Time Series Analysis for Temperature Prediction

DATS 6313

Liang Gao

George Washington University

10 December 2023

Content

1 Overview

2 Preprocessing the Data

3 Stationarity

4 Time Series Decomposition

5 Holt-Winter Method

6 Feature Selection & Regression

7 Base Models

8 SARIMA Model

9 Residual Analysis

10 Conclusion

List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

Overview

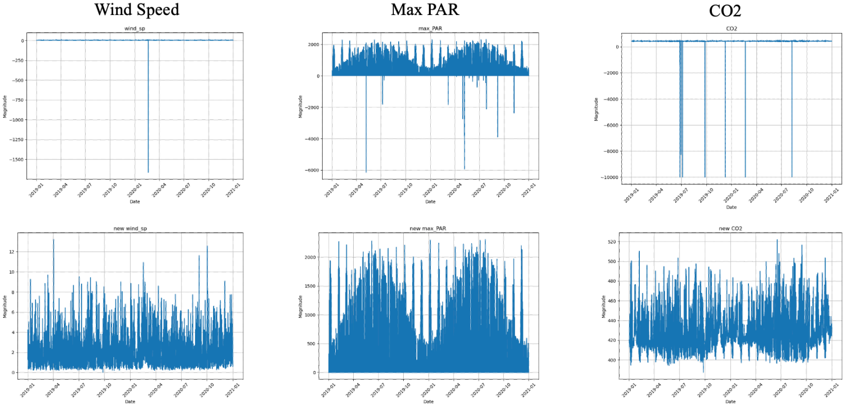
- The Jena Weather dataset was recorded every 10 minutes(2004 - 2020)
- Choose 3 years (2018 - 2020) of data, and average the data in each hour
- The dataset has 22 columns including " date" and 21 numerical variables(e.g. atmospheric pressure, Relative Humidity, Vapor pressure).
- The dependent variable is the temperature in Celsius

List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

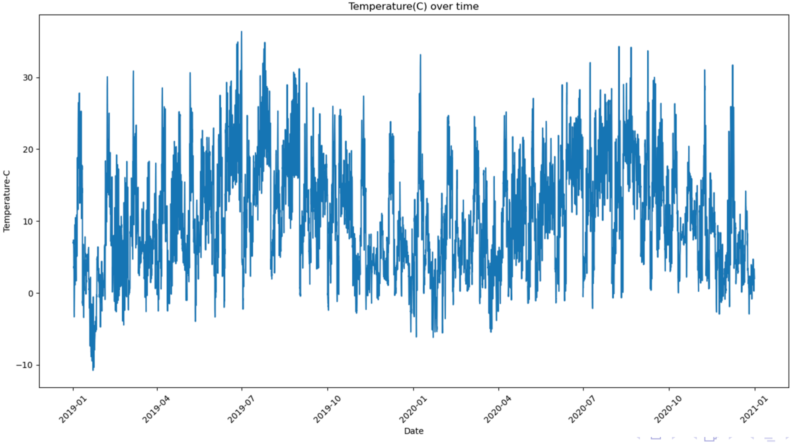
Preprocessing the Data

- Use the Drift method to fill in the missing values
- Outliers fix: Average method for CO2 & Wind speed; Naive for max PAR



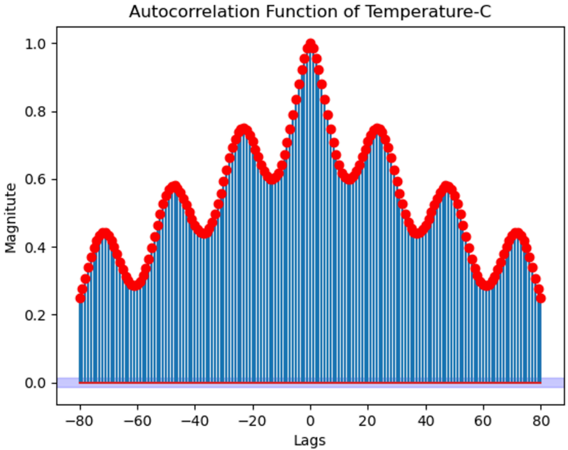
Preprocessing the Data

- Temperature over Time



Preprocessing the Data

- ACF of Temperature



Preprocessing the Data

- The observation is 14,036 in the train set(80%) and 3,509 in the test set(20%)

