Carnegie Mellon University Africa

6/15/2023

IMDEA Homework Report

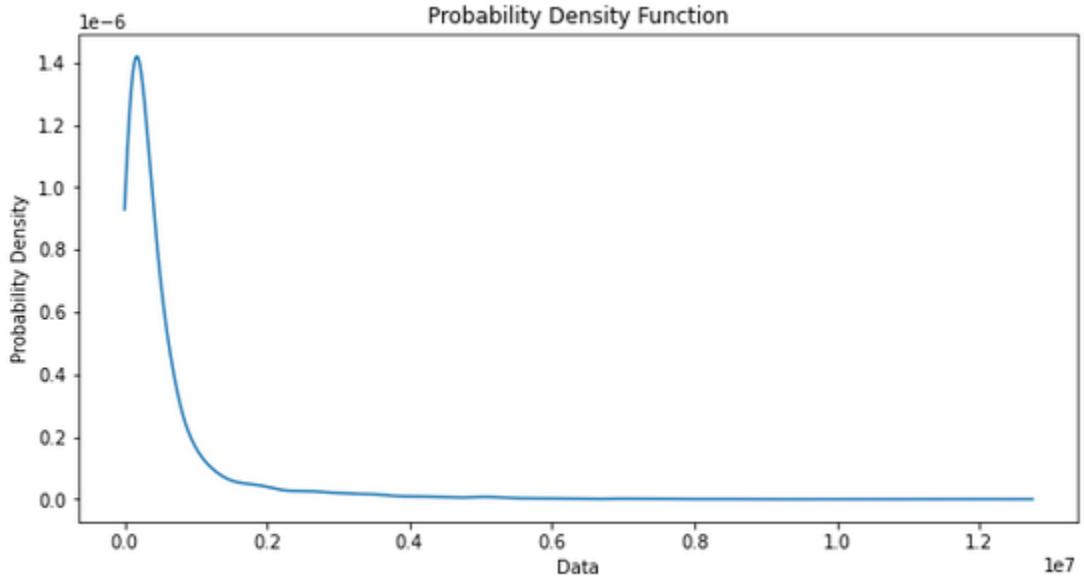
Arlette Bidossessi Houndji

email: houndji.arlette@gmail.com

Task I

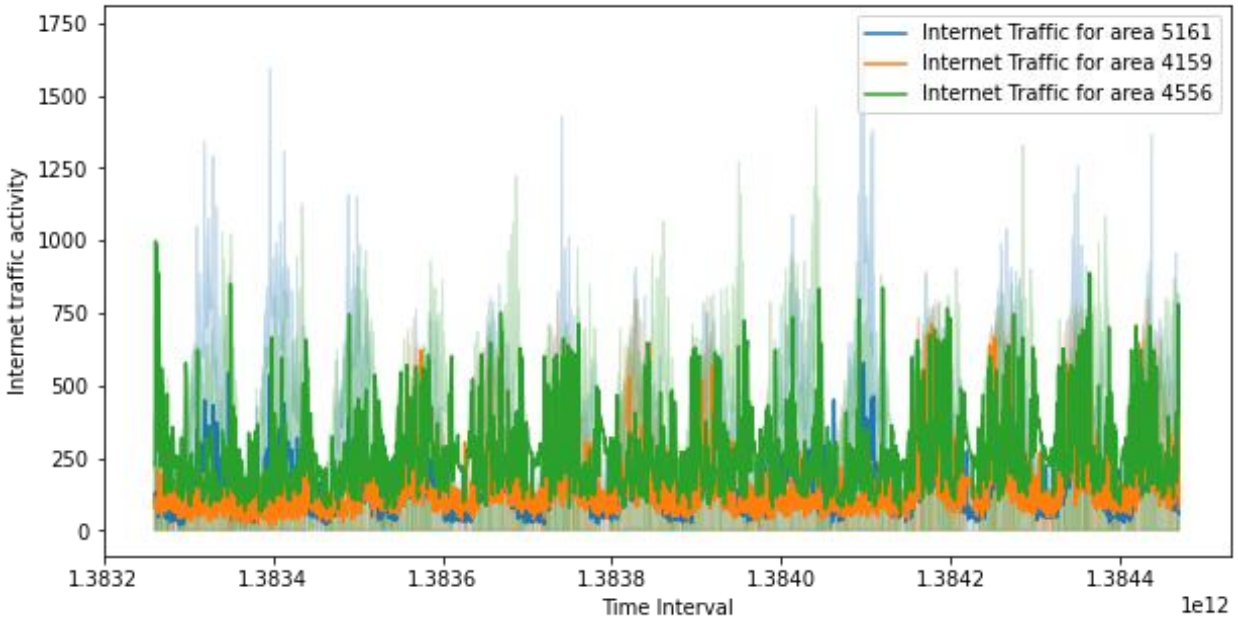
1) Description of the hardware and software setup to process the data:
   Computer: HP Probook 450 GZ
   Processor: Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz 2.30 GHz
   RAM: 16GB
   Architecture: x64
   OS: Windows 11 Pro
   Python: Jupyter Notebook (Anaconda3)

