10/10 Truly excellent!

Math 7241-Markov Chain Project: Modeling Wind Speed Data

Josh Galloway (galloway.j@northeastern.edu),
Northeastern University

December 2020

1 Introduction

Working from publicly available data weather station data, the student constructed a finite-state Markov chain to model average daily wind speed at Boston Logan Airport [NCfEIN20]. The model was tested via comparison of autocorrelations from the model simulation and original dataset and against its two-step state prediction by χ^2 test. Results were mixed but overall the model produced a reasonable-looking simulation. However, the cyclical nature of the data makes it challenging to produce a high quality model from a Markov chain.

2 Dataset

The dataset is comprised of average daily wind speed from weather station USW00014739 located at in Boston, MA at Logan International Airport. The data ranged from 01 January 2000 to 01 January 2020 with a total data count of 7306 measurements with mean of 10.936 miles-per-hour (MPH), standard deviation of 3.759 MPH, minimum of 2.24 MPH and maximum of 38.03 MPH. The raw data is presented in Figure [1]. The data was cleaned to exclude the measurements over 27 MPH as they are readily observed as outliers. This filtering only removed 7 data points and gives a final data count of 7299 measurements. Overall, dataset was very well constructed and contained no missing values or errors.

3 Model Derivation

Starting with the cleaned and conditioned dataset, several steps followed to build a working model. The data was first read into a Python, Jupyter notebook environment for

Josh, this is the best report I have seen in several years! It is awesomely good. The project was limited in scope but clearly your work could be pushed much further, for example (as you suggest) building a model which is a sum of a deterministic piece (for long-term trends) and a stochastic piece for short-term fluctuations. Let me know if you would be interested in pursuing this as an 'extra-curricular' activity!

Well done!!

Z.

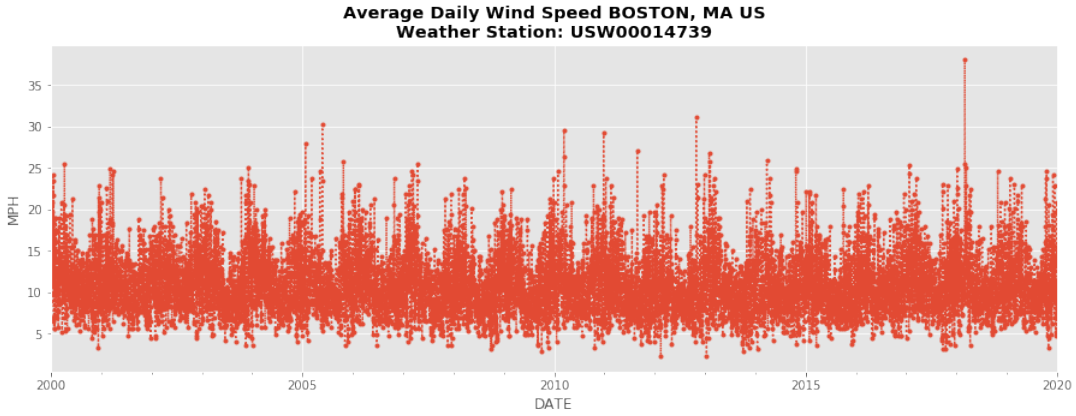

Figure 1: Raw Data

processing. Following this, a histogram was constructed with 10 bins representing the quantization of the original continuous data for representation as a 10-state Markov chain. The number of states selected was done in a somewhat ad hoc fashion as application of Sturge's Rule suggested a state size of 8 but the results produced a distribution with nearly 40% of the data in one bin, so the state space size was increased to 10. Figure 2 shows the results of the quantization process and the empirical distribution/occupation frequencies resultant from the step.

