Predicting precipitation in Darwin: An experiment with Markov chains

John Boncek and Sig Harden
Troy University, Montgomery Campus, Alabama
<jboncek@troy.edu>
<sharden39277@troy.edu>

As teachers of first-year college mathematics and science students, we are constantly on the lookout for simple classroom exercises that improve our students' analytical and computational skills. One such project, *Predicting Precipitation in Darwin*, is outlined below. In this project, students:

- analyze and manipulate raw precipitation data;
- build a prediction model using a Markov chain;
- predict the long term distribution of precipitation-free and rainy days in Darwin, Northern Territory, Australia;
- use a chi-square test to evaluate the effectiveness of the model they have constructed;
- improve their prediction model.

Beyond access to the Internet (to obtain the raw data) and a computer spreadsheet program or calculator, no special equipment is required. If the data is downloaded in advance, a well-prepared junior or senior high-school mathematics (or science) class should be able to perform this exercise in approximately 30–45 minutes of class time.

Mathematical preliminaries: Markov chains

A Markov chain is a sequence of identical trials, each of which can result in exactly one of a finite number of outcomes, called states. As the trials progress, the probability of moving from one state to another depends only on the state in which you are currently found.

In most applications, Markov chains are represented by a state transition matrix, P. In this matrix, the entry in the (i,j)th position (row i, column j) is the probability that you will move to state j in the next trial if you are currently in state i. Properly constructed, the sum of each row in the matrix is one.

Due to the properties of matrix multiplication, the (i,j)th entry in the matrix P^2 is the probability that you will move from state i to state j over the course of two trials; the (i,j)th entry in P^3 is the probability you will move from state i to state j in three trials, and so on.

If there is a positive integer n such that all the entries in the state transition matrix P^n are positive, the Markov chain is said to be regular. Regular Markov chains have the property that the limit $\lim\limits_{n\to\infty} P^n$ exists. Each row in the limiting matrix represents the stable state vector of the system. The stable state vector is used to determine the eventual distribution of the data among each of the possible states.

A simple example

Let us consider the Markov chain whose state transition matrix is given by

$$P = \begin{bmatrix} 0.25 & 0.75 \\ 0.5 & 0.5 \end{bmatrix}$$

and let us assume that each trial in this Markov chain results in our being in either state i or state j. Since all the entries in this matrix are positive, the Markov chain it describes is regular. There are two ways for us to determine $\lim\limits_{n\to\infty} P^n$.

The first approach is using a graphing calculator or computer to calculate successively higher powers of the matrix P^n. At some point, the values returned will no longer change. For example, using a TI-85 graphing calculator and rounding to four decimal places, we obtain:

$$P^2 = \begin{bmatrix} 0.4375 & 0.5652 \\ 0.375 & 0.625 \end{bmatrix}$$

$$P^5 = \begin{bmatrix} 0.3994 & 0.6006 \\ 0.4004 & 0.5996 \end{bmatrix}$$

$$P^{10} = \begin{bmatrix} 0.4000 & 0.6000 \\ 0.4000 & 0.6000 \end{bmatrix}$$

Higher powers of P return the same values as those obtained when calculating P^{10}. Thus,

$$\lim\nolimits_{n\to\infty} P^n = \begin{bmatrix} 0.4 & 0.6 \\ 0.4 & 0.6 \end{bmatrix}$$

A second approach is to use a theorem from Markov analysis that tells us that the entries in the limiting matrix can be determined by solving the following system of equations:

$$\begin{aligned} 0.25i + 0.35j &= i \\ 0.75i + 0.5j &= j \\ i + j &= 1 \end{aligned}$$

The solution of this system is $i = 0.4$, $j = 0.6$. Note that these are the same values you obtain using the graphing calculator technique outlined above.

These values give us important information about the long-term behaviour of our series of trials. If we allow the series to continue for a long enough period of time, we expect 40% of the trials to end up residing in state i and 60% of the trials ending up in state j.

Our experiment

We will study precipitation data from the Darwin airport for the years 1999–2008. We will assume that this information forms a Markov chain: each day either has measurable precipitation or it does not, the probability of it raining on a given day depends only if it rained on the previous day, and the probability of moving between rainy and rain-free days remains constant over the period under consideration.