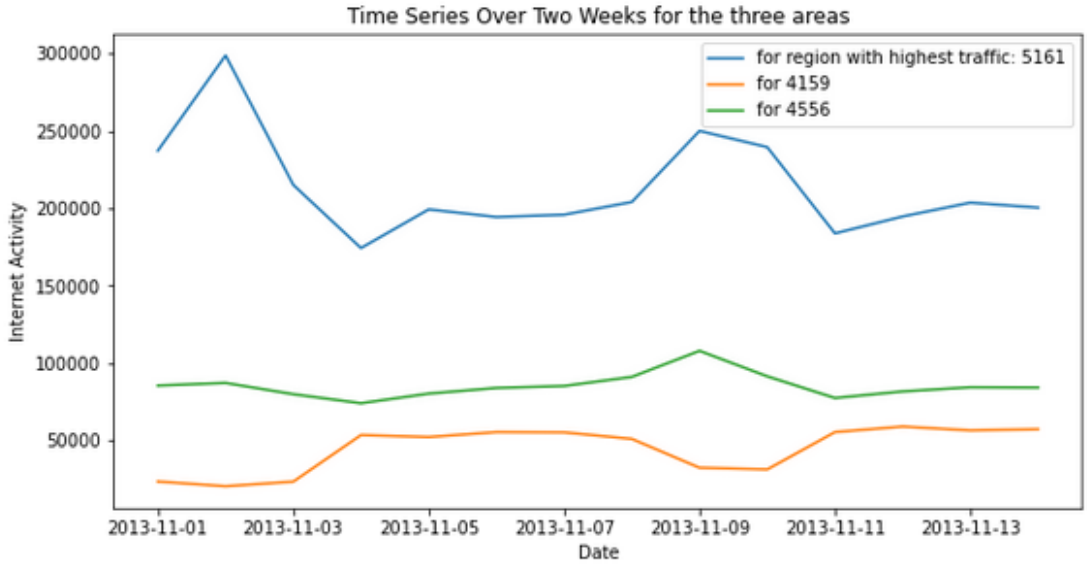
2) Plotting the PDF:



   The PDF shows that the data points exhibit different behaviors across different probability values. The tail of the distribution decays slowly, indicating a higher probability of observing values far from the mean. This could be because of the heterogenous traffic activities observed from the different areas in the city of Milan. Some zones have a higher internet usage than others.

3) Time series of network traffic during the first two weeks in the target city.

