Weather Forecasting

Modeling the Temperature in Jena, Germany

Edison M. Murairi

Final Report of Time Series Models & Analysis

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Data Science Program
The George Washington University
Washington DC, United States
December 12, 2023

Time Series Analysis and Models Homework 5 Report

Edison Murairi

December 12, 2023

Abstract

This study explores time series modeling for weather forecasting, focusing on temperature prediction in Jena, Germany. Analyzing a comprehensive dataset spanning seven years, we employ various models, including Holt-Winters, Multiple Linear Regression, and SARIMA. Feature selection techniques, such as Principal Component Analysis, refine predictive capabilities. The SARIMA model emerges as the most robust, demonstrating exceptional accuracy with a Mean Squared Error of 10^{-17}. This research provides valuable insights and a refined methodology for precise temperature forecasting in Jena.

Contents

List of Figures

List of Tables

1 Introduction

Time series models and analysis have emerged as indispensable tools in unraveling the complexities inherent in sequential data. This dynamic field of study has witnessed a surge in interest owing to its applications across diverse domains, from finance and economics to healthcare and environmental science. In this review, we delve into the key aspects of time series modeling, exploring its methodologies, challenges, and contemporary advancements.

At the heart of time series analysis lies the recognition that data points are not isolated entities but are interconnected over time, forming a sequence that often exhibits temporal patterns and dependencies. Traditional statistical models such as autoregressive integrated moving average (ARIMA) and seasonal decomposition of time series (STL) have long been stalwarts in capturing these patterns. ARIMA, in particular, excels in modeling linear trends and stationary time series, making it a reliable choice in various practical applications.

However, the advent of machine learning has brought forth a new era in time series analysis, ushering in sophisticated models that leverage the power of neural networks. Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) have demonstrated remarkable success in capturing long-term dependencies and non-linear patterns, propelling them to the forefront of time series modeling. Their ability to learn from sequential data has led to significant breakthroughs in forecasting accuracy, especially in domains with intricate and dynamic temporal structures.

In light of the significant success of time series models in the recent years, our final project is aimed at applying these techniques to weather forecasting. In particular, our goal is to model temperature in the city of Jena in Germany. Such a study contributes to the broader field of weather forecasting, enhancing our ability to predict and understand temperature dynamics with implications for sectors ranging from agriculture to urban planning.

2 Dataset

The 'Jena Weather' dataset forms the core of this project, offering a practical snapshot of weather conditions in Jena, Germany, over a span from January 1st, 2009, to January 1st, 2016. This dataset comprises a total of 420,551 measurements, recorded at 10-minute intervals. The dependent variable we seek to predict is the Temperature (Celcius) while Table 1 shows the independent variables in the dataset. There are 13 independent variables, all of which are numerical. One of the variable, Temperature in (Kelvin), representing the absolute temperature is colinear with the dependent variable[1]. We therefore drop the absolute temperature among our predictors. The dataset contains 420,551 measurements, recorded every 10 minutes. Since we do not expect a significant variation every 10 minutes, we downsample the measurements to every 12 hours, and obtain. Figure 1 shows the Temperature with time. The plot shows a strong seasonality, most likely 365 days and a small trend. Figure 2, on the other hand, shows the auto-correlation function (ACF) and the partial auto-correlation function

Variable	Explanation	Note
Pressure (mbar)	Atmospheric pressure	
Temperature	Temperature in Kelvin	Colinear with T (celcius)
Tdew (Celsius)	Temperature in Celsius relative to humidity	
rh (%)	Relative humidity	
VPmax (mbar)	Saturation vapor pressure	
VPact (mbar)	Vapor pressure	
VPdef (mbar)	Vapor pressure deficit	
sh (g/kg)	Specific humidity	
H_2O C (mmol/mol)	Water vapor concentration	
ρ (g/m^3)	Airtight	
' wv (m/s)	Wind speed	
max. wv (m/s)	Maximum wind speed	
wd (deg)	Wind direction	

