Note: probabilities and simulation

Peter Rowlett

The way the weather data were produced is that I used a matrix of transition probabilities to simulate data. Starting with a transition matrix $\mathbf{P}$, here is the process I used to simulate the data:

- Calculate the long-term probabilities, $\mathbf{s}_n$.
- Generate a random number from 0-1 and use the long-term probabilities to turn this into a type of weather ('Sunny' or 'Rainy' for one type of data, 'Strong wind', 'Light wind' or 'None' for the other) for the first time-step.
- Then multiply the current simulated state by $\mathbf{P}$ to get the next state, and repeat until 52 rows of data are generated[1].

In the example given in the notes, we used the data from Town 1 on sun and rain and obtained the following transition probability matrix:

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0.481 \\ 0.5 & 0.519 \end{bmatrix}$$

In the exercises, you used the data from Town 2 and obtained the following transition matrix:

$$\mathbf{P} = \begin{bmatrix} 0.696 & 0.25 \\ 0.304 & 0.75 \end{bmatrix}$$

The conclusion given in the answer is:

> "*It seems likely that the probabilities used to generate the data for Town 2 are not the same as the probabilities used to generate the data for Town 1, because the transition probabilities based on the data differ substantially.*"

This is definitely the right conclusion, however it may interest you to learn that the data were in fact generated using the same transition probabilities! I generated the data for all towns using

$$\mathbf{P} = \begin{bmatrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{bmatrix}.$$

It seems that Town 1 is a bit of an outlier, in that by chance its numbers of sun and rain transitions came out pretty even. This is partly an artefact of the fact we are working with probabilities and that there is relatively little data here (51 transitions per Town).

We would expect with large numbers of transitions that the transition probabilities would be closer to those used to generate the data.

If you want to test with more data and were sure the underlying probabilities are the same, you could combine all the data sets into one big data set and count transitions. This has the limitation that the transition from week 52 of Town 1 to week 1 of Town 2 (etc.) is not a genuine transition (week 1 of Town 2 is randomised based on the long-run probabilities, not on a simulated transition from the previous week), so we are adding some bad data into our calculations, but this may be worth it if we can get enough of an advantage by combining the data.

For example, if you combine the data on sun and rain from all five towns into one set of $52 \times 5 = 260$ days, you can obtain the following transition counts.

[1] It may amuse you to learn that this number is 52 because I had it as weeks and only quite late in the process of writing these notes did I realise that the weather doesn't stay the same all week and change the column header to days.

		Now	
		Sunny	Rainy
Next	Sunny	76	50
	Rainy	51	82

Resulting in the transition matrix:

$$\mathbf{P} = \begin{bmatrix} 0.598 & 0.379 \\ 0.402 & 0.621 \end{bmatrix}.$$

This is still not the same as the probabilities used to generate the data, though it is closer.