We begin by obtaining daily precipitation data from the Darwin airport for the years 1999–2008 from the Australian Weather News archive website: www.australianweathernews.com/data/archive/14GA. A plot of these data is shown in Figure 1.

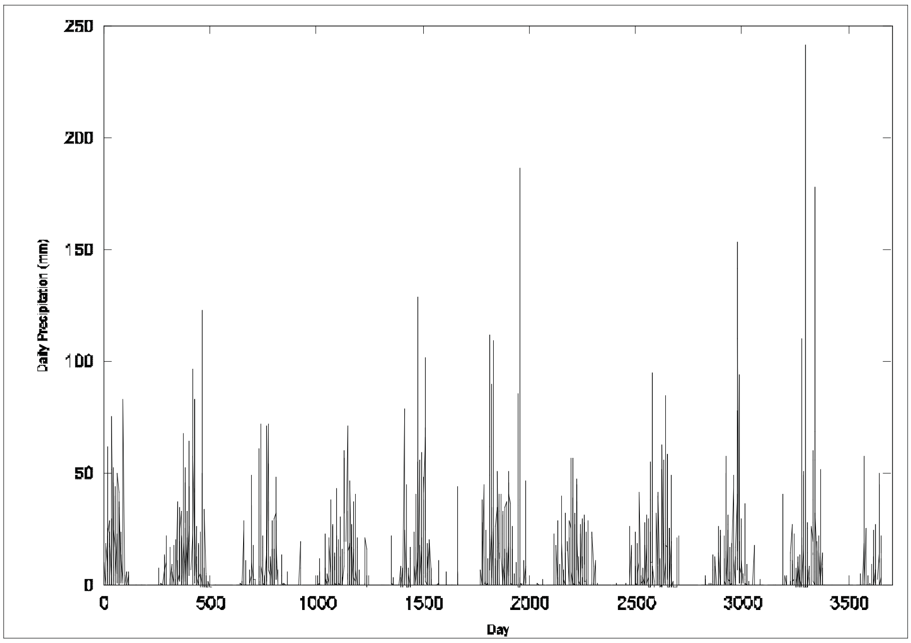

Figure 1. Darwin airport, daily precipitation, 1999–2008.

Exercise 1

We will determine if the daily precipitation data from 2008 can be used to model the data for the entire period 1999–2008. For purposes of this experiment, we assume that the data is representative for all ten years under consideration. We begin by analysing the 2008 data, which reveals the following:

Table 1

Number of precipitation-free days	256
Number of days with precipitation	110

Comparing the precipitation level of each day in the year to its successor, we determine the following about the paired data:

Table 2

	Second day precipitation-free	Second day had precipitation
First day precipitation-free	224 pairs	32 pairs
First day had precipitation	32 pairs	77 pairs

From this we compute the state transition matrix.

Table 3

	Second day precipitation-free	Second day had precipitation
First day precipitation-free	0.875	0.125
First day had precipitation	0.2936	0.7064

Thus, in 2008, if a given day was rain-free, there was an 87.5% chance the next day would be rain-free as well. On the other hand, if the day was rainy, there was a 70.64% chance the next day was rainy as well.

Long term analysis

Given the state-transition matrix above, the long-term behaviour of the Markov chain it describes can be determined by solving the system of equations:

$$\begin{aligned} 0.875i + 0.2936j &= i \\ 0.125i + 0.7064j &= j \\ i + j &= 1 \end{aligned}$$

or by calculating higher and higher powers of the state transition matrix using a graphing calculator or computer program. In either case, we obtain $i = 0.7014$ and $j = 0.7014$. Interpreting this result, we predict that over a long period of time, 70.14% of all days at the Darwin airport will be precipitation-free and 29.86% of them will see rain.

There were 3653 days during the period 1999–2008. Our model predicts that 70.14% of them, or 2562 of them, should be precipitation-free and 1091 should have measurable precipitation. Here is how the forecast compares to the actual data.