Table 1: Independent variables in the dataset

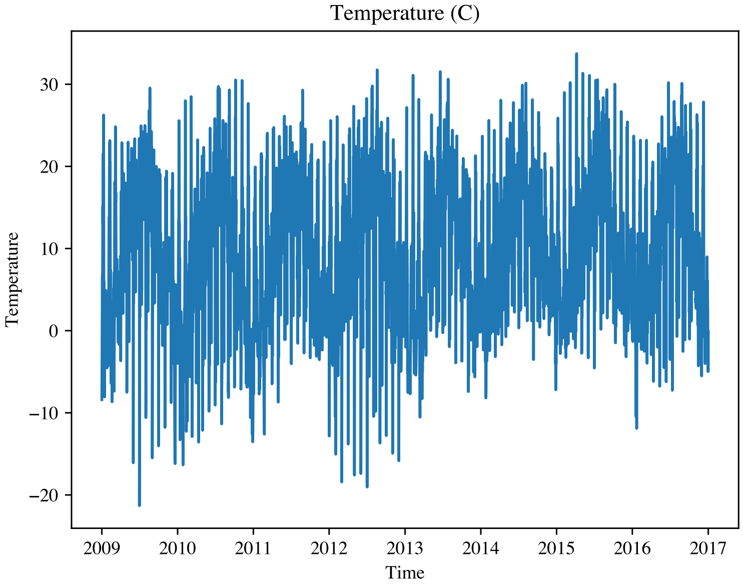


Figure 1: Temperature with Time

[1] T (Kelvin) = T (Celcius) + 273

(PACF). Both the ACF and PACF plots show that there are strong correlations every 30 days, in addition to the yearly pattern we have seen in Figure 1.

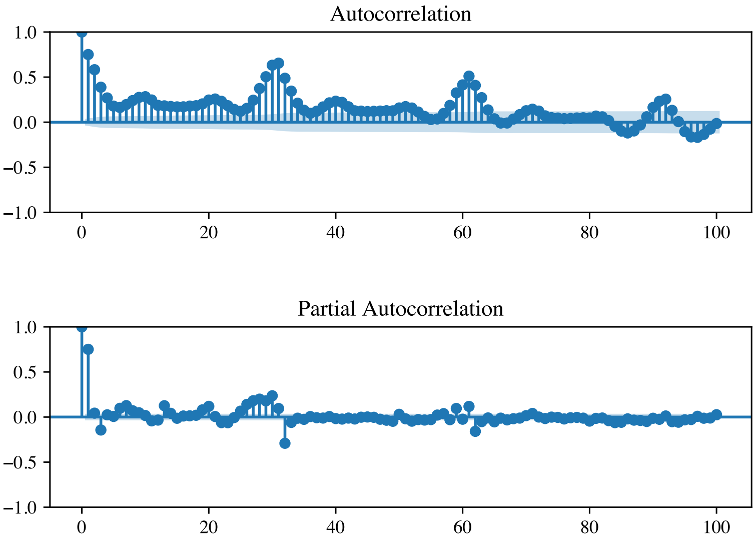


Figure 2: ACF and PACF of Temperature

Finally, Figure 3 shows the heatmap of all the variables. The figure shows a strong correlations of the dependent variable temperature (T) with pressure (p) and the temperature relative to the humidity (T_{dew}).