List of Content

1 Overview

2 Preprocessing the Data

3 Stationarity

4 Time Series Decomposition

5 Holt-Winter Method

6 Feature Selection & Regression

7 Base Models

8 SARIMA Model

9 Residual Analysis

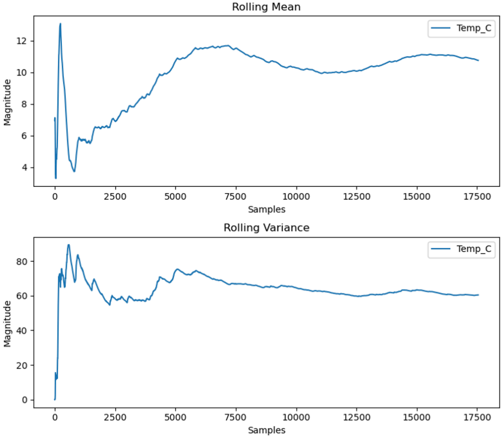
10 Conclusion

Stationarity

- The target variable passes the ADF test with a p-value of 0.00 but fails to pass the KPSS test with a p-value of 0.02
- The rolling mean and variance of temperature in Celsius, which stabilize once all samples are included
- The target variable dataset is weak-stationary

Stationarity

- Rolling mean & Variance of Temperature in Celsius

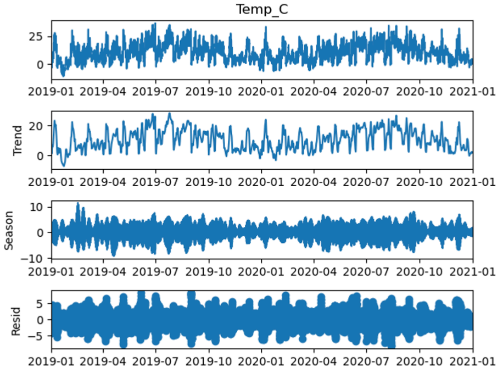


List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

Time Series Decomposition

- The strength of the trend is 94.37%, and the strength of the seasonality is 74.79%

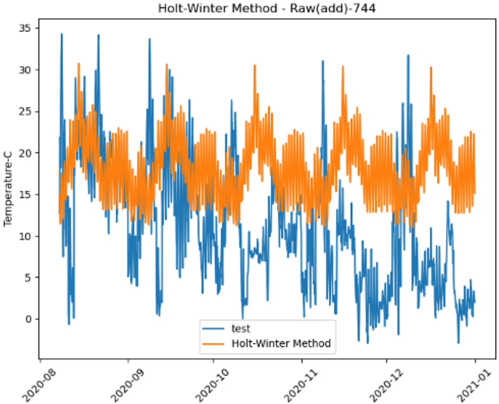


List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

Holt-Winter Method

- This method captures most seasonality but not the trend



List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

Co-linearity check

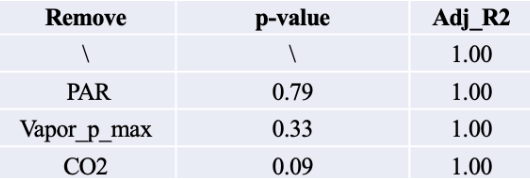
- All the singular values are greater than 0, but the last few singular values are relatively small compared to the first largest one
- The condition number is 1,409,780.69 and is highly greater than 1,000
- Both results indicate severe co-linearity among some independent variables

Principal Component Analysis(PCA)

- The threshold for the PCA feature selection is a variance ratio of less than 0.95
- 7 features are chosen
- Adjusted R-squared - 0.982
- Mean of error - 0.005
- Variance of error - 0.017
- MSE - 0.017
- All the coefficients are statistically significant with p-values less than 0.05

Backwards Stepwise Regression

- Started with the model containing all independent variables, removed one predictor with the highest p-value at a time. 3 features were deleted



