# Weather Forecasting

## Time Series Analysis and Models Final Project – Fall 2023

Edison M. Murairi

The George Washington University and University of Maryland College Park

December 12, 2023

# Table of Contents

# Introduction

## Problem Statement

Predict the temperature of a region based on physical indicators.

- City: Jena in Germany
- Area: 44.31 sq mi (114.76 km$^2$)
- Population: 110,502

## Dataset

Weather information collected every 10 minutes between January 1st, 2009 and January 1st 2016

- 12 Numerical variables: Pressure, Temperature relative to humidity, relative humidity, saturation vapor pressure, vapor pressure, vapor pressure deficit, specific humidity, water vapor concentration, airtight, wind speed, maximum wind speed, wind direction
- Categorical variables: None
- Downsample data (every 12 hours): 5839 observations.

# Table of Contents

# Temperature Time Series Plot

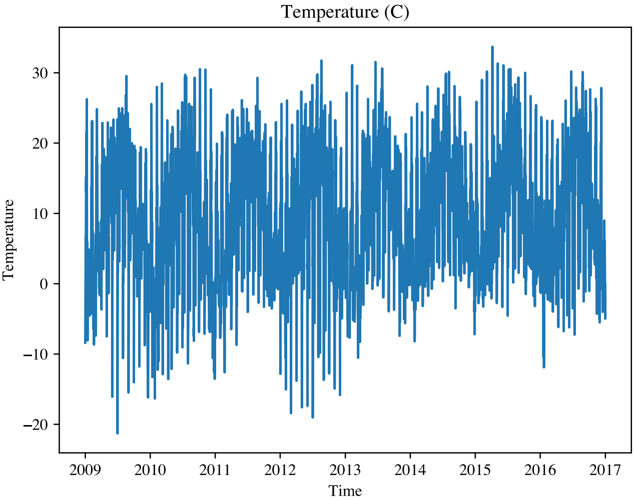



Figure: Raw data

# Temperature ACF Plot

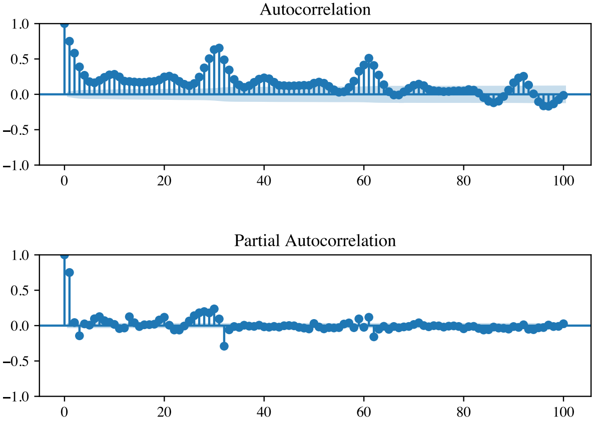
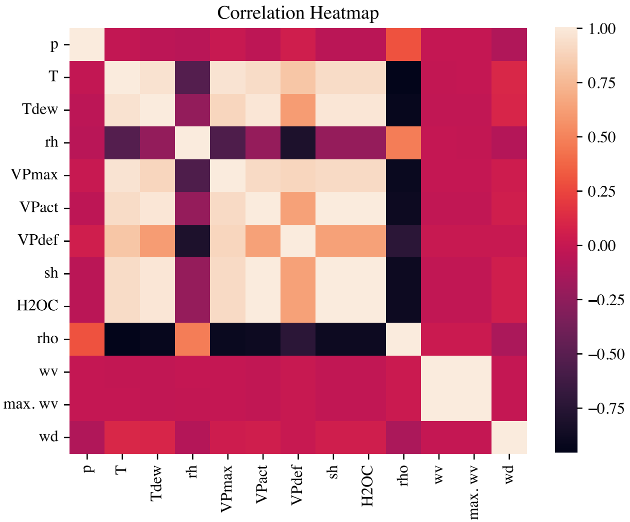