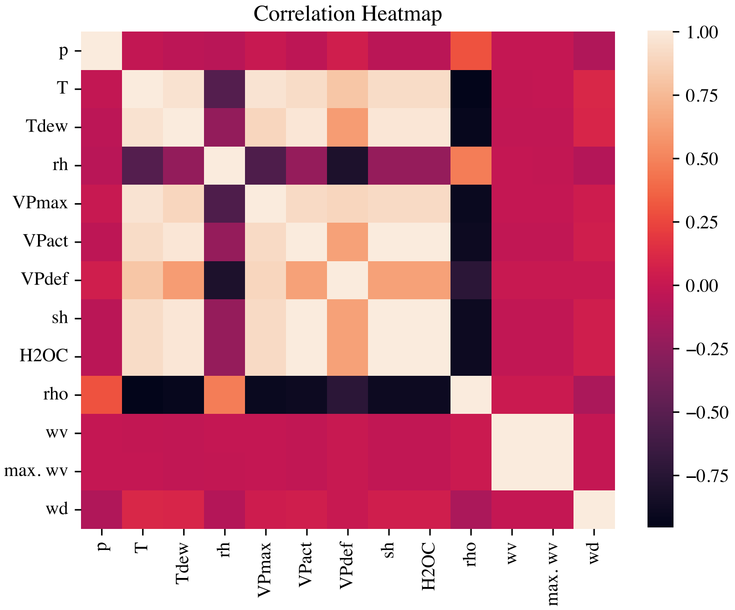


Figure 3: Heatmap of all the variables

Moreover, there a are signs of colinearity, for examole between H_2OC and sh, $VPact$ and sh, and $VPact$ and H_2OC, only to name a few. These colinearities will be explored in more details during the linear regression. For the rest of this work, we split the dataset into a training set (80%), and a test set (20%).

Test	Test Stats.	P-Value	Crit. Val 1%	Crit. Val. 5%	Crit. 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 2: ADF and KPSS test results

3 Stationarity

In this section, we test whether the data is stationary. To do so, we first plot rolling mean and variance, Fig. 4 shows the both the rolling mean and variance appear to have converged to a constant value, suggesting that the dataset is stationary. We will confirm this behavior with the ADF test and the KPSS test.

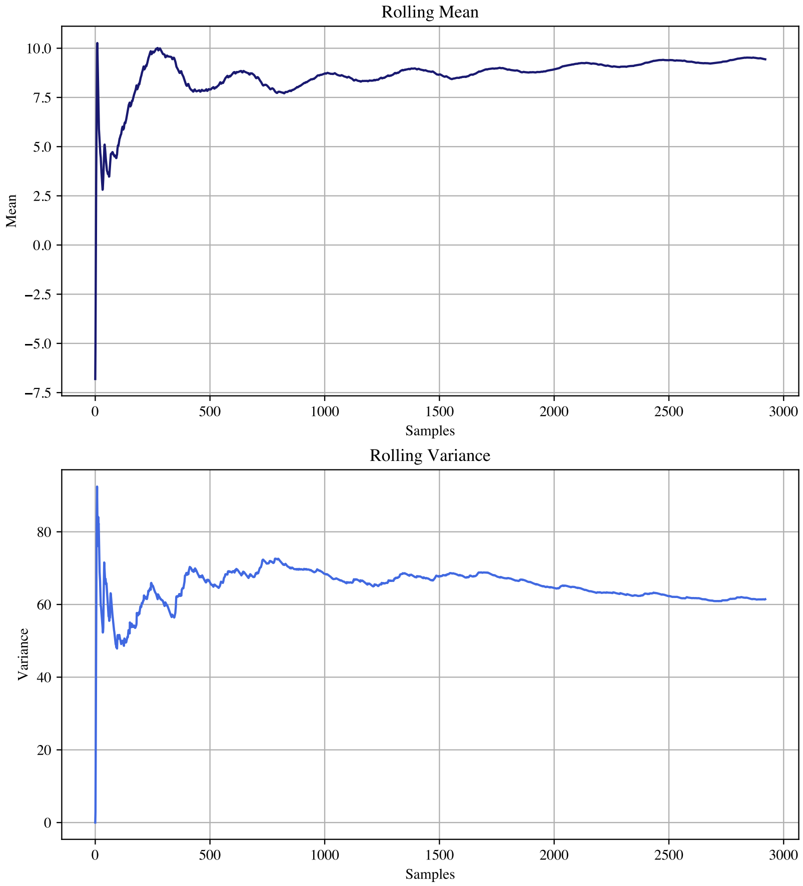


Figure 4: Rolling mean and variance

- ADF Test: The null hypothesis of the ADF test is that the dataset is non-stationary. We find a $P - value > 0.05$, suggesting to reject the null hypothesis and conclude that the dataset is stationary.
- KPSS Test The null hypothesis of the KPSS test is that the dataset is stationary. We find a $P - value < 0.05$, suggesting not to reject the null hypothesis and conclude that the dataset is stationary.

4 Time series Decomposition

In this section, we will decompose the dataset into its trend and seasonality component. Figure 5 shows the result of the decomposition.

