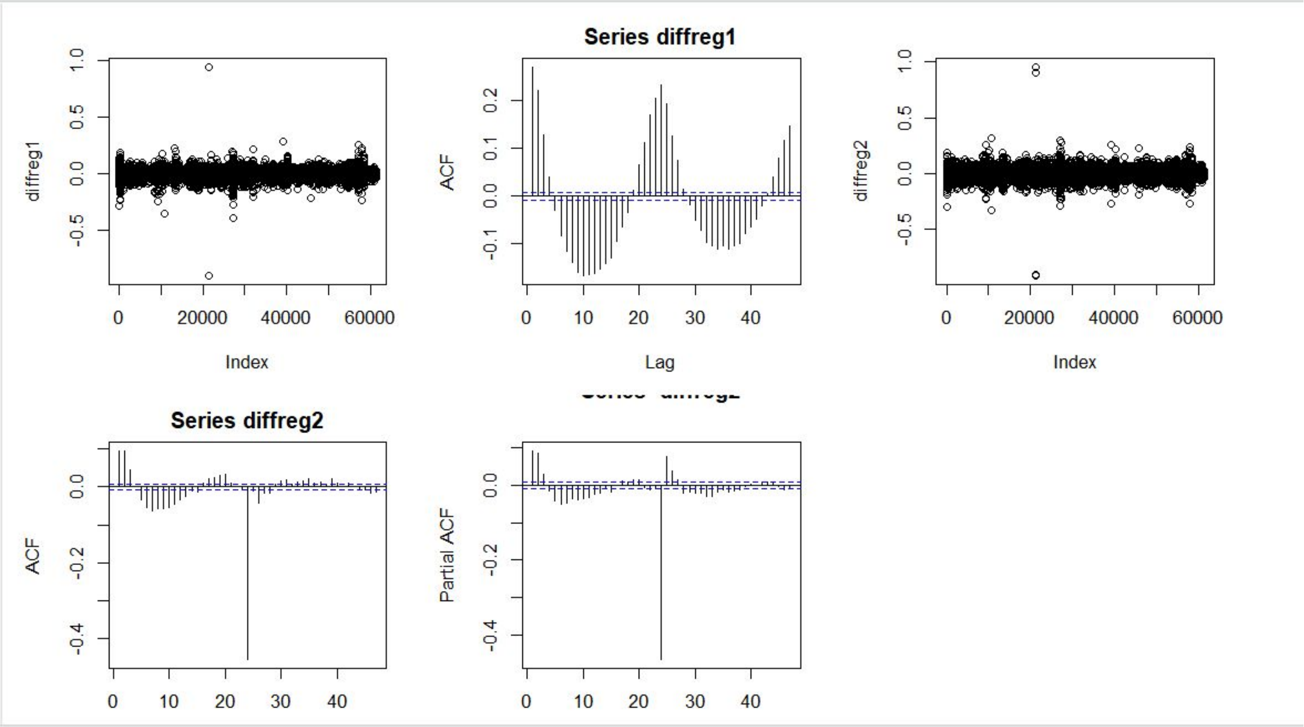
FINAL REGRESSION MODEL



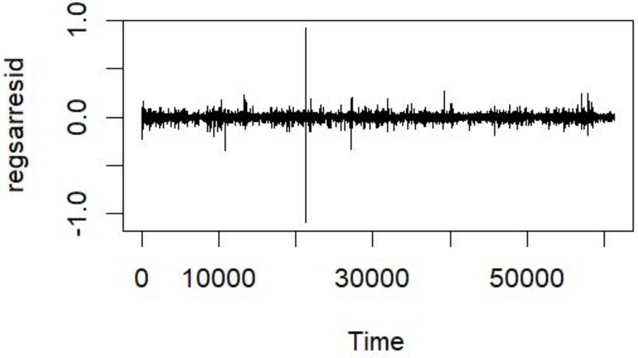
REGRESSION + BOX JENKINS

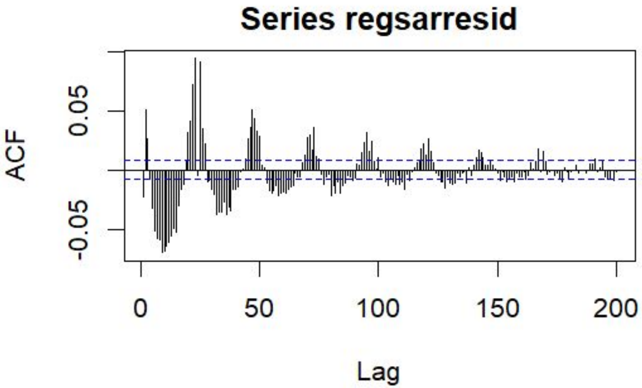
- ACF shows clear trend and seasonality
- Differencing shows lag 1 and lag 24 applicable
- We try SARIMA on the residuals