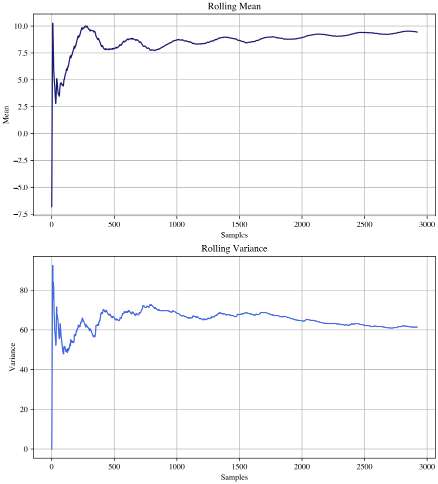

Figure: ACF and PACF of Temperature

# correlation Heatmap



# Table of Contents

# Rolling Mean and Variance



# ADF and KPSS Test

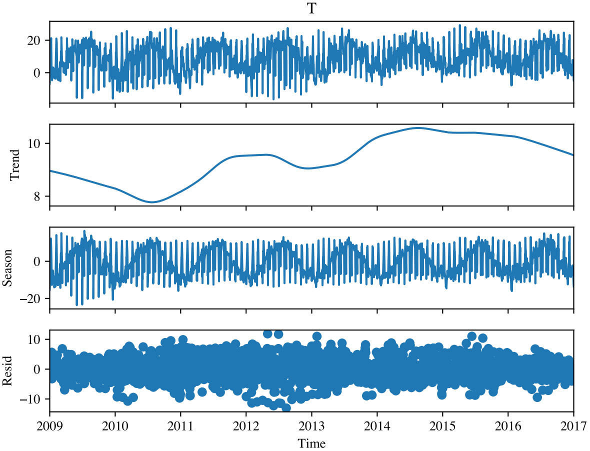
| Test | Test Stats. | P-Value | C. Val $1\%$ | C. Val. $5\%$ | C. Val. 10% |
|---|---|---|---|---|---|
| ADF | -3.210 | 0.019 | -3.433 | -2.863 | -2.567 |
| KPSS | 0.446 | 0.057 | 0.739 | 0.574 | 0.347 |

Table: ADF and KPSS test results

Results: Stationary

# Table of Contents

# Trend-Seasonality Decomposition



# Strengths of Trend and Seasonality

$F_T$ and $F_s$ measure the strength of the trend and seasonality component respectively.

$$F_T = 0.0784$$
$$F_S = 0.8362$$

# Differencing

- We will perform $s = 365$ days later for SARIMA

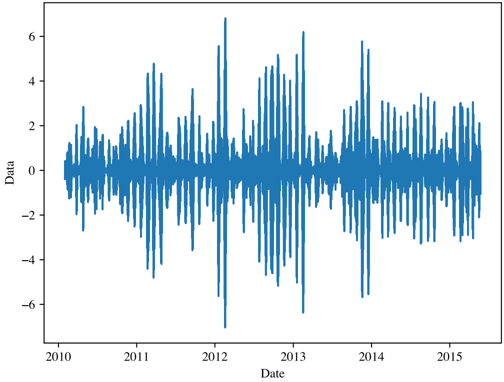



Figure: Differenced data

# ADF and KPSS Test

| Test | Test Stats. | P-Value | C. Val $1\%$ | C. Val. $5\%$ | C. Val. 10% |
|---|---|---|---|---|---|
| ADF | -14.164 | 0.000 | -3.433 | -2.863 | -2.567 |
| KPSS | 0.277 | 0.100 | 0.739 | 0.463 | 0.347 |