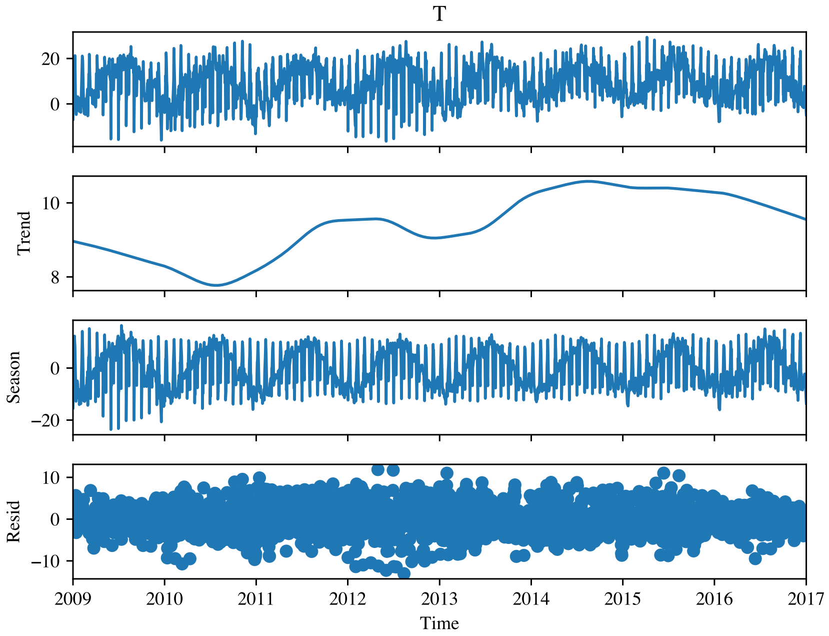


Figure 5: Trend seasonality decomposition

Given this decomposition, we found that

- The strength of trend for this data set is 0.0784
- The strength of seasonality for this data set is 0.8362

This decomposition shows a much stronger seasonality than trend as we expected from Figure 1. Even though we found that the dataset was stationary, the result of the ADF test was not convincing. Moreover, we expect weather patterns to repeat every ~ 365 days. Therefore, we perform a seasonal differencing of $s = 365$. We confirm that the dataset is still stationary and the Table 3 below show the ADF and KPSS test results after differencing.

Test	Test Stats.	P-Value	Crit. Val 1%	Crit. Val. 5%	Crit. 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 3: ADF and KPSS test after seasonal differencing with $s = 365$ days

5 Holt Winter Method

In this section, we use the Holt Winter Method to model the original dataset. The Holt-Winters model [14, 8], also known as the Triple Exponential Smoothing method, is a robust and widely used time series forecasting technique. It extends the basic exponential smoothing model to handle seasonality and trends in data. What sets the Holt-Winters model apart is its ability to capture and predict time series data with both upward or downward trends and seasonal patterns [5]. The model incorporates three smoothing equations for the level, trend, and seasonality components, each weighted with smoothing parameters. It is therefore a suitable candidate for our modeling. Figure 6 shows the result

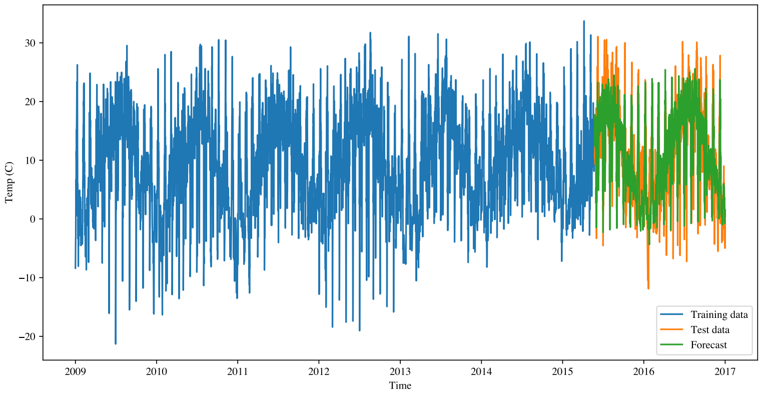


Figure 6: Holt Winter Model

of the Holt winter method. The model shows a good performance with mean-square error 16.400. This performance will be compared to that of other models to select the final model.