Remove	p-value	Adj_R2
\	\	1.00
PAR	0.79	1.00
Vapor_p_max	0.33	1.00
CO2	0.09	1.00

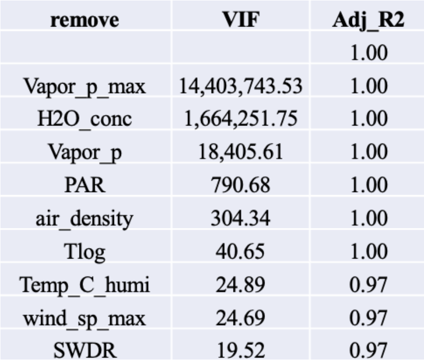
- Remove features with small coefficients(confidence interval for " rain time" is [-0.001, -0.000])

Backwards Stepwise Regression

- 8 features are chosen
- Adjusted R-squared: 1
- Mean of error: 0.001
- Variance of error & MSE: less than 0.00001
- All the coefficients are statistically significant with p-values less than 0.05
- Problem: The condition number of the regression model is 4.1e+03, which indicates strong multi-collinearity or other numerical problems

Variance Inflation Factor(VIF))

- The threshold for the VIF value is 10
- removed one predictor with the highest VIF value at a time(deleted 9 features)



remove	VIF	Adj_R2
		1.00
Vapor_p_max	14,403,743.53	1.00
H2O_conc	1,664,251.75	1.00
Vapor_p	18,405.61	1.00
PAR	790.68	1.00
air_density	304.34	1.00
Tlog	40.65	1.00
Temp_C_humi	24.89	0.97
wind_sp_max	24.69	0.97
SWDR	19.52	0.97

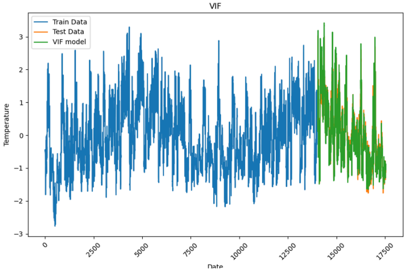
- Delete one insignificant feature & 6 features with small coefficients

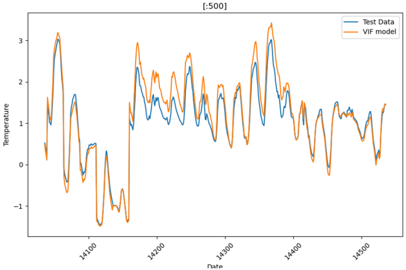
Variance Inflation Factor(VIF)

- 3 features are chosen
- Adjusted R-squared: 0.971
- Mean & Variance of error: 0.031
- MSE: 0.030
- All the coefficients are statistically significant with p-values less than 0.05

Final Regression Model(VIF)

- Model derived from VIF has fewer features and no multi-collinearity problem
- Model performance

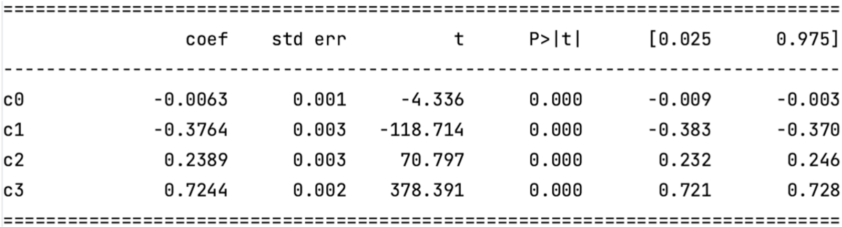




Final Regression Model(VIF)

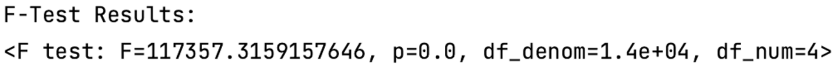
- Hypothesis tests: F-test & T-test

T-test



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            -0.0063      0.001     -4.336      0.000      -0.009      -0.003
c1            -0.3764      0.003   -118.714      0.000      -0.383      -0.370
c2             0.2389      0.003     70.797      0.000       0.232       0.246
c3             0.7244      0.002    378.391      0.000       0.721       0.728
==============================================================================
```

F-test