Table: ADF and KPSS test after seasonal differencing with $s = 365$ days

Result: Stationary

# Table of Contents

# Holt Winter Method

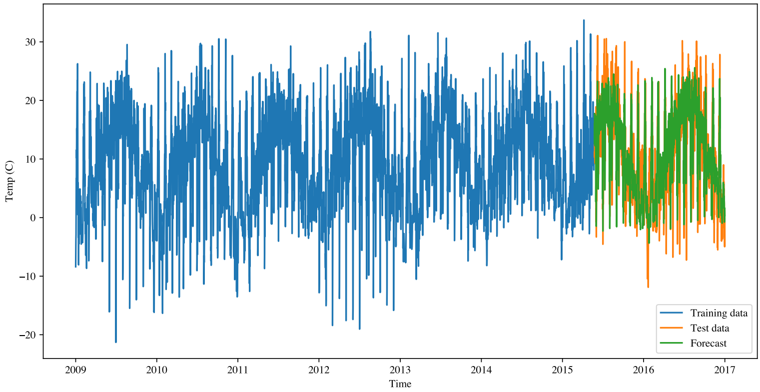



Figure: Holt Winter Model

# Table of Contents

# Frame Title

We perform our feature selection using Principal Component Analysis.

- Condition number $= 206021.99$

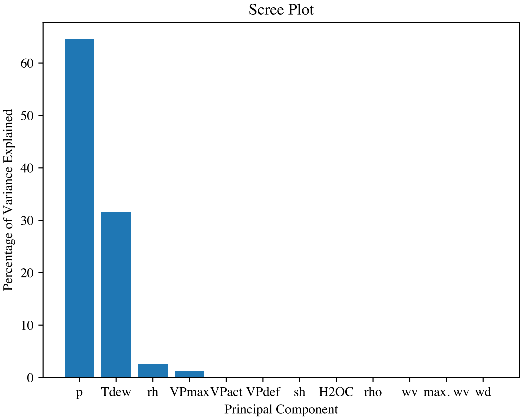
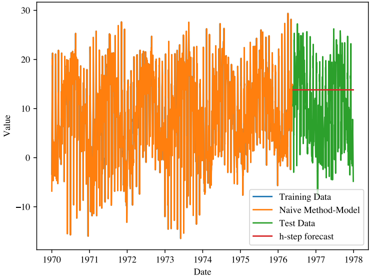


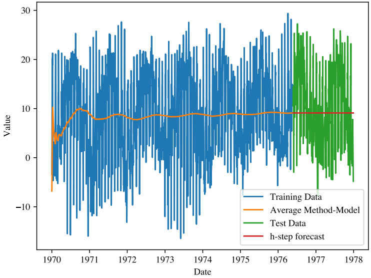

Figure: Percentage of Variance Explained by each variable

# Table of Contents

# Naive and Average Method



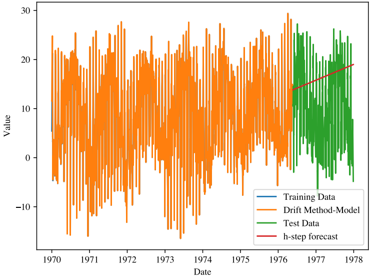

(a) Naive Method



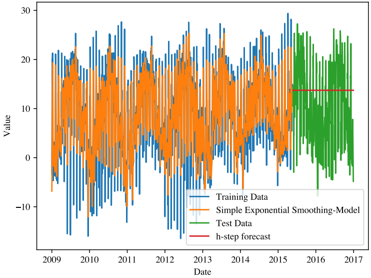

(b) Average Method

Figure: Training data, testing data, one-step prediction and h-step prediction for the naive method and average method models.

# Drift and Simple Exponential Smoothing (SES) Method




(a) Drift Method




(b) SES Method

Figure: Training data, testing data, one-step prediction and h-step prediction for the drift method and SES models.

# Base Models MSE

| Model | One-Step MSE | One-Step Q | H-Step MSE | H-Step Q |
|---|---|---|---|---|
| Naive | 30.851 | 11201 | 64.135 | 209.847 |
| Average | 62.775 | 87705.827 | 57.778 | 209.847 |
| Drift | 0.071 | 5871 | 91.607 | 226.600 |
| SES | 33.469 | 40104.049 | 63.622 | 209.847 |

Note, the fact that some h-step Q values are identical is odd although it is not obvious what may cause that, and perhaps it might be due to rounding.