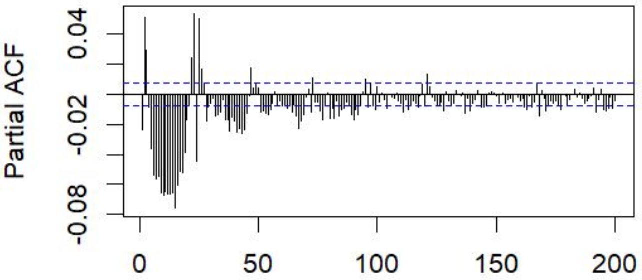
DIFFERENCING TO JUSTIFY SARIMA



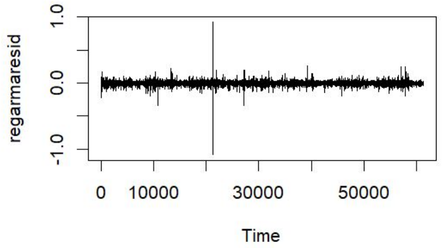
RESIDUALS OF FITTED SARIMA MODEL - ARMA PROPOSED

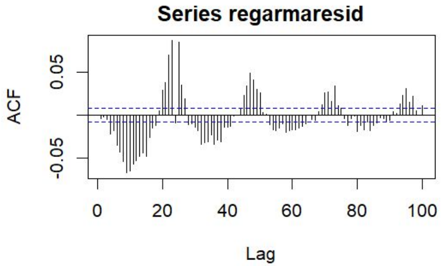


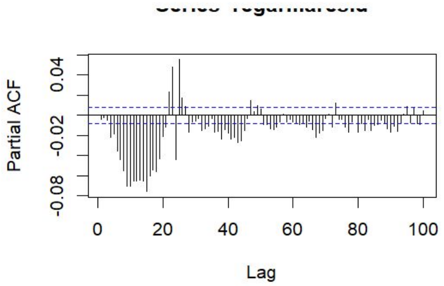




ARMA - CONTINUED





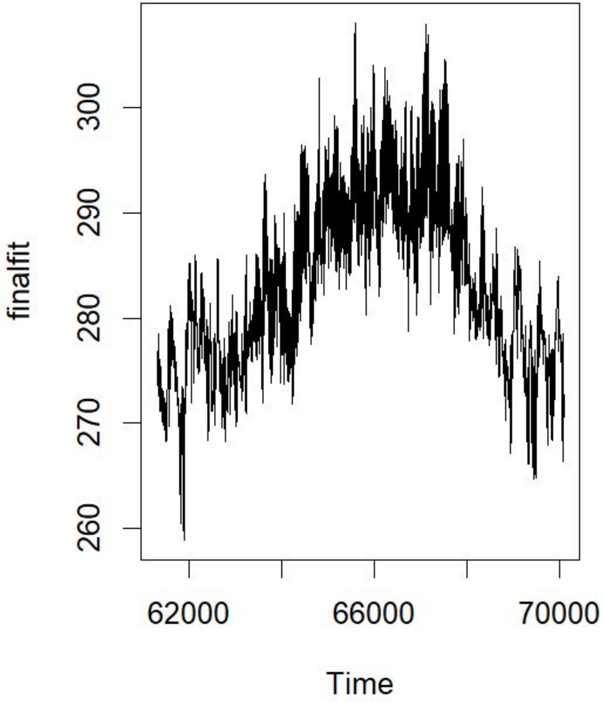


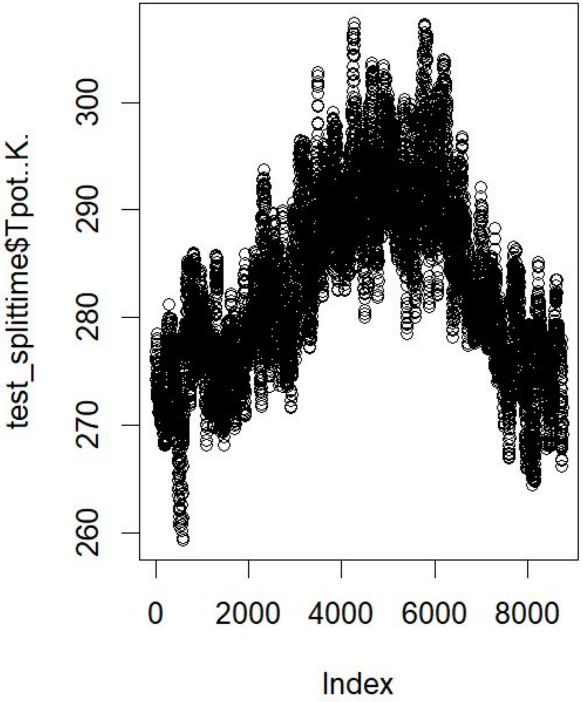
- Different ARMA models were applied on the SARIMA residuals
- Combinations: p = 1,2,3; q = 1,2,3
- Did not improve ACF/PACF
- Conclusion: Stick to Regression + SARIMA

