To Ski or Not to Ski:
Estimating Transition Matrices to Predict Tomorrow's Snowfall Using Real Data

Michael A. Rotondi
The University of Western Ontario, London, Canada

Journal of Statistics Education Volume 18, Number 3 (2010)
www.amstat.org/publications/jse/v18n3/rotondi.pdf

Key Words: Precipitation, Snowfall, Transition Matrix, Weather Models, Forecasting.

Abstract

Using historical data from the Global Historical Climatology Network (GHCN)-Daily database, the use of Markov chain models is presented to predict a 'Snow Day' at eight national weather stations. This serves as a variation of the classic Markov chain precipitation example, predicting a significant snow depth tomorrow from today's snow depth conditions. Stations near Seattle WA, Denver CO, Milwaukee WI, Chicago IL, New York NY and Boston MA, were included as they represent major urban centers, while stations in Montana and North Dakota were added to improve geographical coverage. Estimates of the appropriate transition matrices ($\hat{P}_i$) are provided, as well as a sample of code in the R statistical programming language to enable construction of similar examples for other geographical areas.

1. Introduction

The application of Markov chains to model precipitation data was introduced over 40 years ago (Chin 1977; Gabriel and Neumann 1962; Katz 1977). Despite its overall simplicity, the first order Markov model (based upon the previous day's rainfall) remains a suitable technique for the modeling of precipitation data in many geographical areas (Schoof and Pryor 2008). As such, variations of the 'Markov weather model' are found in a variety of textbooks (e.g., Ross 2003), as well as the potentially influential Wikipedia page for 'Examples of Markov Chains' (Wikipedia 2010).

In light of their instructive popularity, the purpose of this data set and sample code is to provide a seasonal variation of the Markov chain weather model, where a 'Snow Day' tomorrow is predicted from the current snow depth conditions today. Within this framework, estimates of the Markov transition matrix P are presented for eight American weather stations, as estimated from the Global Historical Climatology Network (GHCN)-Daily reports.

To ensure consistency in the outcome of interest, a 'Snow Day' is defined as a day where at least 50 mm (2 inches) of accumulation (snow depth) is observed. This definition is plausible, as all weather stations are located in areas of significant snowfall. In this manner, a 'Snow Day' implies a generous amount of snow on the ground (accumulation), rather than simply observing snowfall. In addition, the described example emphasizes forecasting snowfall over the last two weeks of December, coinciding with end of term examinations and Christmas break. Note that the included example can easily be modified for other weather outcomes, (e.g., maximum daily temperature, or precipitation), as well as alternative months of interest using the accompanying code.

A general introduction to the GHCN database and a summary of the included weather stations is presented below, followed by a brief introduction to Markov chains. The paper concludes with general results and potential classroom applications. Appendices include a worksheet of sample questions (Appendix A), and a detailed example illustrating the described calculations and computer code (Appendix B). Note that readers are assumed to have some familiarity with R (R Development Core Team 2010). However, students do not need prior experience with any programming language to benefit from these examples.

2. Data Description

The included data were obtained from the GHCN-Daily database. General information regarding this resource is paraphrased from the GHCN home page: http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/index.php, while details related to the included weather stations

are presented below. Note that the data files for the eight included weather stations are available from the *JSE* web site. For consistency with other journal articles, data files were saved using the .dat extension.

You may access these data files by clicking on the name of the file in the following table. Note that the data files for the eight included weather stations are quite large. Also note that if you download the files directly from the GHCN home page they will have a .dly extension while if you download the data files from the *JSE* site they will have a .dat extension. The R code to read in the Central Park data (see Appendix B) is written for downloading files from the *JSE* site. The only change needed in the code if you download files from the GHCN home page is to change the .dat extension to .dly.

Data File	Data File
BlueHill.dat	CentralPark.dat
Charleston.dat	DelNorte.dat
Glendive.dat	Medford.dat
SedroWooley.dat	Willow.dat

2.1 General Information

The GHCN-Daily database contains extensive (daily) temperature, precipitation, and snowfall records from around the world. These data records are obtained from numerous sources and are subjected to extensive quality control measures.