# Table of Contents

# Multiple Linear Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | T | **R-squared (uncentered):** | 0.957 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.957 |
| **Method:** | Least Squares | **F-statistic:** | 2.612e+04 |
| **Date:** | Mon, 11 Dec 2023 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 20:38:12 | **Log-Likelihood:** | -5449.0 |
| **No. Observations:** | 2336 | **AIC:** | 1.090e+04 |
| **Df Residuals:** | 2334 | **BIC:** | 1.091e+04 |
| **Df Model:** | 2 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P> \|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **p** | 0.0039 | 6.36e-05 | 61.845 | 0.000 | 0.004 | 0.004 |
| **Tdew** | 1.1207 | 0.008 | 145.375 | 0.000 | 1.106 | 1.136 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 150.236 | **Durbin-Watson:** | 1.058 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 178.641 |
| **Skew:** | 0.667 | **Prob(JB):** | 1.62e-39 |
| **Kurtosis:** | 3.235 | **Cond. No.** | 148. |

Figure: Multiple Linear Regression Results

# Multiple Linear Regression Tests

1. F Test

Test if each coefficient is significant and different from zero

| F Test Stat. | P Value | DF_denom | DF_num |
|---|---|---|---|
| 17902.69 | 0.0 | $2.33 \times 10^3$ | 1 |

2. T tests

Test if the two values are significantly different

| | coef. | std. err. | t | $P > \|t\|$ | [0.025 0.975] |
|---|---|---|---|---|---|
| $c_0$ | -1.1208 | 0.008 | -132.708 | 0.000 | -1.137 -1.104 |

# Table of Contents

# GPAC Table

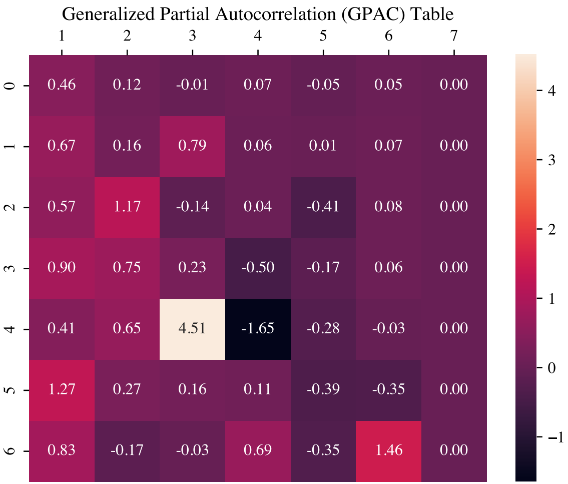



Figure: GPAC Table after differencing $\nabla^{30}\nabla_{365}$

# Orders selection

We select

- AR order $\hat{n}_a = 1$ and MA order $\hat{n}_b = 0$
- AR order $\hat{n}_a = 2$ and MA order $\hat{n}_b = 3$

# Table of Contents

# Parameters Determination

LM Algorithm: $a_1 \approx -0.46622$ and $-0.50609 < a_1 < -0.42636$

| | | | |
|---|---|---|---|
| **Dep. Variable:** | T | **No. Observations:** | 1971 |
| **Model:** | ARIMA(1, 0, 0) | **Log Likelihood** | 20880.626 |
| **Date:** | Mon, 11 Dec 2023 | **AIC** | -41755.252 |
| **Time:** | 19:21:55 | **BIC** | -41738.493 |
| **Sample:** | 01-01-2010 | **HQIC** | -41749.094 |
| | - 05-25-2015 | | |
| **Covariance Type:** | opg | | |

| | **coef** | **std err** | **z** | **P> \|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 2.933e-10 | 0.001 | 3.75e-07 | 1.000 | -0.002 | 0.002 |
| **ar.L1** | 0.4650 | 2.72e-08 | 1.71e+07 | 0.000 | 0.465 | 0.465 |
| **sigma2** | 1e-10 | 4.46e-11 | 2.240 | 0.025 | 1.25e-11 | 1.87e-10 |

| | | | |
|---|---|---|---|
| **Ljung-Box (L1) (Q):** | 6.60 | **Jarque-Bera (JB):** | 253.16 |
| **Prob(Q):** | 0.01 | **Prob(JB):** | 0.00 |
| **Heteroskedasticity (H):** | 0.97 | **Skew:** | -0.16 |
| **Prob(H) (two-sided):** | 0.72 | **Kurtosis:** | 4.73 |

