Adaptations of Markov Chains for Weather Forecasting

Ricardo Trujillo RET@CPP.EDU
California State Polytechnic University Pomona

Abstract

Leveraging its probabilistic nature, Markov Chains present a robust tool for modeling, predicting, and forecasting weather patterns. Through the use of historical data collected by CALMAC, this project explores 3 versions of a Markov Chain: a Traditional Markov Chain, a Second-Order Markov Chain, and a Seasonal Markov Chain. Each model was evaluated based on its ability to capture the probabilistic relationship of past and present weather states. This report highlights the performance capabilities of each model through its respective MSE while addressing potential tradeoffs. With its highly accurate predictive performance, the Seasonal Markov Chain model was used to forecast the unobserved months of 2024 to analyze monthly ratios of each weather state, underscoring the importance for weather preparation in various sectors.

Keywords: Markov Chain, Second-Order Difference Equation, Time Series, Weather Forecasting, Seasonality

1 Introduction

Weather prediction plays a vital role across sectors, including but not limited to agriculture and risk management. Markov Chains, a stochastic process transitioning between states based on probabilities, offers a potent tool for modeling and forecasting weather trends due to their probabilistic features. They can suitably handle the uncertainty of future weather phenomena, whether it be for the next day, week or month.

In this context, the states represent various weather conditions within a defined space, and transitions rely on historical data of meteorological factors. Analyzing the collected data enables a Markov Chain to establish the probabilistic relationships between weather states and estimate their likelihood. In this project, the goal is to leverage the simplicity and efficiency of a Markov Chain to capture weather patterns for the upcoming year.

Furthermore, I propose adaptations to traditional Markov Chain to accommodate for seasonal variations and concurrent weather states. Overall, the Markov Chain framework facilitates understanding of a complex system otherwise known as nature.

2 Data

In this project, historical weather data sourced from CALMAC (1) for Los Angeles spanning 2022 and 2023 was utilized. This dataset encompasses a variety of meteorological factors including wind speed, temperature, and pressure, among others. For the purpose of this analysis, daily averages between the hours of 7:00AM-9:00PM were computed for each variable, representing the typical hours a human experiences.

Specifically, this project focuses on the average values for the following variables: Temperature(F), Rainfall(mm), Sky Cover (tenths), and Pressure(mbar). The following rules set forth the 5 different weather state classification of a given day:

- Rainy: $avg\_rain \geq 1$
- Cloudy: $avg\_sky\_cover \geq 6$
- Chilly: $avg\_temp < 65$
- Partly Cloudy: $avg\_temp \geq 65$ and $avg\_sky\_cover > 5$
- Sunny: $(avg\_temp \geq 65$ and $avg\_sky\_cover \leq 5)$ or $(avg\_pressure \geq 1005)$

2.1 Exploratory Data Analysis

Through the set rules, we can observe a transitioning, polar effect happening in the winter and summer months for Sunny and Chilly Days. Given the drought nature of Los Angeles, it is expected that there are few rainy days — most concentrated in the winter months.

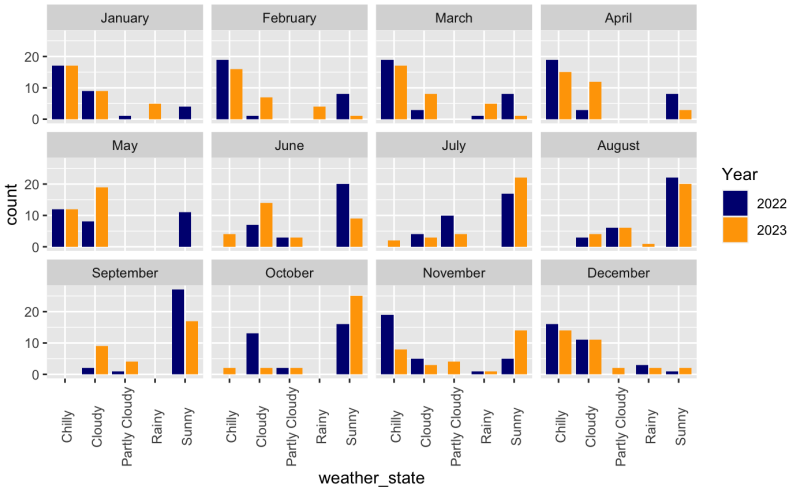

Figure 1: Barchart of Weather States

3 Methodology

Given its flexibility, this report chooses to explore 2 new versions aside from the traditional Markov Chain model: one incorporating the seasonality and the other considering the influence of concurrent days.