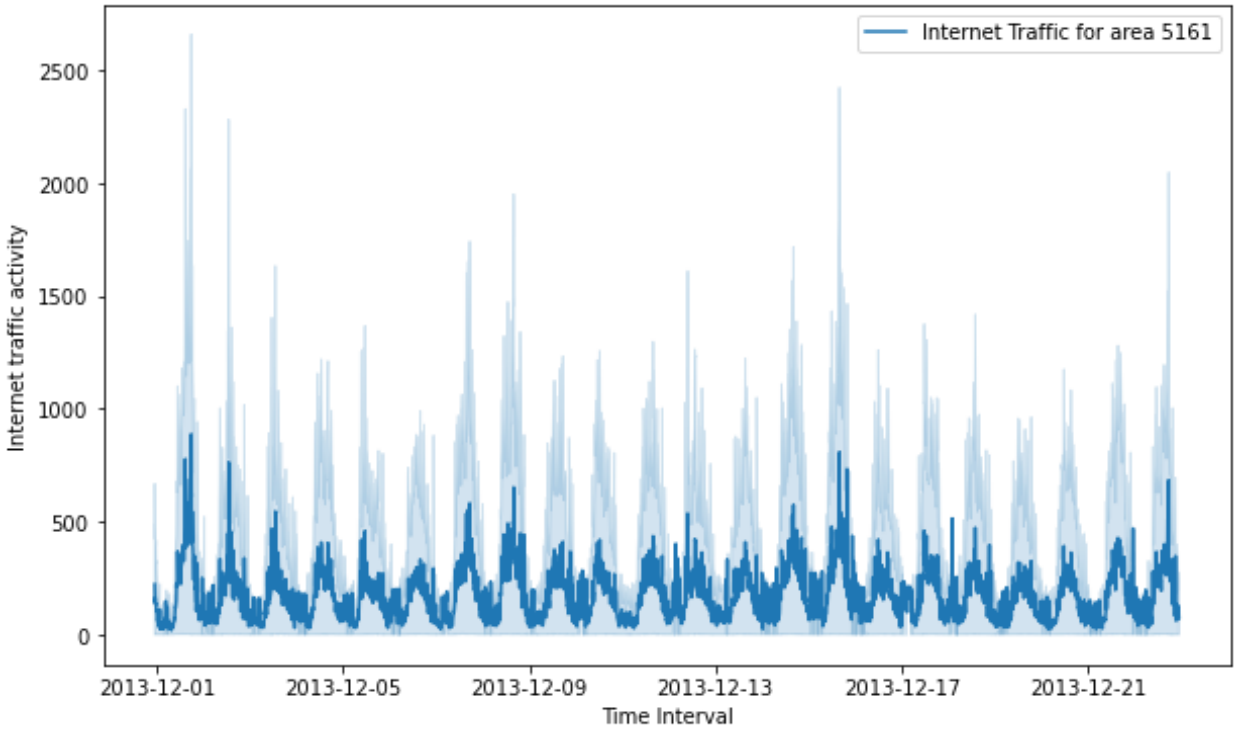




From this time series, we observe that the area 5161 has a a high amount of internet traffic compared to the other two. Area 5161 could be a highly populated area in Milan, or could have a higher technological development.

Task II

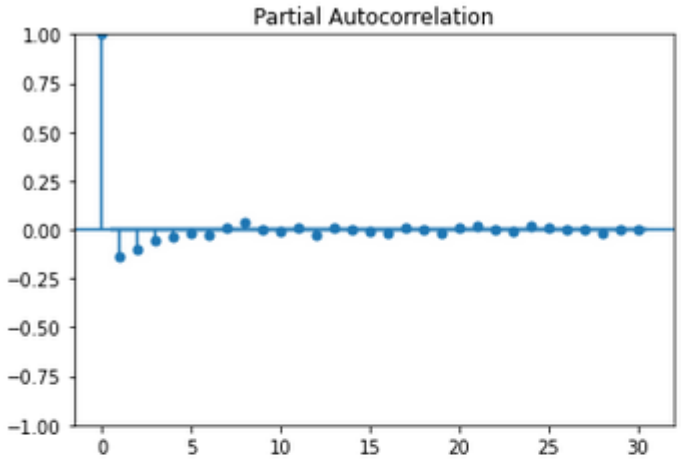
4) To build the algorithm, we first determine the seasonality and stationarity of the time series relative to each area to be examined:
   For area 5161