```
F-Test Results:
<F test: F=117357.3159157646, p=0.0, df_denom=1.4e+04, df_num=4>
```

Final Regression Model(VIF)

- Cross-validation

Subset	MSE	MeanRMSE	R-squared	Adj R-squared
1	0.05	0.22	0.96	0.96
2	0.04	0.19	0.97	0.97
3	0.02	0.15	0.97	0.97
4	0.04	0.19	0.97	0.97
5	0.03	0.16	0.97	0.97

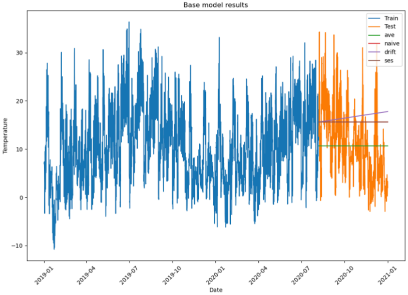
- The consistency of the metrics across different subsets suggests that the model is stable and generalizes well to different subsets of the data

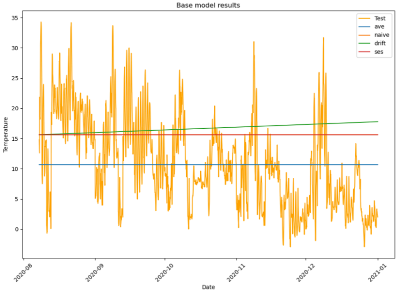
List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

Base Models

- Average, Niave, Drift, and SES





Base Models

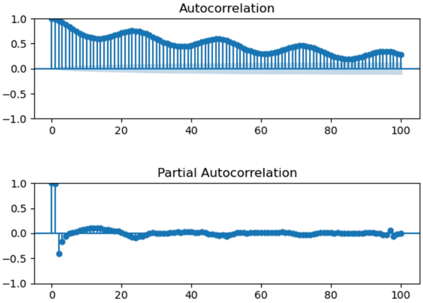
Model	Mean	Variance	MSE
Average	0.39	55.47	55.62
Naive	-4.56	55.47	76.23
Dirft	-5.64	61.04	92.82
SES	-4.56	55.47	76.23

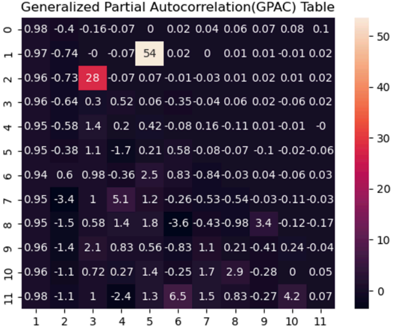
List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

SARIMA Model

- GPAC & ACF/PACF of Raw Dataset



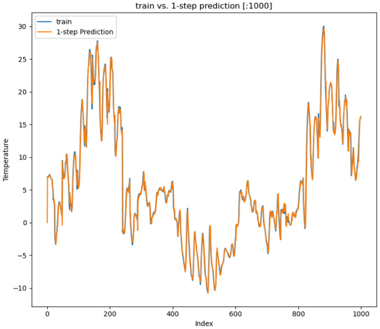

50
40
30
20
10
0

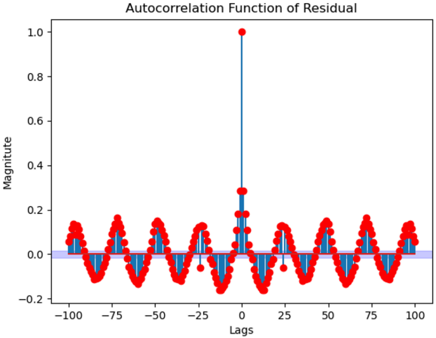
Generalized Partial Autocorrelation(GPAC) Table

	1	2	3	4	5	6	7	8	9	10	11
0	0.98	-0.4	-0.16	-0.07	0	0.02	0.04	0.06	0.07	0.08	0.1
1	0.97	-0.74	-0	-0.07	54	0.02	0	0.01	0.01	-0.01	0.02
2	0.96	-0.73	28	-0.07	0.07	-0.01	-0.03	0.01	0.02	0.01	0.02
3	0.96	-0.64	0.3	0.52	0.06	-0.35	-0.04	0.06	0.02	-0.06	0.02
4	0.95	-0.58	1.4	0.2	0.42	-0.08	0.16	-0.11	0.01	-0.01	-0
5	0.95	-0.38	1.1	-1.7	0.21	0.58	-0.08	-0.07	-0.1	-0.02	-0.06
