Markov-based Time Series Modeling

for Temperature State of Traffic-Network

ABSTRACT

A Markov model is used to estimate the probability that one state transits to another state after a given time period . In Markov Chain we predict the position of future state based only on the present state(most recent state). Markov models can be an effective way of prediction in time series. In this project, a Markov based time series model is built to analyse the state of temperature of "Metro Interstate Traffic Volume" data set taken from the UCI ML repository.

DATA DESCRIPTION

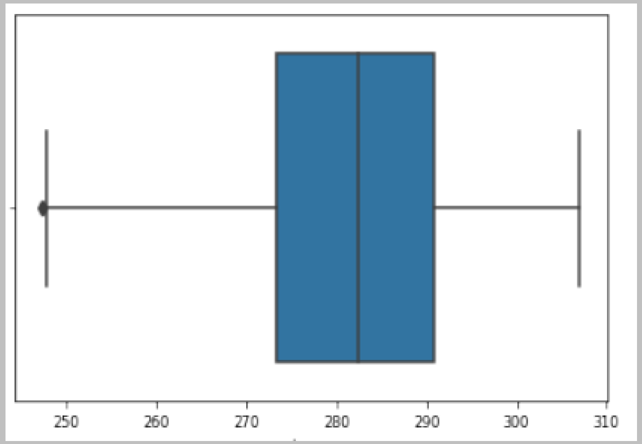
This dataset talks about traffic network conditions by integrating archived and real-time data under various external conditions, including holiday, temperature, cloud coverage and weather conditions between the year of 2012 and the year of 2018.

	A	B	C	D	E	F	G	H	I	J
1	holiday	temp	rain_1h	snow_1h	clouds_al	weather_m	weather_d	date_time	traffic_volume	
2	None	288. 28	0	0	40	Clouds	scattered	10/2/2012 9:00	5545	
3	None	289. 36	0	0	75	Clouds	broken cl	10/2/2012 10:00	4516	
4	None	289. 58	0	0	90	Clouds	overcast	10/2/2012 11:00	4767	
5	None	290. 13	0	0	90	Clouds	overcast	10/2/2012 12:00	5026	
6	None	291. 14	0	0	75	Clouds	broken cl	10/2/2012 13:00	4918	
7	None	291. 72	0	0	1	Clear	sky is cl	10/2/2012 14:00	5181	
8	None	293. 17	0	0	1	Clear	sky is cl	10/2/2012 15:00	5584	
9	None	293. 86	0	0	1	Clear	sky is cl	10/2/2012 16:00	6015	
10	None	294. 14	0	0	20	Clouds	few cloud	10/2/2012 17:00	5791	
11	None	293. 1	0	0	20	Clouds	few cloud	10/2/2012 18:00	4770	
12	None	290. 97	0	0	20	Clouds	few cloud	10/2/2012 19:00	3539	
13	None	289. 38	0	0	1	Clear	sky is cl	10/2/2012 20:00	2784	
14	None	288. 61	0	0	1	Clear	sky is cl	10/2/2012 21:00	2361	
15	None	287. 16	0	0	1	Clear	sky is cl	10/2/2012 22:00	1529	
16	None	285. 45	0	0	1	Clear	sky is cl	10/2/2012 23:00	963	
17	None	284. 63	0	0	1	Clear	sky is cl	10/3/2012 0:00	506	
18	None	283. 47	0	0	1	Clear	sky is cl	10/3/2012 1:00	321	
19	None	281. 18	0	0	1	Clear	sky is cl	10/3/2012 2:00	273	
20	None	281. 09	0	0	1	Clear	sky is cl	10/3/2012 3:00	367	
21	None	279. 53	0	0	1	Clear	sky is cl	10/3/2012 4:00	814	
22	None	278. 62	0	0	1	Clear	sky is cl	10/3/2012 5:00	2718	
23	None	278. 23	0	0	1	Clear	sky is cl	10/3/2012 6:00	5673	
24	None	278. 12	0	0	1	Clear	sky is cl	10/3/2012 8:00	6511	
25	None	282. 48	0	0	1	Clear	sky is cl	10/3/2012 9:00	5471	
26	None	291. 97	0	0	1	Clear	sky is cl	10/3/2012 12:00	5097	
27	None	293. 23	0	0	1	Clear	sky is cl	10/3/2012 13:00	4887	
28	None	294. 31	0	0	1	Clear	sky is cl	10/3/2012 14:00	5337	