3.1 Traditional Markov Chain

The first model explored was the Traditional Markov Chain — formulaically expressed as $x_t = P \cdot x_{t-1}$, where P is the one-step transition matrix[A], x_{t-1} is the one-hot encoded representation of the previous state, and x_t denotes the probability distribution of the 5 weather states [1]. Due to the randomness presented, the algorithm was ran over a total of 50 simulations. Each simulation saved the monthly ratio for each state, summed the ratios, and then were averaged. This resulted in a more stable and consistent prediction that expressed the proportion of days being a particular weather state.

In the corresponding Ratio Matrix [B], the ratio between the known and predicted were similar for the following: Cloudy in January and March, Partly Cloudy for January, and Rainy for April. On the other hand, the model tended to be biased towards Sunny such that the ratios were higher than what was expected for each month. Additionally, the model is not able to capture the ratio for Rainy days for the first 3 months, which are Los Angeles' rainy season. When compared with the expected, or actual, ratios, the model obtains an MSE of 0.1612 or 16%.

3.2 Second Order Markov Chain

Expanding on the idea of the Traditional Markov Chain, this adapted Second-Order Difference equation leverages the idea that the previous two days will influence the

1. [Chilly, Cloudy, Partly Cloudy, Rainy, Sunny]

current weather state directly. It captures the idea of interpreting the likelihood of the weather given the same or a variation of two weather states. For example, if it has been Sunny the last two days, what would be the likelihood that the current day will also be Sunny. Thus, we represent the aforementioned difference equation as the following:

$$x_t = c_1 A x_{t-1} + c_2 B x_{t-2}$$
$$\mathrm{c}_1 + c_2 = 1$$

where Matrix A and B $^{[A]}$ denote the one and two step transition matrices, respectively, and x_{t-2} represents the one-hot encoded representation of the weather states from two time-steps before.

Exploring the different possibility of (c_1, c_2) pairs in Δ changes of 0.1, the following trend expresses $(c_1, c_2) = (0.4, 0.6)$ having the lowest MSE.

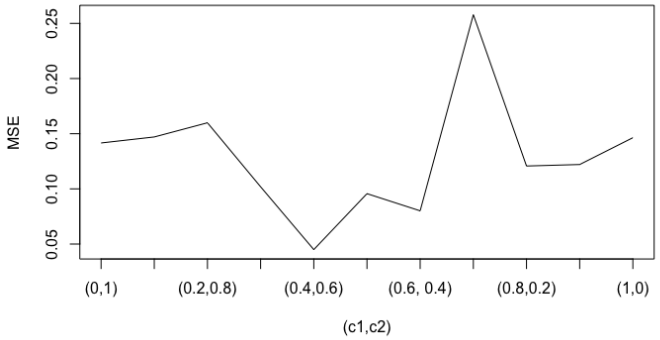

Figure 2: MSE vs (c_1, c_2)

Moving forward with $c_1 = 0.4$ and $c_2 = 0.6$, the model was ran over a total of 50 simulations, where once again the ratios were saved, summed, and averaged for each month. As shown in the corresponding Ratio Matrix $^{[B]}$, the predicted values were similar to the expected values in the following ways: Cloudy for January, March and April, Chilly for February, and Rainy for January. Once again, we see that the second order is still prone to over-predicting Sunny days for these months and its inability to capture the Rainy days for February and March. Then, when compared with the expected, or actual, ratios, the model obtains an MSE of 0.1169 or 12%.