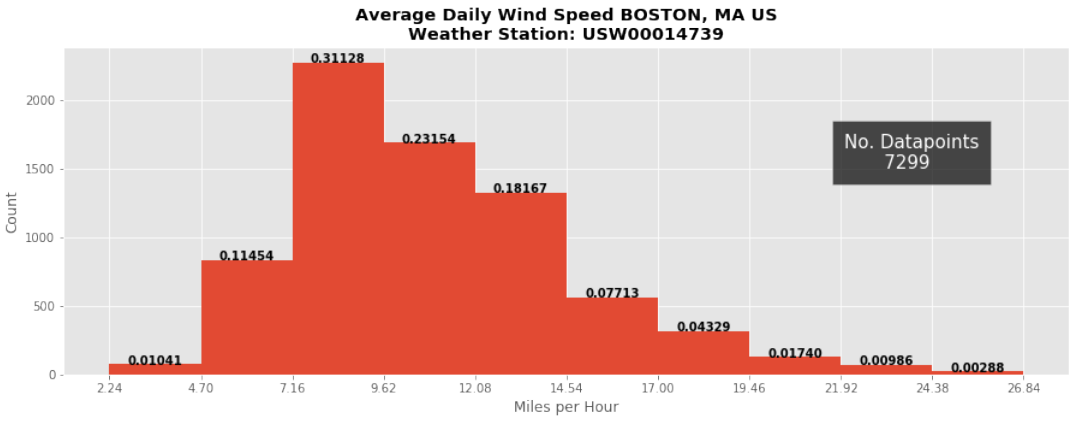

Figure 2: Empirical Distribution and Occupation Frequencies

After the data was quantized into states, the single-step transition frequencies were counted and normalized such that the total jump probability out of each state sums to 1. This was then tabulated into the model transition matrix (P) and is shown in Table 1. Figure 3 shows the network graph for the transition matrix.

The transition matrix was then multiplied by itself repeatedly and, at each multiplication, compared to previous iteration by taking the Frobenius norm of $P^{(n-1)} - P^n$ to ascertain when the stationary distribution was reached. Upon calculation of P^6, the result of the stopping calculation was less than the chosen tolerance of 10^{-12} and the stationary distribution was produced. Comparing the stationary distribution to the empirical distribution shows that the two are roughly equal as compared in Table 2. However, there is some significant deviation between the two in states 2-6 but only on the magnitude of hundredths or thousandths.

(i, j)	1	2	3	4	5	6	7	8	9	10
1	0.0921	0.2500	0.2237	0.1711	0.1447	0.0395	0.0658	0.0132	0.0000	0.0000
2	0.0347	0.2309	0.3218	0.2141	0.1065	0.0490	0.0311	0.0096	0.0024	0.0000
3	0.0093	0.1557	0.3364	0.2628	0.1375	0.0584	0.0206	0.0108	0.0064	0.0020
4	0.0052	0.0909	0.3115	0.2799	0.1838	0.0748	0.0265	0.0145	0.0104	0.0026
5	0.0040	0.0659	0.2468	0.2937	0.2087	0.0992	0.0548	0.0159	0.0087	0.0024
6	0.0080	0.0462	0.1529	0.2914	0.2182	0.1449	0.0892	0.0255	0.0207	0.0032
7	0.0032	0.0411	0.1203	0.2120	0.2595	0.2025	0.1139	0.0316	0.0063	0.0095
8	0.0000	0.0551	0.1181	0.2047	0.2283	0.1732	0.1260	0.0472	0.0472	0.0000
9	0.0000	0.0000	0.0563	0.1127	0.1831	0.2254	0.1549	0.1831	0.0423	0.0423
10	0.0000	0.0000	0.0455	0.2273	0.0909	0.1818	0.1818	0.1364	0.0455	0.0909