6 Feature selection

In section 2, we discussed possible colinearity issues within the predictors variables. In this section, we will primarily use the principal component analysis (PCA) to select the most relevant features in explaining the variance in the dependent variable. Below, we will provide a short review of PCA. We point the reader to Refs. [10, 3, 15], for a more detailed discussion.

Principal Component Analysis (PCA) stands as a powerful technique for feature selection, particularly in scenarios where datasets contain a high degree of multicollinearity among variables. PCA transforms the original features into a new set of uncorrelated variables, known as principal components, that capture the maximum variance in the data. By retaining only the principal components contributing significantly to the variance, PCA effectively reduces the dimensionality of the dataset, simplifying subsequent analyses without sacrificing crucial information.

First, we perform the SVD decomposition to compute the condition number.

$$\kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} = 206021.98 \tag{1}$$

Such a large condition number suggests a severe case of co-linearity. Finally, Figure 7 shows the proportion of the variance explained by each variable. From Figure 7, we select Pressure (p) and the temperature relative to the humidity (T_{dew}).

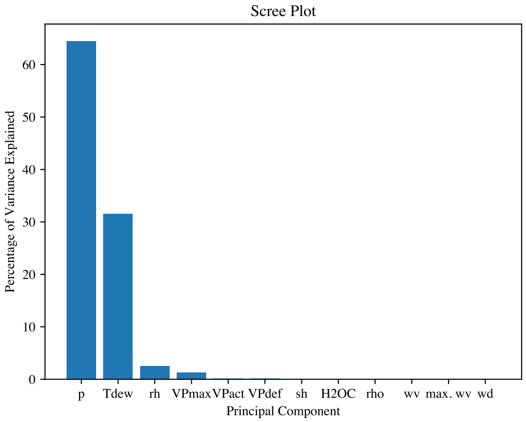


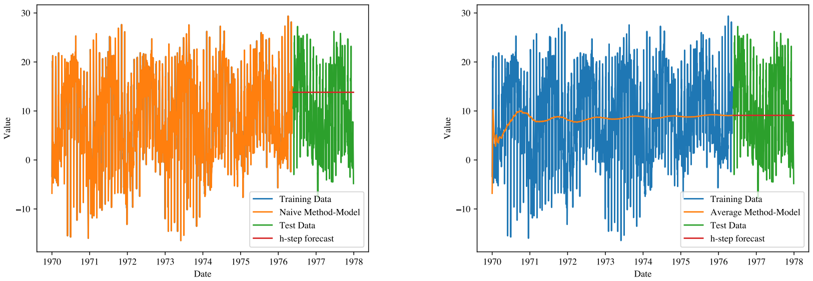
Figure 7: Percentage of Variance Explained by each variable

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

Table 4: Metrics for Base Models

7 Base Models

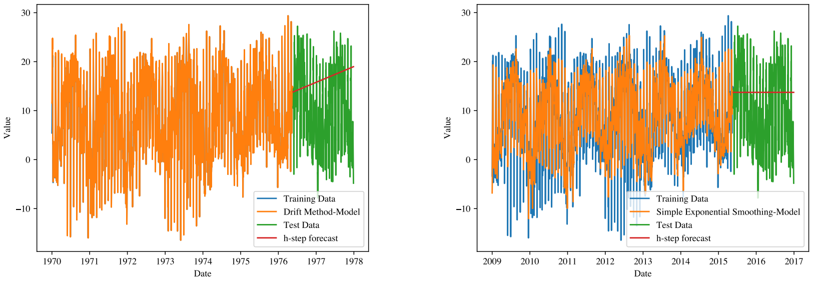
In this section, we will use the naive method, average method, drift and simple exponential smoothing to model the data, see e.g. Refs. [13, 7] for a review. Figures ?? show the plot while Table ?? show the metrics.



(a) Naive Method

(b) Average Method

Figure 8: Training data, testing data, one-step prediction and h-step prediction for the naive method and average method models.