Table 4

	Forecast	Actual
Number of precipitation-free days	2562	2416
Number of days with precipitation	1091	1237

Our model over-predicts the number of precipitation-free days and under-predicts the number of rainy days. The question is: Is our model statistically accurate?

Chi-square goodness of fit test

A standard statistical method used to see if observed data fit a mathematical prediction is the chi-square goodness of fit test. We adopt the null hypothesis, H_0 = the observed data does not differ from the expected value versus the hypothesis H_1 = the observed data differs from the expected values. Using the chi-square goodness of fit test with 1 degree of freedom, we find

$$\chi^2 = \frac{(2416-2562)^2}{2562} + \frac{(1237-1091)^2}{1091} = 27.8581$$

The critical value for chi-square at the 95% confidence level is 3.8412, so we reject the null hypothesis. By conventional statistical criteria, the difference between our forecast and the observed data is considered to be significant.

What went wrong?

One of the assumptions in Markov analysis is that the probabilities in the state transition matrix are constant for all time. Table 5 indicates that the precipitation levels observed in Darwin in 2008 are not representative of the period 1999–2008.

Table 5. Number of precipitation-free days per year.

Year	Precipitation-free days
1999	224
2000	220
2001	238
2002	258
2003	252
2004	244
2005	250
2006	244
2007	230
2008	256

The mean number of precipitation-free days per year over the ten-year period is 241.6, with a standard deviation of 13.29. This indicates that 2008 was a particularly dry year. It is not surprising, therefore, that our forecast over-predicts the number of dry days during the period 1999–2008.

Exercise 2

The data in the previous table indicates that the average number of precipitation-free days per year during the two year period 2007–2008 was 243. This seems to be close to the average number of precipitation-free days per year for the ten-year period. Use the data for the two-year period 2007–2008 and repeat the forecast process completed in Exercise 1 and see if the results improve.

Table 6. Paired data.

	Second day precipitation-free	Second day had precipitation
First day precipitation-free	413 pairs	73 pairs
First day had precipitation	73 pairs	171 pairs

Table 7. State-transition matrix.

	Second day precipitation-free	Second day had precipitation
First day precipitation-free	0.8498	0.1502
First day had precipitation	0.2992	0.7008

Long term analysis

$$\begin{aligned} 0.8498i + 0.2992j &= i \\ 0.1502i + 0.7008j &= j \\ i + j &= 1 \end{aligned}$$

yields $i = 0.6658$ and $j = 0.3342$; that is, we expect 66.58% of all days to be precipitation-free and 33.42% of all days to have some amount of measurable precipitation. This gives us a ten-year forecast of 2432 precipitation-free days and 1221 rainy days.

Here is how the forecast compares to the actual data.

Table 8

	Forecast	Actual
Number of precipitation-free days	2432	2416
Number of days with precipitation	1221	1237

While this forecast still over-predicts the number of dry days by 16, it is obvious that our forecast is much closer to the actual data than the forecast we obtained in Exercise 1. Moreover, this forecast passes the chi-square goodness of fit test:

$$\chi^2 = \frac{(2416-2432)^2}{2432} + \frac{(1237-1221)^2}{1221} = 0.3419$$

This value is much less than the critical value of 3.4812. At a 95% level of confidence, we cannot say that our forecast results are statistically different from the observed data. In other words, this forecast model works, while the forecast model obtained in Exercise 1 did not.

Observations and comments

The two exercises demonstrate the importance of knowing your data before constructing a mathematical model. If we had known that 2008 was an unusually dry year before we used it to construct our Markov analysis, we would have known (or at least suspected) that its state transition matrix would not be representative for the ten-year period, and we would not have wasted our time using it for that purpose. This could have been detected easily if we had constructed a histogram or table of the yearly data, or inspected Figure 1 more closely.

Additional exercises can be constructed to determine if data from any of the other nine years can be used to construct an even better (and statistically acceptable) forecast than the one we obtained in Exercise 2.

Further reading

Lial, M. A. Greenwell, R. N. & Ritchey, N. P. (2008). *Finite mathematics* (9th ed.). Boston, MA: Pearson Addison Wesley.

Triola, M. F. (1998). *Elementary statistics* (7th ed.). Reading MA: Addison Wesley Longman.