Table 1: Model Transition Matrix

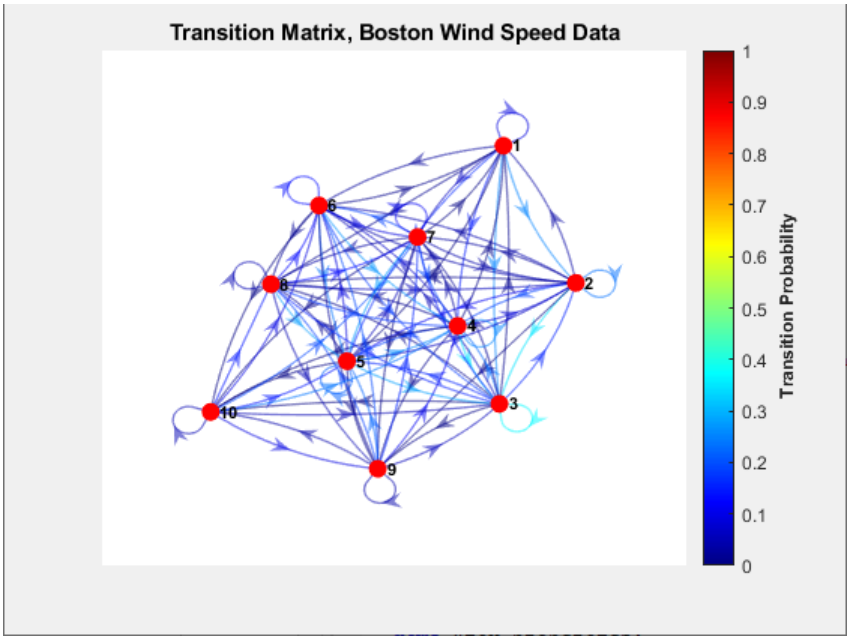

Figure 3: Network Graph of Transition Matrix

State	Bins	Empirical Distribution	Stationary Vector	Absolute Error
1	[0.0, 4.7]	0.0104	0.0104	1.3294e-06
2	[4.7, 7.16]	0.1145	0.1145	4.4789e-06
3	[7.16, 9.62]	0.3113	0.2790	3.2323e-02
4	[9.62, 12.08]	0.2315	0.2638	3.2234e-02
5	[12.08, 14.54]	0.1817	0.1727	9.0102e-03
6	[14.54, 17.0]	0.0771	0.0862	9.0677e-03
7	[17.0, 19.46]	0.0433	0.0433	1.7240e-05
8	[19.46, 21.92]	0.0174	0.0174	4.8283e-06
9	[21.92, 24.38]	0.0099	0.0097	1.3381e-04
10	[24.38, 27.0]	0.0029	0.0030	1.3772e-04

Table 2: Comparison of Empirical and Stationary Distributions

4 Simulation

Utilizing the calculated single-step transition matrix, an new simulated dataset was produced. This was done by randomly generating a number $\in (0, 1]$ and using the step probabilities in the current state to set a range in which the random number falling selects the next state. For example, suppose the chain were in state 1, and the random number 0.3 was generated. State 1's transition probabilities are 0.0921 for jumping back to State 1 and 0.25 for jumping to State 2. The range for jumping back to State 1 is $(0, 0.0921]$, and the range for jumping to State 2 is $(0.0921, 0.25 + 0.0921] = (0.0921, 0.3441]$, so in this case, the random number of 0.3 would cause the simulation to jump to State 2 on the next step. The resultant simulation is compared to the original time series data quantized into states in Figure [4]. Qualitatively, the Markov chain model simulation produces a reasonable representation of the original dataset although it is apparent that the cyclical peaks and troughs are not found in the simulation as they are in the original data.