(a) Drift Method

(b) SES Method

Figure 9: Training data, testing data, one-step prediction and h-step prediction for the drift method and SES models.

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.

8 Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical modeling technique used to analyze the relationship between a dependent variable and multiple independent variables, see e.g. Refs. [6, 1], . Unlike simple linear regression, which considers only one predictor, MLR accommodates the complexity of real-world scenarios where outcomes are influenced by multiple factors. The model assumes a linear relationship between the variables, with coefficients representing the strength and direction of their impact on the dependent variable.

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.

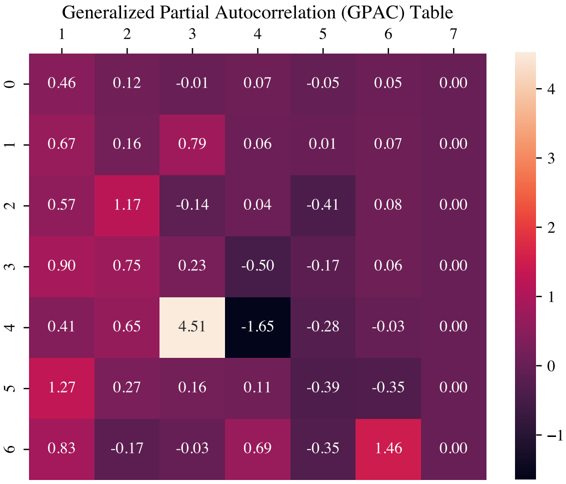
Table 5: Multiple Linear Regression Results

We perform a MLR with indepenent variables chosen from PCA: pressure and temperature relative to humidy, and dependent variable as temperature in Celcius. Table 5 shows the results.

The model also exhibits a good performance with MSE of 6.217 to be compared to that of other models. Moreover, we performed T test and F test for the significance of the regression parameters. We found that the parameters are significant.

9 ARMA/ARIMA/SARIMA/Multiplicative Model: Order determination

The Generalized Portmanteau Autocorrelation (GPAC) table is an essential tool in the process of order determination for Autoregressive Moving Average (ARMA) models, see e.g. Refs. [2, 9, 4]. The GPAC table aids in identifying the appropriate orders (p, q) for the autoregressive (AR) and moving average (MA). Figure 10 shows the GPAC Table for the differenced dataset.



	1	2	3	4	5	6	7
0	0.46	0.12	-0.01	0.07	-0.05	0.05	0.00
1	0.67	0.16	0.79	0.06	0.01	0.07	0.00
2	0.57	1.17	-0.14	0.04	-0.41	0.08	0.00
3	0.90	0.75	0.23	-0.50	-0.17	0.06	0.00
4	0.41	0.65	4.51	-1.65	-0.28	-0.03	0.00
5	1.27	0.27	0.16	0.11	-0.39	-0.35	0.00
6	0.83	-0.17	-0.03	0.69	-0.35	1.46	0.00

Figure 10: GPAC Table after seasonal differencing $s = 365$ days

From this figure, we have two choices for the orders:

- First choice:

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 - 05-25-2015	**HQIC**	-41749.094
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

Table 6: Parameters Estimation using statsmodels

  - AR order: $\hat{n}_a = 1$
  - MA order: $\hat{n}_b = 0$

- Second choice: It is difficult to make a choice different from what we have made. Another choice we can make, although not very good is. We will explore this choice only if the first one does not succeed.
  - AR order: $\hat{n}_a = 2$
  - MA order: $\hat{n}_b = 3$

10 Parameters Estimation: Levenberg Marquardt (LM) algorithm

In this section, we use the LM algorithm to estimate the parameters of the model. The Levenberg-Marquardt algorithm is an optimization method widely used for solving nonlinear least squares problems. It combines features of both the steepest descent and Gauss-Newton methods, offering robust convergence, especially in ill-conditioned situations, see e.g. Refs. [11, 12].

Our first model is $AR(1)_{s=365}$ and MA(0). Using our implementation of the algorithm, we obtain the results:

$$a_1 = -0.46622 \tag{2}$$

$$-0.50609 < a_1 < -0.42636 \tag{3}$$

This results shows that the coefficient is significant. Moreover, this coefficient also agrees with the results of the 'SARIMAX' fit function implemented in 'statsmodels', see Table 6.