DATA PREPRATION

- In this step temperature column was chosen to analyse, and hourly data points of year 2017 were selected. At this stage data points were visualized in a box plot , temperature under 247.36 and over 316 is considered as an outlier. IQR approach was used to find outliers and corresponding rows were removed.

$$Upper = Q3 + 1.5 \times IQR$$
$$Lower = Q1 - 1.5 \times IQR$$




```
Old Shape:  (10581, 2)
New Shape:  (10579, 2)
```

DATA MAPPING

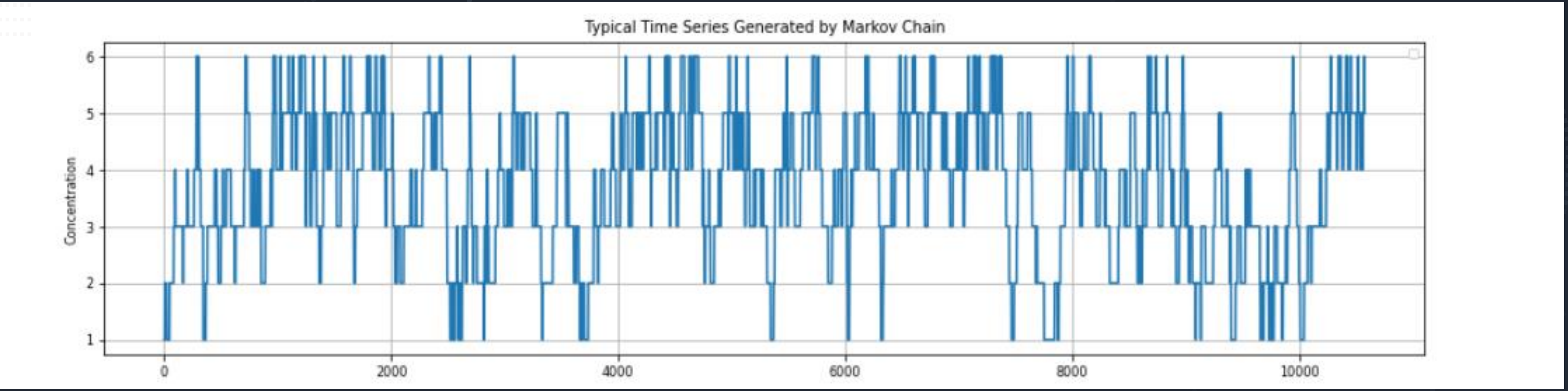
- In order to map our data into a Markov chain, we first calculate the minimum and maximum and then get the range of our data points which is approximately 59.23. Therefore, we decide to choose 6 states and allocate each temperature to a specific state.
- Associated computation :

```
df_t17['states'] = df_t17['temp'].apply(lambda x: ((x-min_value)//10))
```

STATE	1	2	3	4	5	6
TEMP	248<T<258	258<T<268	268<T<278	278<T<288	288<T<298	298<T<308

Building the Markov Chain

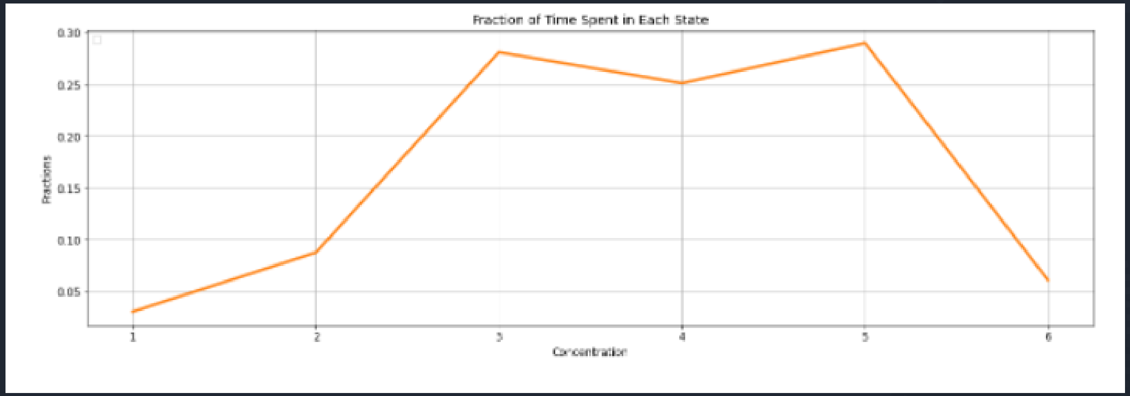
- The Markov states are {1,2,3,4,5,6} . Here is a time series graph of the 10000 datapoints taken from 2017’s data.