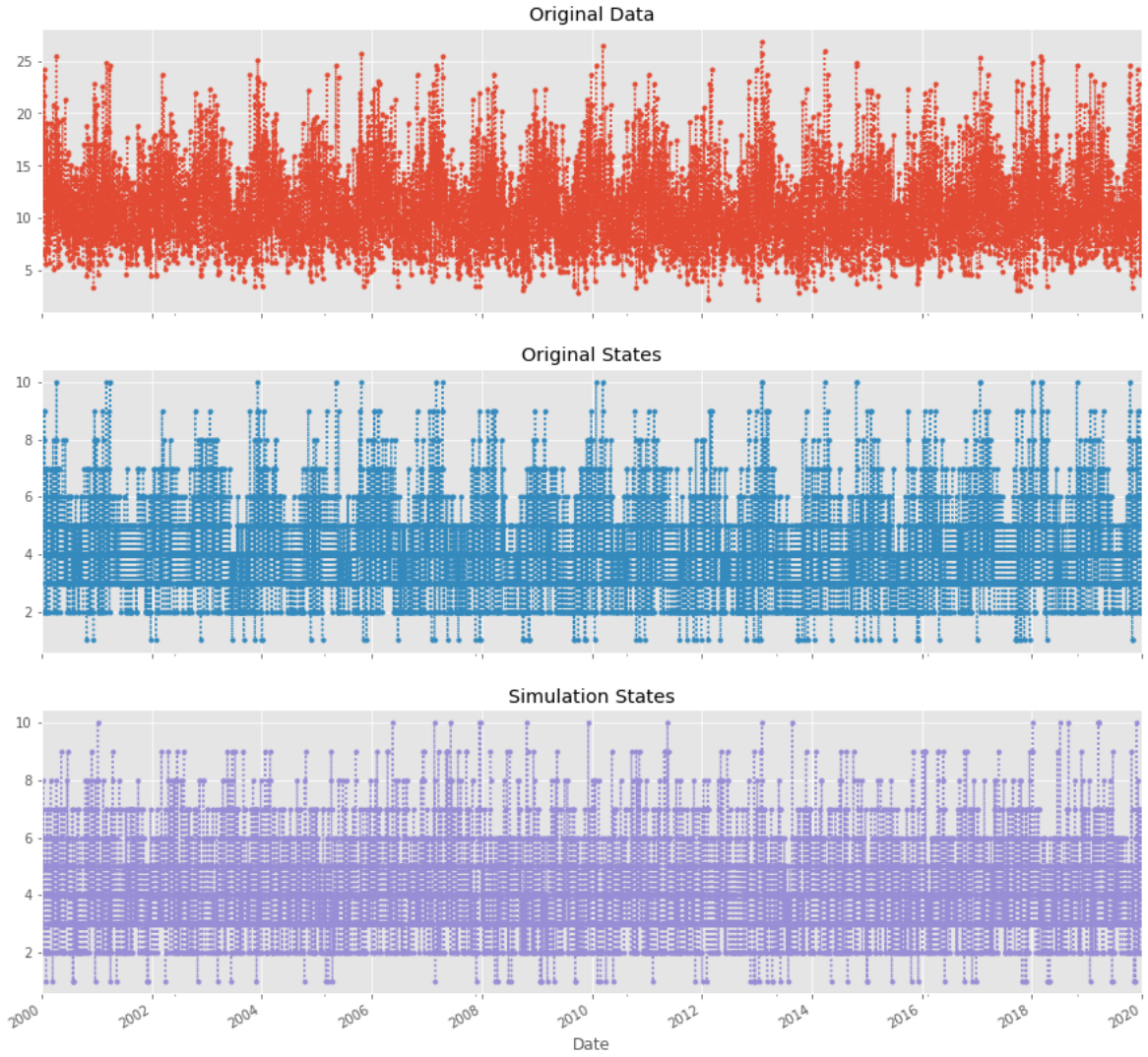

Figure 4: Comparison of Original Time Series and Simulation

5 Model Evaluation

To evaluate the model and resulting simulation, the autocorrelation function (ACF) was calculated for both series. The results are shown in Figure 5. The original data has nonzero lag correlations throughout the range of the ACF. This is expected as the data is clearly cyclical. Fitting a simple sinusoid via taking a Fast-Fourier transform (FFT), fitting the data and converting back to time domain, reveals that the data cycles with period of 1 year (as would be expected), peaks in late January, and troughs in late August. Figure 6 shows this relationship. However, the Markov chain simulation has ACF correlations of practically zero for all values. This is also expected as by definition the Markov chain is only dependent on the previous step's state so the ACF should be close to zero after just one lag. This is a fundamental flaw in trying to represent this dataset with a Markov chain, as the wind speed at Boston Logan is clearly not a Markov process.