PREDICTION

- ARIMAX for Regression with SARIMA residuals proved computationally difficult
- Method: Predict on regression model + SARIMA's forecast on regression residuals from training set, forecasted ahead a year
- Fit looked better, but MSE still proved higher than just regression

REGRESSION + SARIMA

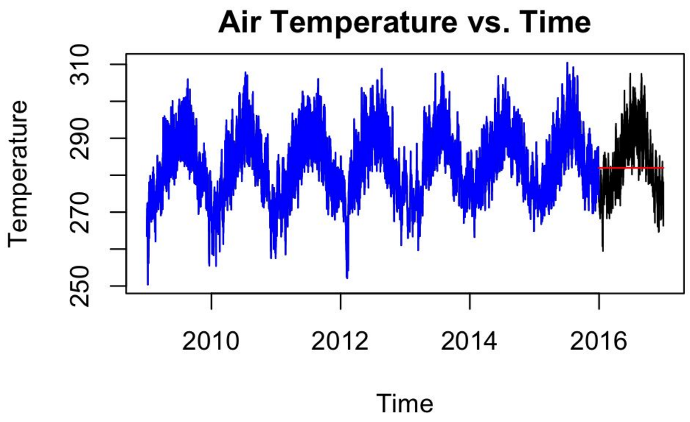




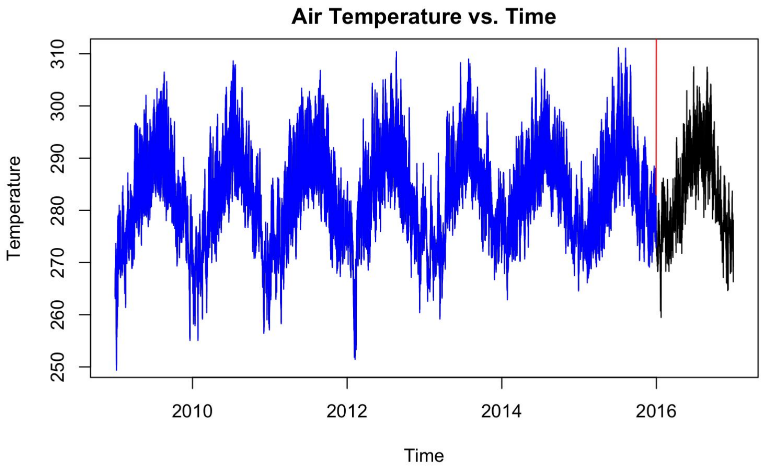
CONCLUSION

For our dataset, just regressing on time and other variates was sufficient, despite leaving information in the residuals.