11 Forecast Function

Now, we develop the forecast function. Our model is given by

$$(1 + a_1\, q^{-1})\,(1 - q^{-s})y_t = \varepsilon_t \tag{4}$$

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

$$\begin{aligned}(1 + a_1\, q^{-1})\,(y_t - y_{t-s}) = \varepsilon_t \\ y_t - y_{t-s} + a_1\,(y_{t-1} - y_{t-s-1}) = \varepsilon_t \\ y_{t+h} = y_{t+h-s} - a_1\,(y_{t+h-1} - y_{t+h-s-1}) + \varepsilon_{t+h}\end{aligned} \tag{5}$$

Then, we have

$$\hat{y}_t(h) = \mathrm{E}[y(t+h-s)] - a_1\,\mathrm{E}\,[y(t+h-1)] + a_1\,\mathrm{E}\,[y(t+h-s-1)] \tag{6}$$

- For $h = 1$:

$$\hat{y}_t(1) = y(t+1-s) - a_1 y(t) + a_1 y(t-s) \tag{7}$$

- For $2 \leq h \leq s$

$$\hat{y}_t(h) = y(t+h-s) - a_1 \hat{y}_t(h-1) + a_1 y(t+h-s-1) \tag{8}$$

- For $h > s$:

$$\hat{y}_t(h) = \hat{y}_t(h-s) - a_1 \hat{y}_t(h-1) + a_1 \hat{y}_t(h-s-1) \tag{9}$$

12 Residual Analysis

Now, we perform the residual analysis to determine whether the model derived in sec. **??** is appropriate. We obtained the result as follows:

- 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

Figure 11 shows the ACF and PACF plot of the residuals, consistent with a white noise.

13 Final Model

We will use MSE as the metric to select the final model among those studied.

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	10^{-17}

Table 7: MSE for all the models

We see in table 7 that the SARIMA model with $\hat{n}_a = 1$, $\hat{n}_b = 0$ with seasonality $s = 365$ days perform much better than all the other models; it is therefore our final model.

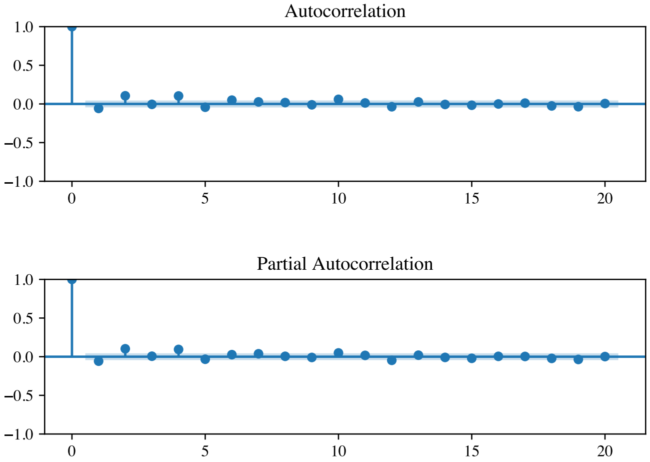


Figure 11: ACF and PACF of residuals

14 H-Step ahead Predictions

Having derived the forecast function in sec. 11, we will now perform h-step predictions. For simplicity, we perform the h-step prediction for one year. Figure 12 shows the result. The figure shows that the forecast strongly agrees with the testing dataset.