Figure: Order determination with $\hat{n}_a = 1$ and $\hat{n}_b = 0$

# Table of Contents

# Forecast Function

$$(1 + a_1\, q^{-1})\,(1 - q^{-s}) y_t = \varepsilon_t \tag{1}$$

where $a_1 = -0.46622$ and $s = 365$ days. We can rewrite

$$\begin{aligned} (1 + a_1\, q^{-1})\,(y_t - y_{t-s}) &= \varepsilon_t \\ y_t - y_{t-s} + a_1\,(y_{t-1} - y_{t-s-1}) &= \varepsilon_t \\ y_{t+h} = y_{t+h-s} - a_1\,(y_{t+h-1} - y_{t+h-s-1}) &+ \varepsilon_{t+h} \end{aligned} \tag{2}$$

Then, we have

$$\hat{y}_t(h) = \mathrm{E}[y(t+h-s)] - a_1\,\mathrm{E}\,[y(t+h-1)] + a_1\,\mathrm{E}\,[y(t+h-s-1)] \tag{3}$$

# Forecast Function

- For $h = 1$:

$$\hat{y}_t(1) = y(t + 1 - s) - a_1 y(t) + a_1 y(t - s) \tag{4}$$

- For $2 \leq h \leq s$

$$\hat{y}_t(h) = y(t + h - s) - a_1 \hat{y}_t(h - 1) + a_1 y(t + h - s - 1) \tag{5}$$

- For $h > s$:

$$\hat{y}_t(h) = \hat{y}_t(h - s) - a_1 \hat{y}_t(h - 1) + a_1 \hat{y}_t(h - s - 1) \tag{6}$$

# Table of Contents

# Residual Analysis

- With Q = 7.4569 and Qc = 33.9303, the data are uncorellated (white)
- Variance of Error 0.000
- Forecast Error MSE 0.000
- Variance of the Forecast Error 0.000
- Estimated variance of error: 28.96940
- The model is unbiased

# Residual Analysis

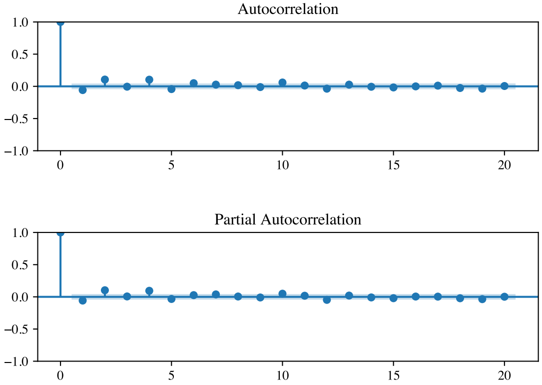



Figure: ACF and PACF Plot of residuals

# Table of Contents

# Model selection

We will use MSE as the metric

| Model | MSE |
|---|---|
| Naive Method | 30.851 |
| Average Method | 62.775 |
| Drift Method | 0.071 |
| SES Method | 33.469 |
| Holt Winter | 16.400 |
| Linear Regression | 6.217 |
| SARIMA | $\sim 10^{-17}$ |

We select **SARIMA**

# Final Model

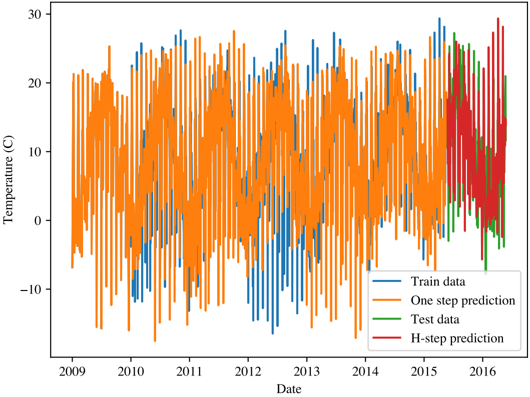



Figure: Final Model