SMOOTHING METHODS



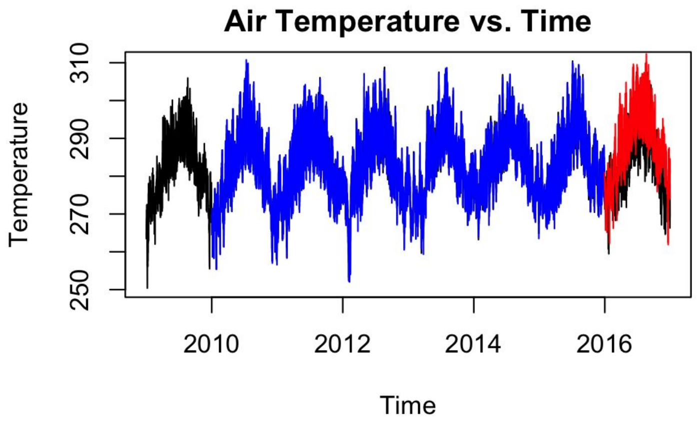
Exponential Smoothing



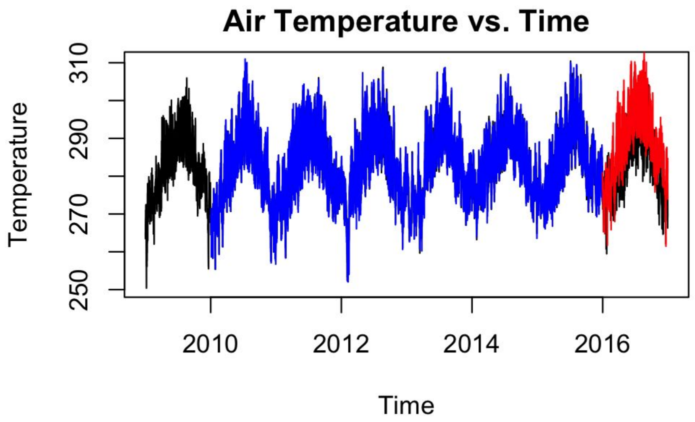
Double Exponential Smoothing

- The blue line represents the fit
- The red line represents the prediction for 2016

SMOOTHING METHODS



Additive Holt-Winters



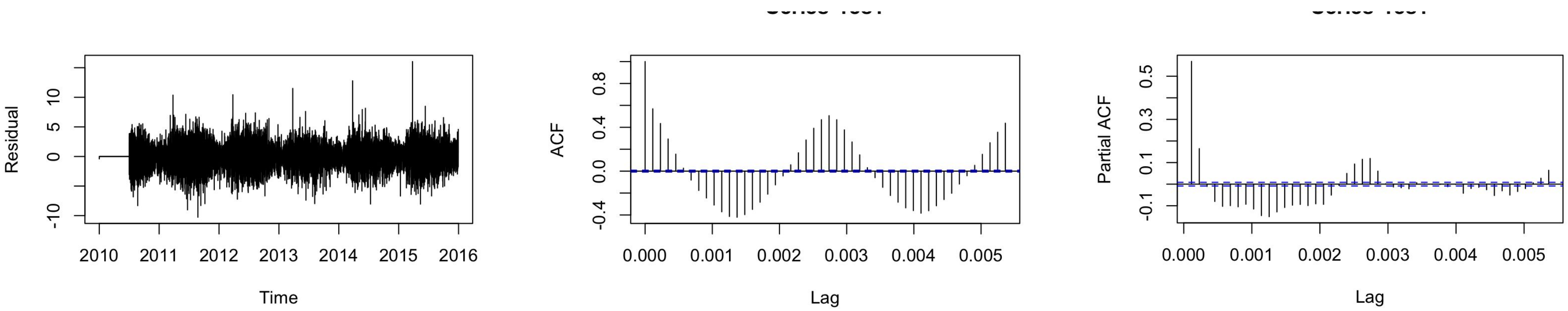
Multiplicative Holt-Winters

- The blue line represents the fit
- The red line represents the prediction for 2016

SMOOTHING METHODS COMPARISON

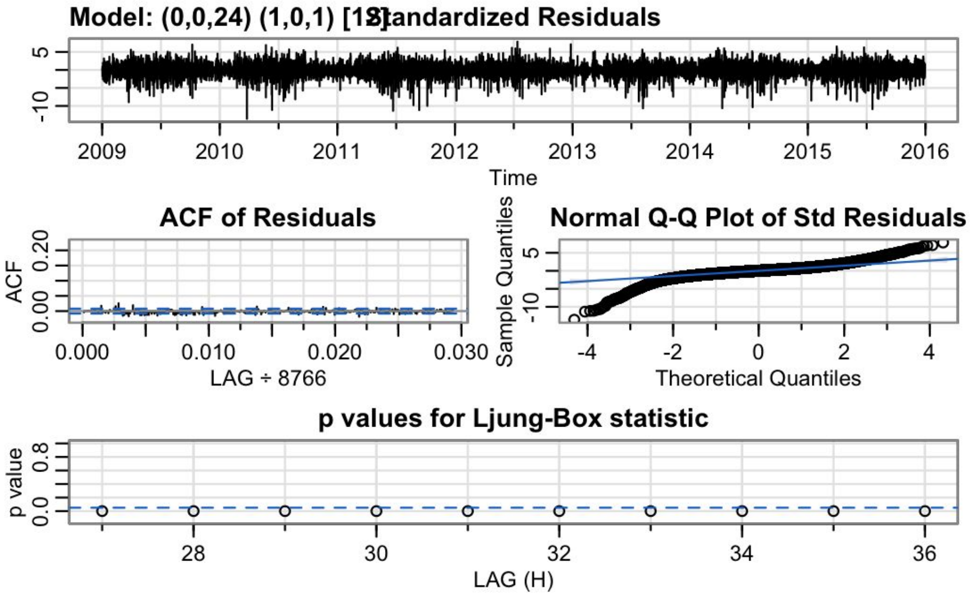
	Simple Exponential	Double Exponential	Additive Holt-Winters	Multiplicative Holt-Winters
Prediction MSE	68.97593	25922282	66.56564	69.65569