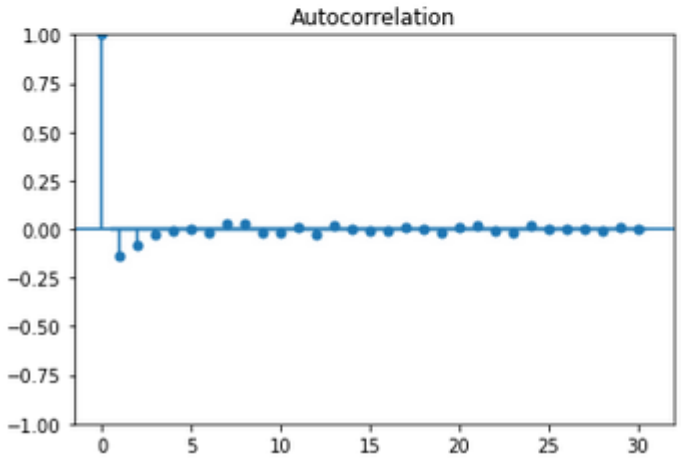


The plot of the internet traffic in area 5161 during the time interval shows seasonality, which is consistent over every 24 hours.
Using the Dickey-Fuller test results, we can say that the data is stationary (test-statistics < critical value).

```
test_stat             -7.057623e+00
p-value                5.324515e-10
lags                   4.800000e+01
num_observations       2.494800e+04
Critical Value (1%)   -3.430612e+00
Critical Value (5%)   -2.861656e+00
Critical Value (10%)  -2.566832e+00
dtype: float64
```

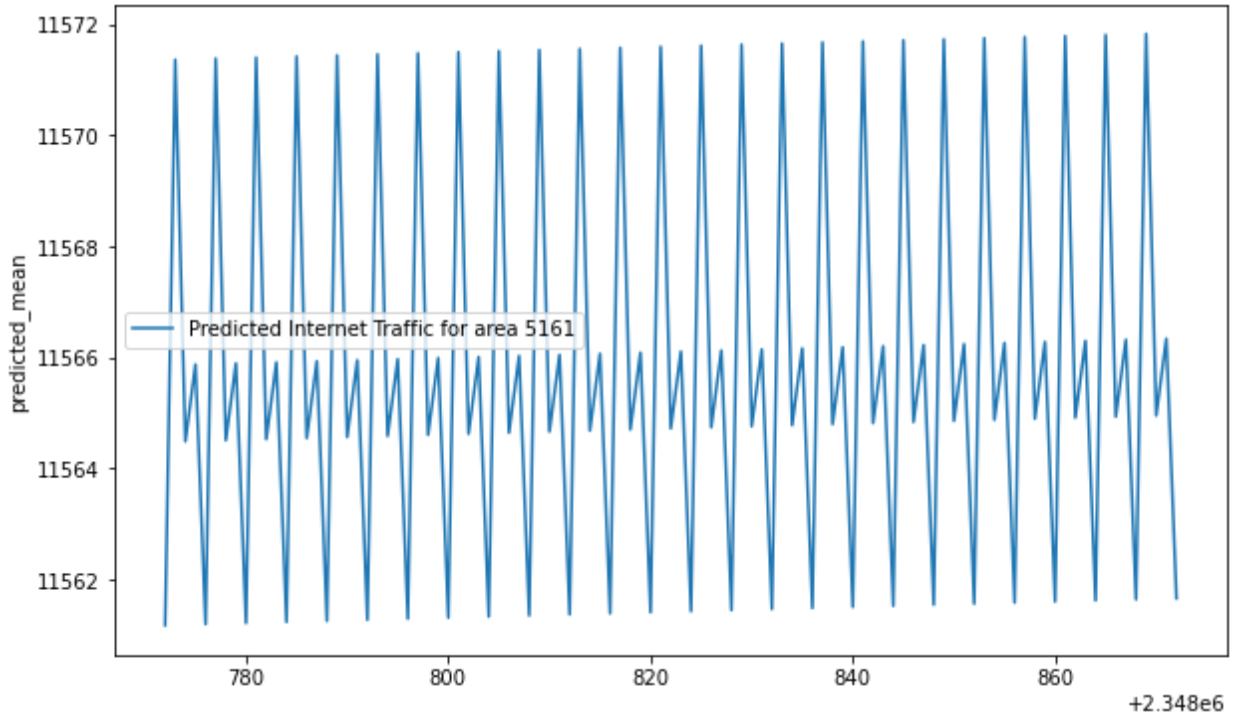
We can therefore build the algorithm and predict the values.
The autocorrelation and partial autocorrelation function graphs gives us the values of p, d, and q.