Empirical distribution from time series

- Next step is to compute the occupation frequencies for each state and turn this into a probability distribution. This is the empirical distribution of our chain.
- Empirical distribution from time series is the fraction of time spent in each state.

number	frequency
320	0.030249
924	0.087343
2972	0.280934
2655	0.250969
3066	0.289819
642	0.060686



Computation of Transition Matrix and Stationary Distribution

Transition Matrix :

To compute the transition matrix, which is essentially a representation of probability that our Markov state goes from one state to another, for this time series which follows the above distribution. We got the following transition matrix as a result

```
[[0.94984326 0.05015674 0.         0.         0.         0.        ]
 [0.01839827 0.93939394 0.04220779 0.         0.         0.        ]
 [0.         0.01345895 0.96433378 0.02220727 0.         0.        ]
 [0.         0.         0.02485876 0.93408663 0.04105461 0.        ]
 [0.         0.         0.         0.03555121 0.93868232 0.02576647]
 [0.         0.         0.         0.         0.12305296 0.87694704]]
```

Stationary Distribution

To compute the limiting probability vector a.k.a the stationary distribution which tell us the long term probability distribution of our markov states.

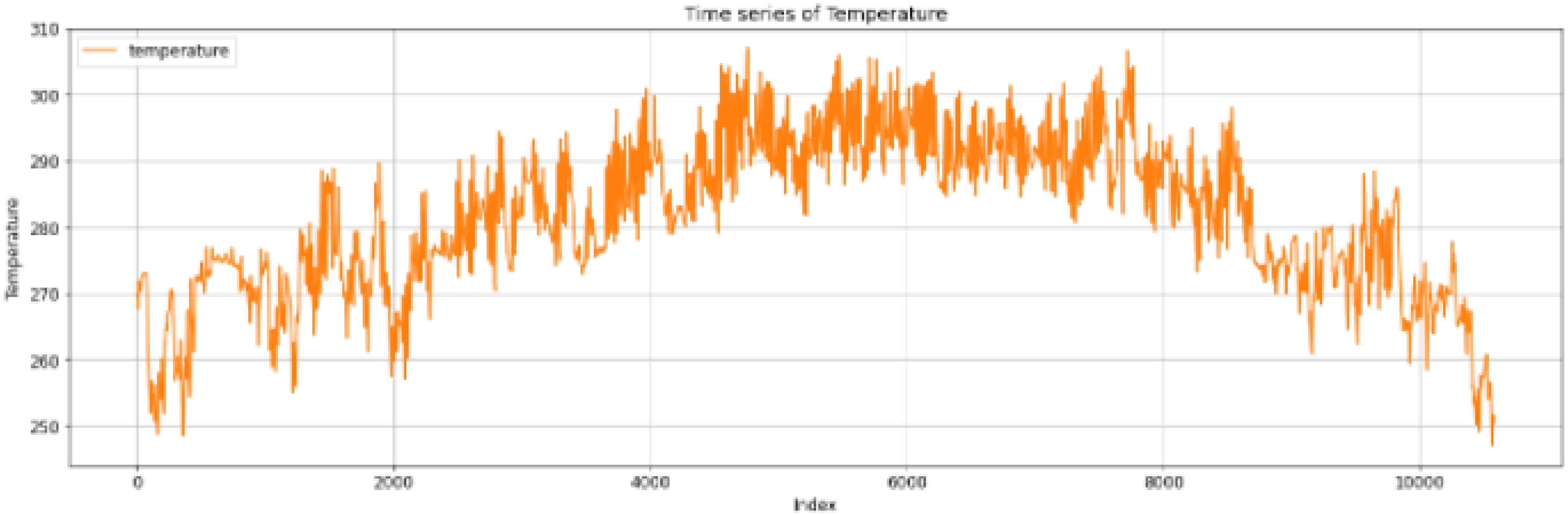
Let W be the stationary vector

W=(w1,w2,w3,w4,w5,w6)

```
([0.03270157, 0.08914993, 0.27957765, 0.24975729, 0.28842028,
  0.06039329])
```

Time Series Plot

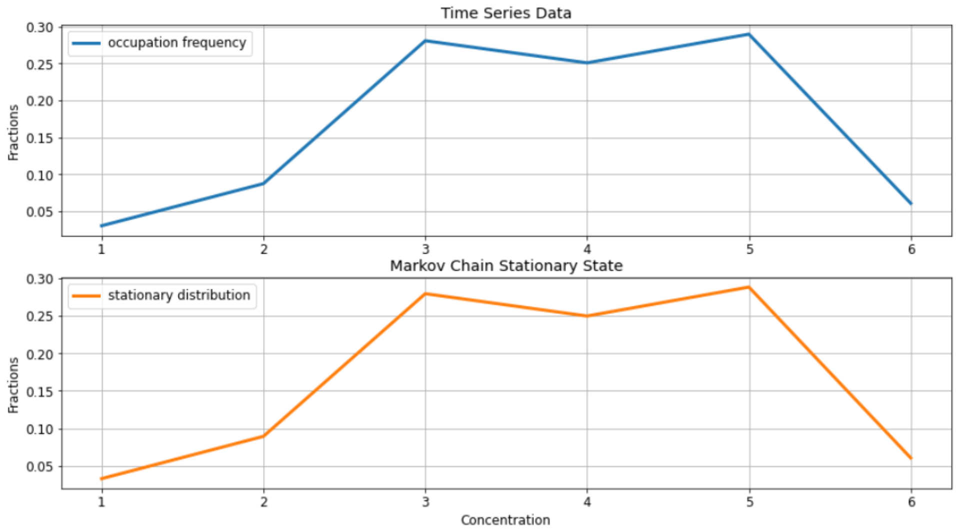
- In this step, considering 10000 datapoints of Year 2017 I investigated the time series plot of the original data. Here x-axis is the index of each data point (corresponding to hourly stamps)and the y-axis is the value of temperature.