6	0.94	0.6	0.98	-0.36	2.5	0.83	-0.84	-0.03	0.04	-0.06	0.03
7	0.95	-3.4	1	5.1	1.2	-0.26	-0.53	-0.54	-0.03	-0.11	-0.03
8	0.95	-1.5	0.58	1.4	1.8	-3.6	-0.43	-0.98	3.4	-0.12	-0.17
9	0.96	-1.4	2.1	0.83	0.56	-0.83	1.1	0.21	-0.41	0.24	-0.04
10	0.96	-1.1	0.72	0.27	1.4	-0.25	1.7	2.9	-0.28	0	0.05
11	0.98	-1.1	1	-2.4	1.3	6.5	1.5	0.83	-0.27	4.2	0.07

(1,0,1) (1,0,0,24)

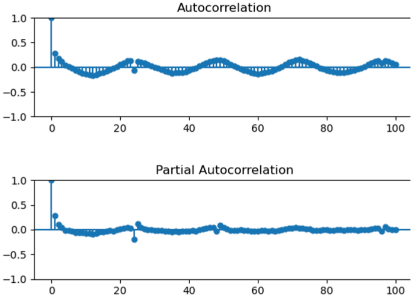
- 1-step prediction & residual ACF

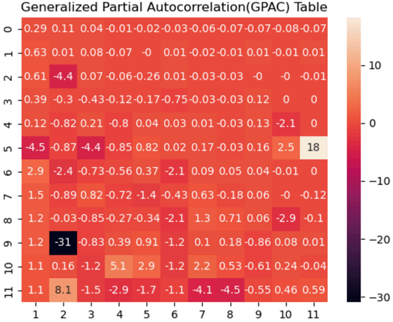




(1,0,1) (1,0,0,24)

- Residual ACF/PACF & GPAC



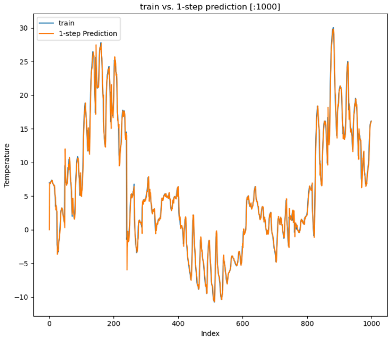


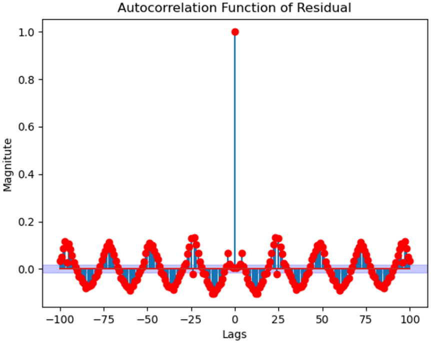
Generalized Partial Autocorrelation(GPAC) Table

	1	2	3	4	5	6	7	8	9	10	11
0	0.29	0.11	0.04	-0.01	-0.02	-0.03	-0.06	-0.07	-0.07	-0.08	-0.07
1	0.63	0.01	0.08	-0.07	-0	0.01	-0.02	-0.01	0.01	-0.01	0.01
2	0.61	-4.4	0.07	-0.06	-0.26	0.01	-0.03	-0.03	-0	-0	-0.01
3	0.39	-0.3	-0.43	-0.12	-0.17	-0.75	-0.03	-0.03	0.12	0	0
4	0.12	-0.82	0.21	-0.8	0.04	0.03	0.01	-0.03	0.13	-2.1	0
5	-4.5	-0.87	-4.4	-0.85	0.82	0.02	0.17	-0.03	0.16	2.5	18
6	2.9	-2.4	-0.73	-0.56	0.37	-2.1	0.09	0.05	0.04	-0.01	0
7	1.5	-0.89	0.82	-0.72	-1.4	-0.43	0.63	-0.18	0.06	-0	-0.12
8	1.2	-0.03	-0.85	-0.27	-0.34	-2.1	1.3	0.71	0.06	-2.9	-0.1
9	1.2	-31	-0.83	0.39	0.91	-1.2	0.1	0.18	-0.86	0.08	0.01
10	1.1	0.16	-1.2	5.1	2.9	-1.2	2.2	0.53	-0.61	0.24	-0.04
11	1.1	8.1	-1.5	-2.9	-1.7	-1.1	-4.1	-4.5	-0.55	0.46	0.59

(1,0,3) (1,0,1,24)

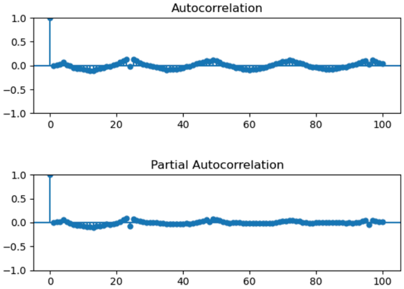
- 1-step prediction & residual ACF

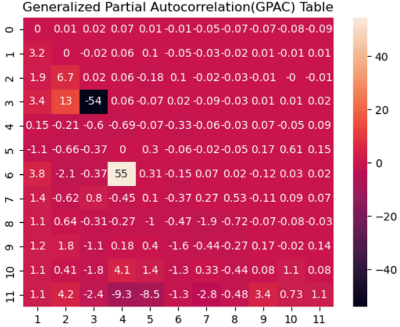