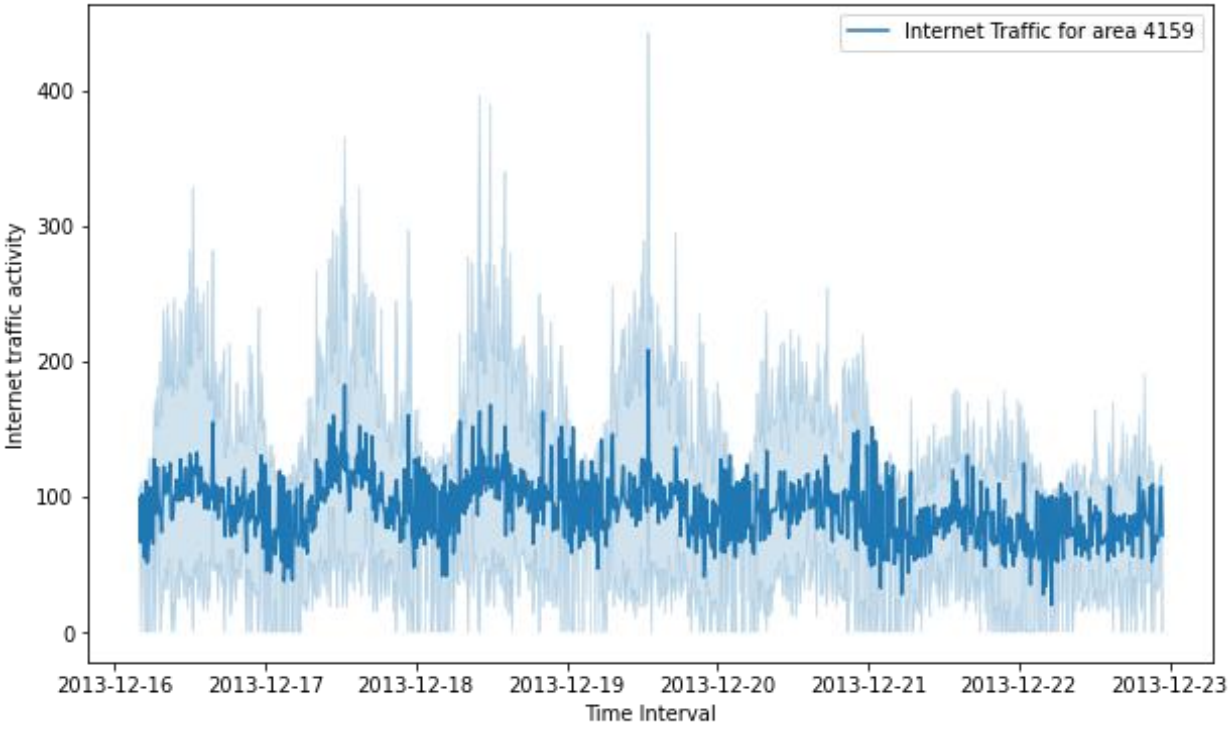


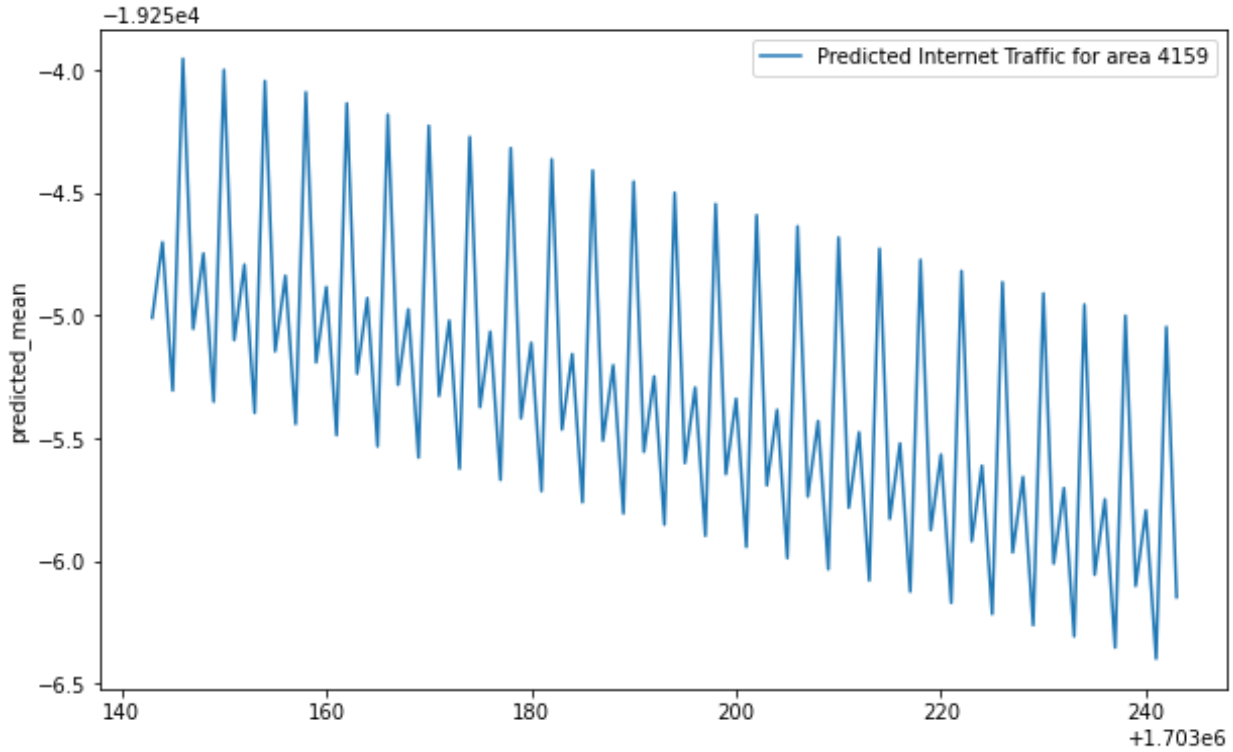
We use the SARIMAX model because it is well suited for predicting seasonal time series. For the algorithm, the frequency m, should be 24 (hours). However, trying to compute it with the present hardware used for the assignment proved challenging. We used a much smaller value, 4.

The predicted traffic for area 5161 for the specified time period is seen on the graph:

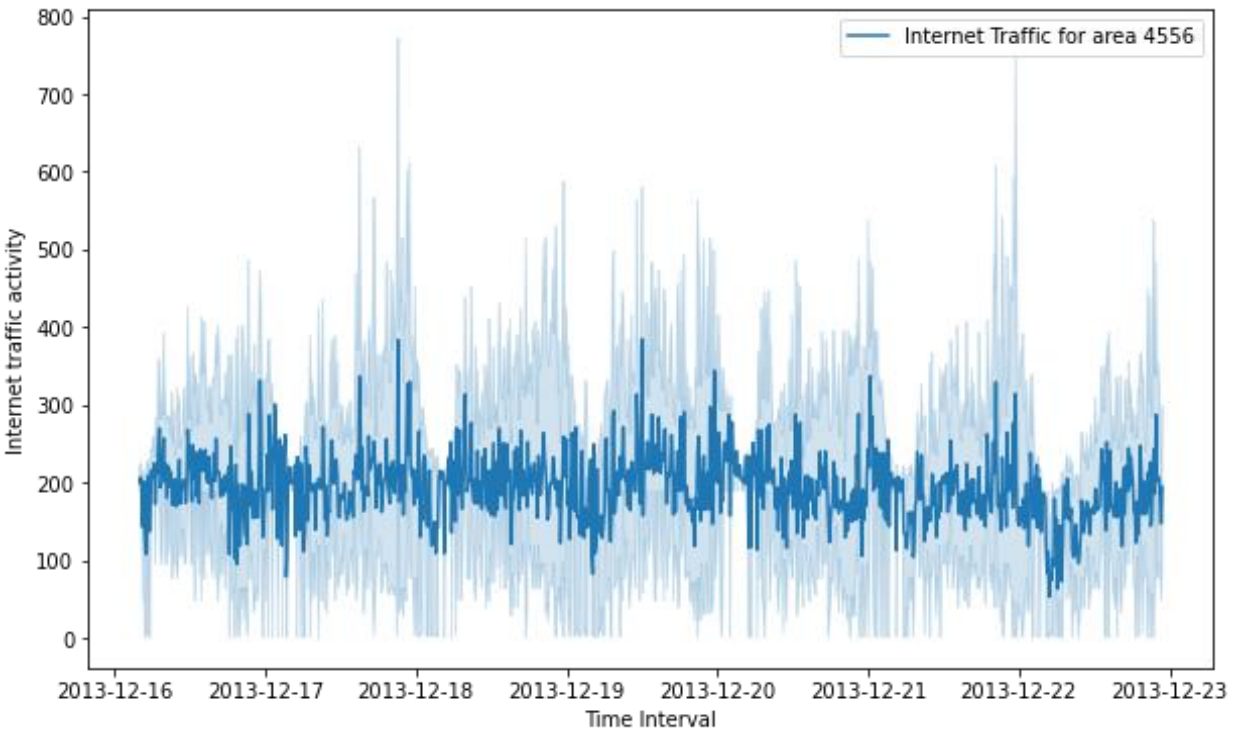


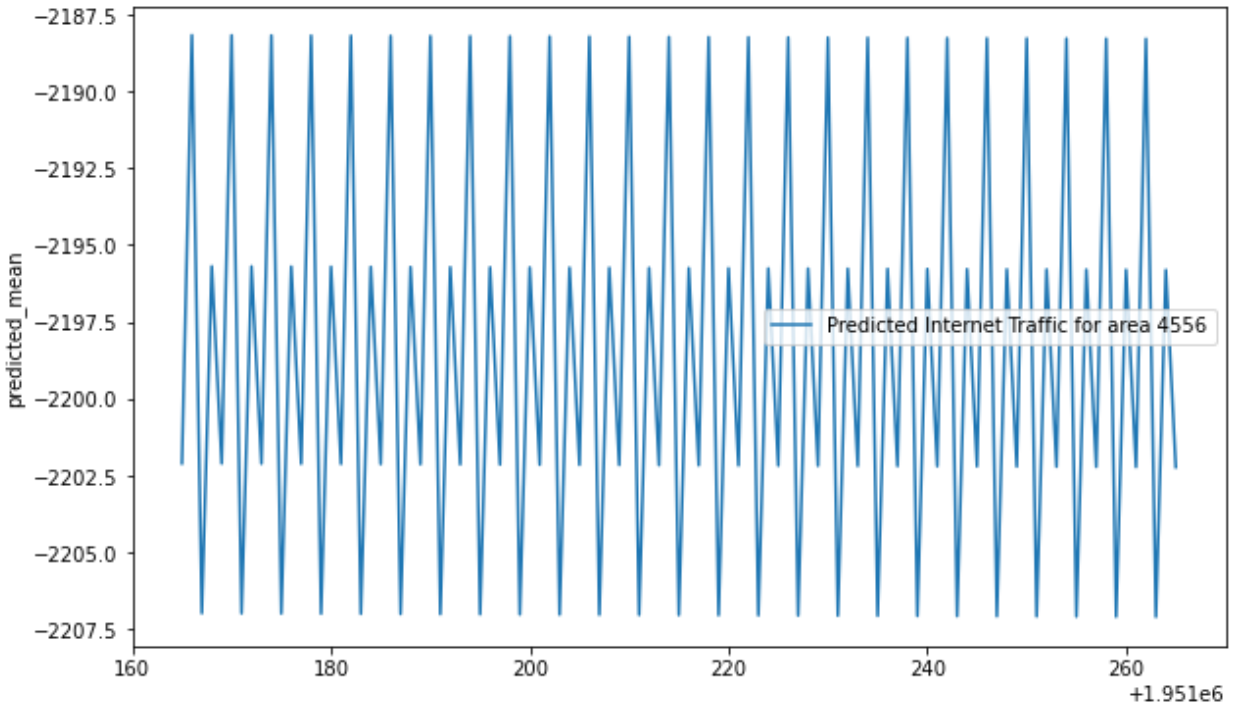
The same procedure is used for the other two areas.
For area 4159, the actual and predicted traffic are as follows:





For area 4556:




predicted_mean
−2187.5
−2190.0
−2192.5
−2195.0
−2197.5
−2200.0
−2202.5
−2205.0
−2207.5
160
180
200
220
240
260
+1.951e6
Predicted Internet Traffic for area 4556