3.3 Markov Chain with Seasons

Lastly, addressing the concern, that we experience 4 different seasons[2] in Los Angeles, each with distinct weather patterns, this adaptation incorporates the effects of a particular day pertaining to a season. While its mathematical expression has not been derived, we can express this idea into its building blocks such that

$$x_t = P_{Winter} \cdot x_{t-1}, \text{ if Month is December, January, February}$$
$$x_t = P_{Spring} \cdot x_{t-1}, \text{ if Month is March, April, May}$$
$$x_t = P_{Summer} \cdot x_{t-1}, \text{ if Month is June, July, August}$$
$$x_t = P_{Fall} \cdot x_{t-1}, \text{ if Month is September, October, November}$$

Note, each seasonal transition matrix $^{[A]}$ was calculated using the respective months. In consideration of how many seasons are in the interested time of exploration, it is important to consider that each matrix will essentially be weighted differently. In other words, it understands that the Summer transition matrix won't have any effect on the output if working with a range of dates between January and April.

With the same setup, this version was ran over a total of 50 simulations where the ratios were saved, summed, and averaged. In the corresponding Ratio Matrix $^{[B]}$, the model captures the expected ratios extremely well. Using $\epsilon \leq 0.1$ as an approximate standard, the seasonal Markov Chain ratio remains relatively close to the expected ratios with the minor exception of the Rainy state in February. Nonetheless, it was able to capture that February did have a higher ratio of Rainy days than other months in the first quarter. When compared with the expected, or actual, ratios, the model obtains an MSE of 0.01534 or 2%.

2. Winter, Spring, Summer, Fall

4 Discussion

Across the 3 Markov Chain adaptations, each providing a unique approach to the weather prediction problem, the performance of each model results in rather adequate results. While it boils down to a trade off in complexity and accuracy, each model has its weaknesses and strengths. For the traditional model, we saw its framework was simple and interpretable; however, the transition probabilities were encompassing of the whole year, unable to express Earth's rotation around the sun. For the second-order difference equation, its framework was a bit more complex, but still at an interpretable level. Again, we have the same concern where Matrix A and B are probabilities of state transitions across the whole year. From a logical perspective, we expect Summer months to be hotter and mostly clear skies compared to Winter months, yet when the transition probabilites are based off the whole year, we consider it to be static weather patterns. Thus, to address this concern, the traditional model was adapted to include seasonal transition matrices for the corresponding months. From a theoretical point of view, the interpretability of this model is easy to follow; however, the complexity of it increases as it involved sectioned periods of time along with varying transition matrices that affect the expected state along with signal weights that either include or remove specific transition matrices. Nonetheless, using this model had its advantages reinforcing scientifically backed proposals that Earth, and in particular Los Angeles, experiences seasonality in the weather. With a significantly large decrease in Ratio Error, the Seasonal Markov Chain Model outperformed its predecessors with an MSE of 2%. It is also worth noting that each model resulted similar steady state distributions where Sunny, Chilly, and Cloudy consisted of values 0.30 ∓ 0.06 and Rainy and Partly Cloudy ranged 0.05 ∓ 0.02 indicated long term behavior of expecting more Chilly, Cloudy, and Sunny days in Los Angeles.

5 Forecasting

Using the Seasonal Markov Chain, we aim to forecast and interpret the expected weather system for the remaining months in 2024[3]. The table below indicates the ratio[4] of weather states for the following months in 2024. We can observe that more than half of each Summer and Fall month will consist of Sunny days and we see the rise in more Chilly days leading up to December.

	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
May	0.50	0.29	0.00	0.04	0.17
June	0.03	0.19	0.16	0.00	0.61
July	0.03	0.20	0.16	0.00	0.61
August	0.03	0.19	0.18	0.00	0.59
September	0.14	0.18	0.08	0.01	0.59
October	0.16	0.21	0.07	0.02	0.54
November	0.18	0.16	0.08	0.01	0.57
December	0.51	0.28	0.02	0.10	0.09

6 Further Works

With the goal of explaining the randomness in weather trends, further exploring the effects of seasonality in Markov Chains would be wise; however, it would be interesting to fuse this idea with the second-order difference equation to identify the effect of current days in each season. Additionally, it would be best to identify and address anomalies in the model. I pose this opportunity to any who wish to further this exploration.