(1,0,3) (1,0,1,24)

- Residual ACF/PACF & GPAC



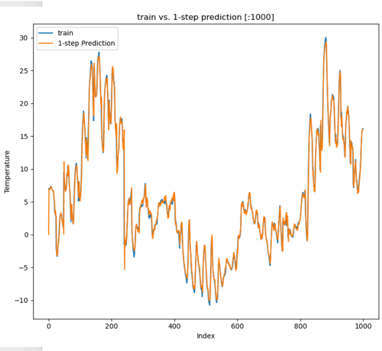


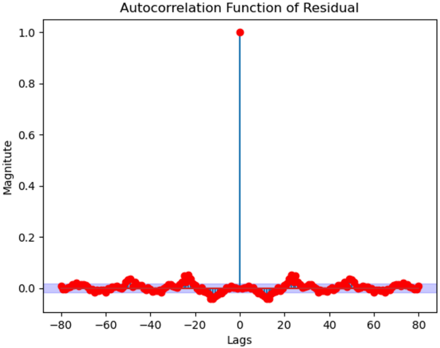
Generalized Partial Autocorrelation(GPAC) Table

	1	2	3	4	5	6	7	8	9	10	11
0	0	0.01	0.02	0.07	0.01	-0.01	-0.05	-0.07	-0.07	-0.08	-0.09
1	3.2	0	-0.02	0.06	0.1	-0.05	-0.03	-0.02	0.01	-0.01	0.01
2	1.9	6.7	0.02	0.06	-0.18	0.1	-0.02	-0.03	-0.01	-0	-0.01
3	3.4	13	-54	0.06	-0.07	0.02	-0.09	-0.03	0.01	0.01	0.02
4	0.15	-0.21	-0.6	-0.69	-0.07	-0.33	-0.06	-0.03	0.07	-0.05	0.09
5	-1.1	-0.66	-0.37	0	0.3	-0.06	-0.02	-0.05	0.17	0.61	0.15
6	3.8	-2.1	-0.37	55	0.31	-0.15	0.07	0.02	-0.12	0.03	0.02
7	1.4	-0.62	0.8	-0.45	0.1	-0.37	0.27	0.53	-0.11	0.09	0.07
8	1.1	0.64	-0.31	-0.27	-1	-0.47	-1.9	-0.72	-0.07	-0.08	-0.03
9	1.2	1.8	-1.1	0.18	0.4	-1.6	-0.44	-0.27	0.17	-0.02	0.14
10	1.1	0.41	-1.8	4.1	1.4	-1.3	0.33	-0.44	0.08	1.1	0.08
11	1.1	4.2	-2.4	-9.3	-8.5	-1.3	-2.8	-0.48	3.4	0.73	1.1

(1,0,3) (2,0,2,24)

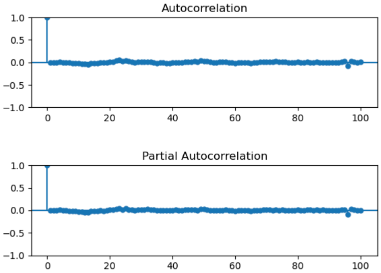
- 1-step prediction & residual ACF

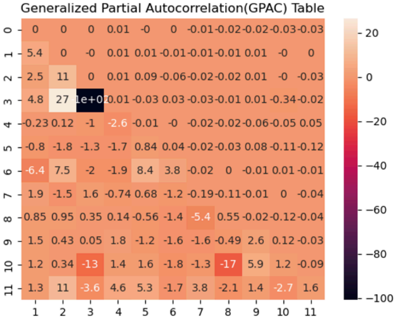




(1,0,3) (2,0,2,24)

- Residual ACF/PACF & GPAC

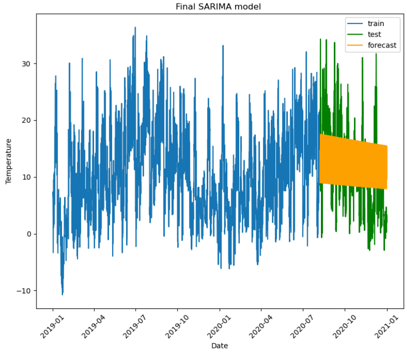


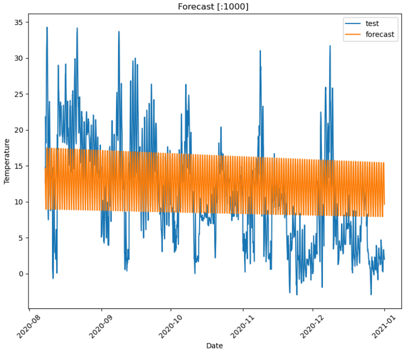

Generalized Partial Autocorrelation(GPAC) Table

	1	2	3	4	5	6	7	8	9	10	11
0	0	0	0	0.01	-0	0	-0.01	-0.02	-0.02	-0.03	-0.03
1	5.4	0	-0	0.01	0.01	-0.01	-0.01	-0.01	0.01	-0	0
2	2.5	11	0	0.01	0.09	-0.06	-0.02	-0.02	0.01	-0	-0.03
3	4.8	27	-1e+02	0.01	-0.03	0.03	-0.03	-0.01	0.01	-0.34	-0.02