Additive Holt-Winters

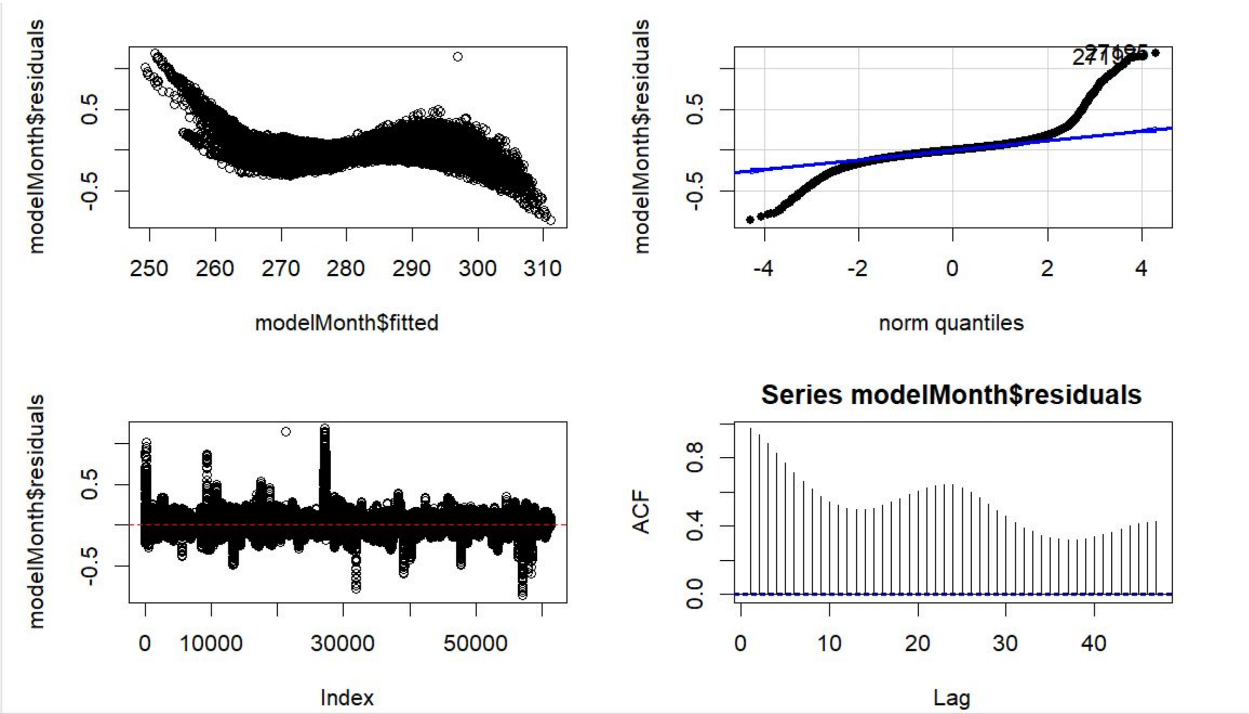


BOX-JENKINS

- Two times SARIMA model applied
- **SARIMA 1:** one-time differencing with lag 24
- **SARIMA 2:** seasonal differencing with monthly seasonality



STATISTICAL CONCLUSION



FINAL CONCLUSION

Chosen model:

(poly(Month) + airtight + atmospheric pressure + saturation vapor pressure + humidity + relative humidity + wind direction)

UNIVERSITY OF WATERLOO

FACULTY OF MATHEMATICS

YOU+WATERLOO

Our greatest impact happens together.