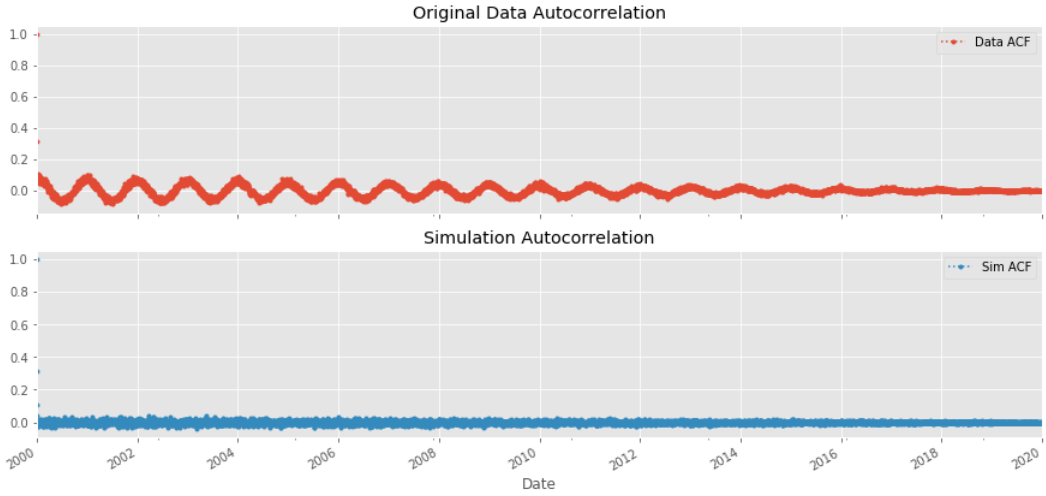

Figure 5: Autocorrelation Function of Original Time Series and Simulation

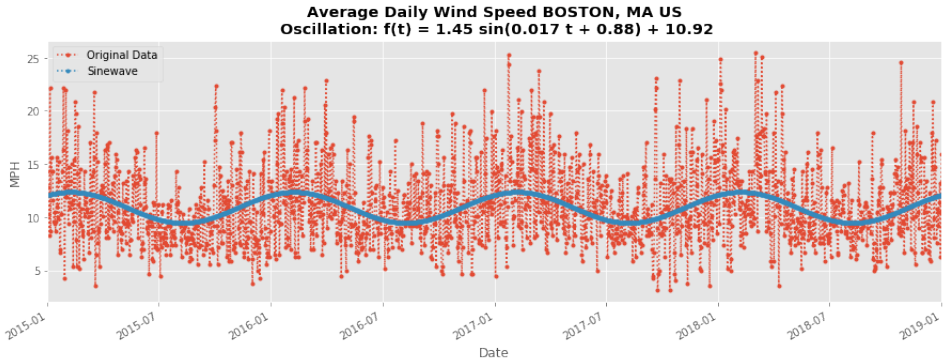

Figure 6: Sinusoid Fit to Original Data

Following the autocorrelation function inspection, a χ^2 test was performed on the the two-step original dataset frequencies versus the expected two step frequencies produced by the model. The model two-step transition matrix was calculated, P^2, and the total residence frequency of each state was multiplied though each row to produce the expected

two-step transition frequencies. The result of this calculation and the tabulation of the two-step transition frequencies for the original data are shown in Table 7. The results of this exercise were used to calculate the Pearson's goodness of fit test statistic and used to perform a χ^2 test with null hypothesis that the model is a good fit for the data at a 5% significance level for each state in the model. The results are shown in Table 3. Only State 2 failed the test and is shown along side the original dataset's two-step transition frequencies in Figure 8 for comparison.

Original Data Two-Step Transition Frequencies

(i,j)	1	2	3	4	5	6	7	8	9	10
1	0.0	11.0	18.0	17.0	14.0	6.0	5.0	3.0	1.0	1.0
2	13.0	120.0	235.0	203.0	116.0	76.0	55.0	11.0	7.0	0.0
3	19.0	244.0	634.0	553.0	340.0	134.0	65.0	26.0	15.0	6.0
4	18.0	220.0	526.0	531.0	344.0	155.0	69.0	42.0	16.0	5.0
5	11.0	128.0	332.0	336.0	222.0	130.0	53.0	24.0	15.0	8.0
6	7.0	57.0	157.0	164.0	127.0	61.0	38.0	9.0	7.0	1.0
7	3.0	34.0	79.0	69.0	61.0	37.0	17.0	9.0	6.0	1.0
8	3.0	15.0	32.0	29.0	19.0	18.0	7.0	1.0	3.0	0.0
9	2.0	7.0	17.0	17.0	11.0	9.0	5.0	2.0	1.0	0.0
10	0.0	0.0	5.0	6.0	6.0	3.0	2.0	0.0	0.0	0.0