The archive includes the following meteorological elements, as well as their corresponding weather abbreviations and units:

- Daily maximum temperature (TMAX: tenths of degrees Celsius)
- Daily minimum temperature (TMIN: tenths of degrees Celsius)
- Precipitation (i.e., rainfall and snow water equivalent) (PRCP: tenths of mm)
- Snowfall (SNOW: mm)
- Snow depth (SNWD: mm)

The data set contains observations of one or more of the above weather conditions at more than 40,000 stations across the world, representing the world's largest collection of daily climatological data. Specifically, a total of 1.4 billion (daily) data values including 250 million values each for maximum and minimum temperatures, 500 million precipitation

totals, and 200 million observations each for snowfall and snow depth are available. As a frame of reference, the *complete* data totals over 1.7 gigabytes of compressed information. Note that all data files are freely available from the GHCN data page, http://www1.ncdc.noaa.gov/pub/data/ghcn/daily.

In light of the breadth of this resource, an appropriate weather station and outcome of interest must be selected. This may be based on a number of factors including: geographic proximity, student interest, or project assignment. In this paper, focus is on the SNWD variable as the outcome of interest. Note that the following data acquisition steps are only required should an instructor wish to incorporate weather stations in addition to the included eight. A brief outline of the procedure is summarized below:

1. Select an appropriate weather station from the *ghcnd-stations.txt*. Locations are sorted by country, geographic latitude/longitude and altitude. Note that all station identifiers begin with a two letter country code, which is detailed in *ghcnd-countries.txt*. For specificity, US denotes the United States. In general, US locations are sorted by state and town, thus the browser's search feature should be used to search for the desired town name. In addition, knowledge of the approximate latitude and longitude of the desired location will confirm that the selected location is correct.

2. Ensure that the desired location is a member of the Historical Climatology Network (HCN) in order to apply the included code. That is, confirm that 'HCN' is listed in the row describing the selected station, as the included functions are designed for use with HCN stations only. Upon selecting the desired city, highlight and copy the location code (beginning with US...in the first column) as this will facilitate searching for the required data file. Note that the location code (e.g. USC00305801) does not correspond to the station name (e.g., Central Park).

3. After selecting the appropriate weather station code, download the corresponding data file (.dly) from the *hcn* subdirectory located in the main web page. The direct URL to the *hcn* directory is: http://www1.ncdc.noaa.gov/pub/data/ghcn/daily/hcn/. Note that .dly files are only available for the HCN stations. The desired data set is listed with a corresponding location code as the file name. Given the volume of data, it is recommended to search for the location code using the browser. The data set may now be saved with an appropriate file name (e.g. 'CentralPark.dat') and read into the R environment.

2.2 Data Layout

The first column represents the appropriate data code and preliminary data. For example, the entry USC00305801187601TMAX in the first row of the Central Park data is comprised of the location code (USC00305801), record year (1876), month (01 = January) and element code, (TMAX). Recall that the possible element codes include: TMAX and TMIN (the maximum and minimum daily temperature), PRCP (precipitation amount), SNOW (observed snowfall) and the variable of interest, SNWD (snow depth). The available data could easily allow for the prediction of other weather conditions using the provided code.

The remaining 124 (31×4) columns of the data set correspond to the maximum of 31 days in each month, where each day occupies four columns, namely the daily record value and three quality control characters. Additional details related to the structure of each data file are available online in the GHCN description file, *readme.txt*. Data files (.dly) can be read into the R environment using the accompanying code in Appendix B.