3. May - December
4. Note: Entries are rounded to 2 decimal points

Appendix A. Transition Matrices

One-Step Transition Matrix (Traditional MC/Matrix A)

(To, From)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
Chilly	0.69	0.24	0.02	0.43	0.07
Cloudy	0.16	0.50	0.35	0.26	0.10
Partly Cloudy	0.00	0.08	0.19	0.09	0.09
Rainy	0.03	0.06	0.00	0.22	0.01
Sunny	0.12	0.12	0.44	0.00	0.74

Matrix B

(To, From)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
Chilly	0.57	0.34	0.08	0.51	0.12
Cloudy	0.25	0.36	0.30	0.24	0.14
Partly Cloudy	0.01	0.05	0.16	0.03	0.10
Rainy	0.03	0.09	0.00	0.19	0.01
Sunny	0.14	0.16	0.46	0.03	0.63

Winter Transition Matrix

(To, From)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
Chilly	0.72	0.32	0.00	0.39	0.35
Cloudy	0.18	0.48	0.40	0.22	0.24
Partly Cloudy	0.02	0.02	0.40	0.00	0.00
Rainy	0.03	0.15	0.00	0.39	0.00
Sunny	0.05	0.03	0.20	0.00	0.41

Spring Transition Matrix

(To, From)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
Chilly	0.69	0.33	0.00	0.64	0.30
Cloudy	0.19	0.56	0.00	0.36	0.12
Partly Cloudy	0.00	0.00	0.00	0.00	0.00
Rainy	0.04	0.07	0.00	0.00	0.03
Sunny	0.09	0.04	0.00	0.00	0.55

Summer Transition Matrix

(To, From)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
Chilly	0.33	0.09	0.00	0.00	0.01
Cloudy	0.17	0.37	0.31	0.00	0.10
Partly Cloudy	0.00	0.29	0.19	1.00	0.14
Rainy	0.00	0.03	0.00	0.00	0.00
Sunny	0.50	0.23	0.50	0.00	0.75

Fall Transition Matrix

(To, From)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
Chilly	0.59	0.15	0.08	0.50	0.05
Cloudy	0.17	0.48	0.38	0.00	0.08
Partly Cloudy	0.00	0.06	0.15	0.50	0.08
Rainy	0.00	0.03	0.00	0.00	0.01
Sunny	0.24	0.27	0.38	0.00	0.79

Appendix B. Ratio Matrices

Expected

(Month, State)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
January	0.58	0.26	0.06	0.06	0.03
February	0.41	0.34	0.00	0.24	0.00
March	0.55	0.29	0.00	0.13	0.03
April	0.60	0.33	0.00	0.03	0.03

Traditional Markov Chain

(Month, State)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
January	0.31	0.22	0.07	0.03	0.36
February	0.29	0.24	0.07	0.03	0.37
March	0.31	0.21	0.05	0.03	0.39
April	0.30	0.26	0.05	0.03	0.36

Second Order Difference Equation

(Month, State)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
January	0.38	0.23	0.06	0.04	0.28
February	0.34	0.24	0.07	0.03	0.32
March	0.34	0.25	0.06	0.04	0.32
April	0.32	0.24	0.06	0.05	0.33

Seasonal Markov Chain

(Month, State)	Chilly	Cloudy	Partly Cloudy	Rainy	Sunny
January	0.56	0.27	0.01	0.10	0.06
February	0.53	0.26	0.02	0.12	0.08
March	0.56	0.28	0.00	0.04	0.12
April	0.52	0.33	0.00	0.04	0.11

References

[1] California Measurement Advisory Council - California Weather Files. (n.d.). https://www.calmac.org/weather.asp

[2] Yutong, Xia. (2021). Applications of Markov Chain in Forecast. Journal of Physics: Conference Series. 1848. 012061. 10.1088/1742-6596/1848/1/012061.