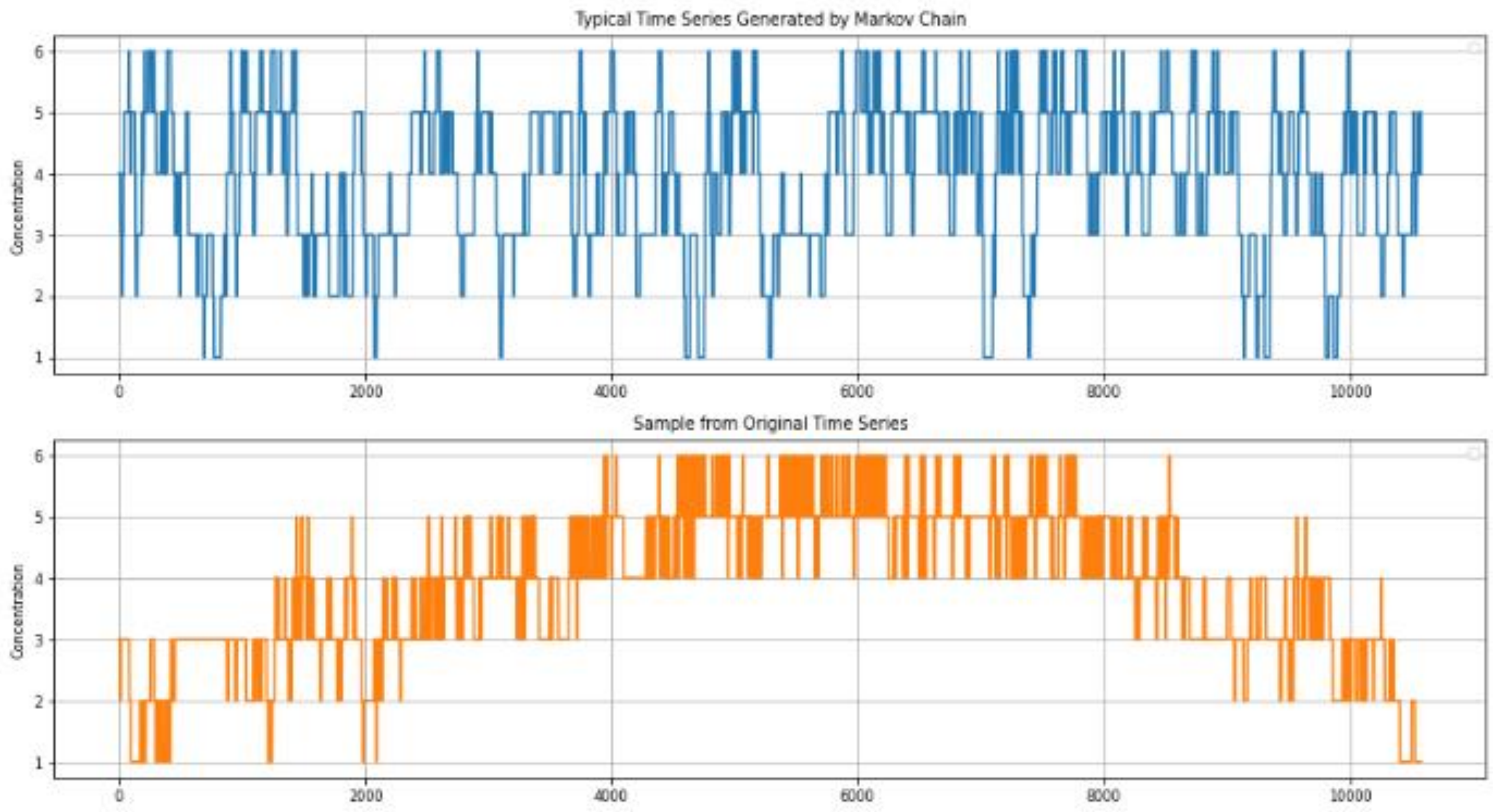
Comparing the empirical distribution of the data set and the stationary distribution of the chain

- Empirical distribution from time series compared to the
  stationary distribution of chain

- The plot of the Markov State vs Normalized Frequency for original time series is similar which is an indication that the frequency distribution of the original data wasn't disturbed.



Comparing Time Series Plot with our simulated Time series



Evaluation for this Model

- Auto-correlation function:

Auto correlation is the correlation between two observations at different points in a time series.

The correlation values for original time series and simulated time series are given in the following table: (we took K=30 , this is a sample of 6 values)

$$\begin{bmatrix} 1.0 & 1.0 \\ 0.978987377545787 & 0.981591011906189 \\ 0.95971090973076 & 0.963998218356827 \\ 0.941636395127476 & 0.947105020131268 \\ 0.924096081951632 & 0.930736518398575 \\ 0.907557396452242 & 0.915184211210319 \end{bmatrix}$$

$$R(k) = \frac{\sum_{i=1}^{N-k}(X_i - \overline{X})(X_{i+k} - \overline{X})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}, k = 1, 2, ..., 30$$

Goodness Of Fit using Chi Test

Using Chi square test, I calculated Chi values and then corresponding P values . P-values for each Markov state is given here ->

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared

O_i = observed value

E_i = expected value

[0. 014234875444839857,
6. 067932897249324e-29,
0. 0013477088948793193,
0. 0,
6. 583736171227538e-29,
1. 718709716907055 7e-29]

Conclusion

- Since all the p-values are less than 0.05(significance level) we will reject the null hypothesis that our dataset is a good fit for Markov model. Hence, we will reject all the states. From the previous table, we see that none of the states can be accepted by our hypothesis. Thus, I do not consider that the Markov chain method produces a good model for this time series.

Thank you!