Expected Values of Two-Step Transtion

(i,j)	1	2	3	4	5	6	7	8	9	10
1	1.6133	11.0884	21.3419	18.7400	12.2395	5.8856	3.3119	1.1031	0.5260	0.1504
2	13.5676	119.2197	247.2743	212.5220	131.0995	61.5885	31.0089	11.7617	6.2908	1.6670
3	24.0963	259.9493	598.9531	533.2640	331.9431	157.2518	75.6116	31.2182	17.6232	5.0895
4	18.0926	218.7902	548.2240	513.2673	330.5382	160.8425	78.2815	32.6769	18.6264	5.6604
5	10.4203	129.4902	343.1634	338.1953	227.2601	114.7767	57.3023	22.5105	12.8928	3.9884
6	4.7655	56.0212	157.1808	167.5864	120.0349	65.2007	34.5021	13.7370	7.4248	2.5465
7	2.1944	25.6446	73.8462	85.0064	62.4111	35.3358	19.4420	6.8413	3.9313	1.3467
8	0.8587	10.2291	28.8905	32.9401	25.0600	14.8608	8.2738	3.5163	1.7294	0.6415
9	0.2927	4.1140	12.6562	17.7833	14.8694	10.1414	6.3342	2.6942	1.5002	0.6145
10	0.0877	1.2564	3.9818	5.4444	4.4345	3.1291	1.9715	0.9413	0.4588	0.2945

Figure 7: Two-Step Transition Frequencies

State	**Test Statistic**	**P-Value**	$\chi^2_{m-1,1-a}$	**Results**
1	11.9048	0.2187	16.919	Accept Null
2	26.5339	0.0017	16.919	Reject Null
3	11.3861	0.2502	16.919	Accept Null
4	6.4891	0.6901	16.919	Accept Null
5	7.3686	0.5988	16.919	Accept Null
6	4.7682	0.8540	16.919	Accept Null
7	8.6676	0.4685	16.919	Accept Null
8	14.0716	0.1198	16.919	Accept Null
9	15.8874	0.0693	16.919	Accept Null
10	3.9141	0.9170	16.919	Accept Null

Table 3: Comparison of Empirical and Stationary Distributions

Original Data Rejected States

State	1	2	3	4	5	6	7	8	9	10
2	13.0	120.0	235.0	203.0	116.0	76.0	55.0	11.0	7.0	0.0

Model Rejected States

	1	2	3	4	5	6	7	8	9	10
2	14.0	119.0	247.0	213.0	131.0	62.0	31.0	12.0	6.0	2.0

Figure 8: Comparison of Rejected States

6 Conclusion

Overall, the model performed well as evinced by χ^2 test. There were, however, some inaccuracies due to the lack of fidelity in the model with respect to the cyclical pattern in the original data which was most noticeable in comparing the autocorrelation functions. Since by definition and design the Markov chain cannot represent these long-term correlations, it was not able to accurately model the data over long periods of time. As previously mentioned, this is due to the data not being dependent exclusively on the previous state as would be required for it to be a Markov process. However, the short-term predictions were very accurate for all but State 2. This short-term evaluation minimizes the affect of the cycling in the data, and explains the difference in results for the two tests. Thus, the model is a very good representation for predicting wind speed in the short-term such as a few days from the current state, but a poor representation for trying to predict wind speed several months from the current state. Objectively, a model such as this would likely be used only to make predictions in the short-term, and in such a case would be well suited to the purpose. Increased accuracy may be obtainable by first subtracting out the modeled cyclical term in the data, building a Markov chain representation from the resultant dataset and then adding back the cyclical term following simulation of the chain. However, that is beyond the scope of this project but could be considered a future work.

References

[NCfEIN20] NOAA: National Centers for Environmental Information (NCEI). Climate data online-order history, 2020. Weather Station-GHCND:USW00014739 data range 2000-01-01 00:00 to 2020-01-01 23:59.