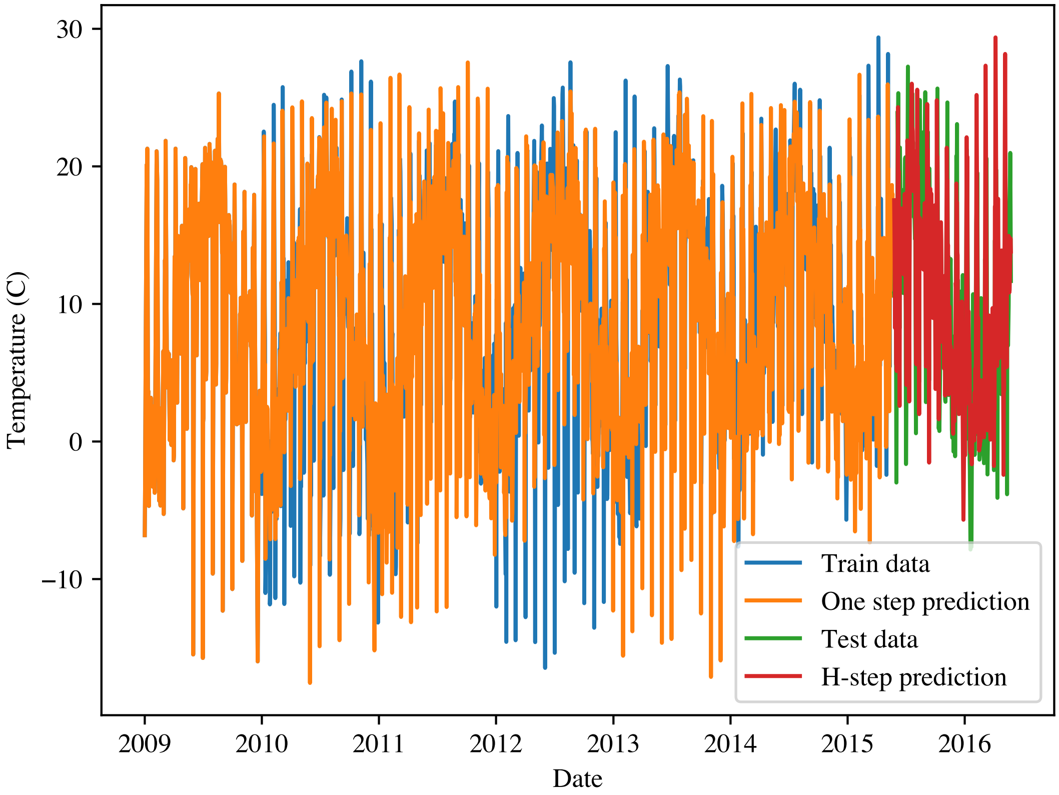


Figure 12: Final Model H step predictions

15 Summary and Conclusion

In this work, we studied weather variations in Jena with the goal of modeling and forecasting its temperature. We began with a detailed exploration of the Jena Weather dataset. Key independent variables were identified, and initial data analysis, including time series decomposition and stationarity tests, was performed. The report then delved into the application of different forecasting models, including Holt-Winters, multiple linear regression, and SARIMA, with an emphasis on order determination using the GPAC table. The Levenberg-Marquardt algorithm was used for parameter estimation. We derived the SARIMA forecasting function, and a thorough residual analysis validates the model's appropriateness. The report concludes with h-step ahead predictions and a comparison of various models based on MSE metrics.

In conclusion, the time series analysis conducted on the Jena Weather dataset has provided valuable insights into temperature dynamics. The SARIMA model, with an AR order of 1, MA order of 0, and a seasonality of 365 days, emerged as the most accurate forecasting model. The h-step ahead predictions showcase the model's ability to capture and forecast temperature variations. This study contributes to the broader field of weather forecasting, offering a robust methodology for predicting temperature patterns in Jena, Germany.

References

[1] L. S. Aiken, S. G. West, and S. C. Pitts. Multiple linear regression. *Handbook of psychology*, pages 481–507, 2003.

[2] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[3] R. Bro and A. K. Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

[4] P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. Springer, 2002.

[5] C. Chatfield. The holt-winters forecasting procedure. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(3):264–279, 1978.

[6] N. R. Draper and H. Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[7] M. Geurts. Book review: time series analysis: forecasting and control, 1977.

[8] C. W. Granger and P. Newbold. Some comments on the evaluation of economic forecasts. *Applied Economics*, 5(1):35–47, 1973.

[9] J. D. Hamilton. *Time series analysis*. Princeton university press, 2020.

[10] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[11] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

[12] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[13] D. C. Montgomery, C. L. Jennings, and M. Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[14] P. R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960.

[15] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

A Linear Regression Parameters Significance

A.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	2.33×10^3	1

A.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