4	-0.23	0.12	-1	-2.6	-0.01	-0	-0.02	-0.02	-0.06	-0.05	0.05
5	-0.8	-1.8	-1.3	-1.7	0.84	0.04	-0.02	-0.03	0.08	-0.11	-0.12
6	-6.4	7.5	-2	-1.9	8.4	3.8	-0.02	0	-0.01	0.01	-0.01
7	1.9	-1.5	1.6	-0.74	0.68	-1.2	-0.19	-0.11	-0.01	0	-0.04
8	0.85	0.95	0.35	0.14	-0.56	-1.4	-5.4	0.55	-0.02	-0.12	-0.04
9	1.5	0.43	0.05	1.8	-1.2	-1.6	-1.6	-0.49	2.6	0.12	-0.03
10	1.2	0.34	-13	1.4	1.6	-1.8	-1.3	-17	5.9	1.2	-0.09
11	1.3	11	-3.6	4.6	5.3	-1.7	3.8	-2.1	1.4	-2.7	1.6

(1,0,3) (2,0,2,24)

- SARIMA model performance





List of Content

1 Overview
2 Preprocessing the Data
3 Stationarity
4 Time Series Decomposition
5 Holt-Winter Method
6 Feature Selection & Regression
7 Base Models
8 SARIMA Model
9 Residual Analysis
10 Conclusion

Residual Analysis

- Box-Pierce test: $Q > Q^*$, fail the test
- Ljung-Box: p-values less than 0.05, fail the test
- Biased model: The estimated mean of the forecast error is -1.64
- Variance of the residual errors is 1.35 & Variance of forecast errors is 47.65
- Perform a zero-pole cancellation operation and there is no zero cancellation

List of Content

1 Overview

2 Preprocessing the Data

3 Stationarity

4 Time Series Decomposition

5 Holt-Winter Method

6 Feature Selection & Regression

7 Base Models

8 SARIMA Model

9 Residual Analysis

10 Conclusion

Conclusion

Method	Variance	Variance improvement (%)	MSE	MSE improvement (%)
Average	55.47	14.1 %	55.62	9.46%
Naive	55.47	14.1 %	76.23	33.94%
Dirft	61.04	21.94 %	92.82	45.74 %
SES	55.47	14.1%	76.23	33.94%
SARIMA	47.65	-	50.36	-