2.3 Included Weather Stations

Each location was selected from the daily Global Historical Climate Network (GHCN 2010), in an effort to represent major urban centers and the central northern states. Included weather stations are shown on a map of the United States in Figure 1, while details pertinent to the 'Snow Day' analysis are summarized in Table 1:

Table 1. Station Characteristics

Station Identifier	Location, State	Latitude	Longitude	Altitude	Earliest SNWD Record	Sample Size
USC00305801	New York, NY	40.78°	-73.97°	40.0'	1912	1346
USC00457507	Sedro Wooley, WA	48.50°	-122.23°	18.0'	1898	1172
USC00243581	Glendive, MT	47.10°	-104.72°	633.0'	1933	1039
USC00329445	Willow City, ND	48.60°	-100.28°	445.0'	1935	555
USC00052184	Del Norte, CO	37.67°	-106.32°	2399.0'	1938	564
USC00475255	Medford, WI	45.13°	-90.35°	448.0'	1889	1067
USC00111436	Charleston, IL	39.47°	-88.18°	207.0'	1901	1373
USC00190736	Blue Hill, MA	42.20°	-71.10°	192.0'	1895	1358

Notes:
- Station Identifiers refer to the GHCN (2010) location codes.
- SNWD: Snow depth

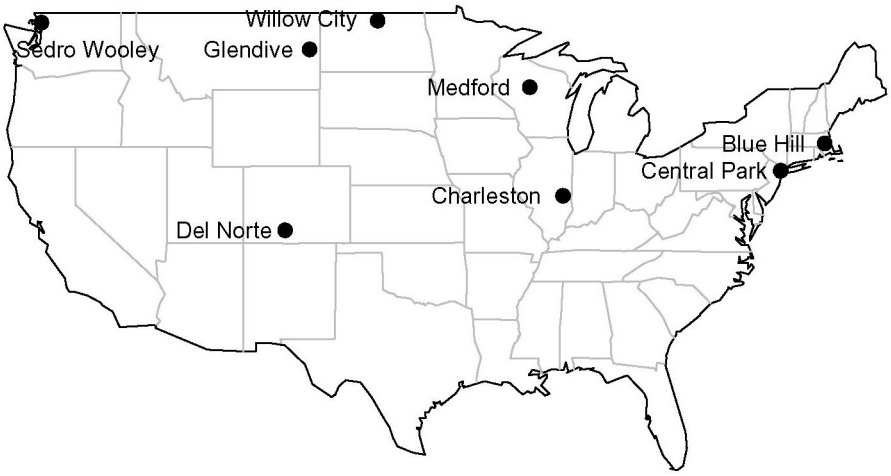

Figure 1. Locations of Included Weather Stations, USA

3. The Markov Weather Model

The Markov chain model incorporates the current snow depth conditions today in the prediction of a significant snow depth tomorrow. The described notation is adapted from Ross (2003), to which the reader may refer for additional information about Markov chains.

As a preliminary introduction, consider a stochastic process $X(t)$ $(t = 0,1,2,\ldots)$ that takes on a finite number of states k. In general, the process $X(t)$ is in state $s = 0,1,2,\ldots,k$ at time $t = n$ if $X(n) = s$. Furthermore, let s_t denote the observed state of the stochastic process at time t. For example, the initial state of the stochastic process is represented by: $X(t = 0) = s_0$. In this application, the process $X(t)$ has two states, namely $X(t) = 1$ if the observed snow depth (*SNWD*) at time t is at least 50 mm (a 'Snow Day') and $X(t) = 0$ if $SNWD < 50$ mm (a 'Green Day').

Furthermore, suppose that for all $t \geq 0$, the *Markov property* holds, that is,

$$\begin{aligned} P(\ X(t+1) = s_{t+1} \mid X(t) = s_t, X(t-1) = s_{t-1}, \ldots, X(1) = s_1, X(0) = s_0\ ) \\ = P(\ X(t+1) = s_{t+1} \mid X(t) = s_t\ ) \end{aligned}$$

In words, this property states that the probability of being in state s_{t+1} at time $t+1$ depends only on the current state s_t. As related to this example, observation of a significant snow depth tomorrow depends only on whether (or not) a significant snow depth is observed today and not on earlier snow depth conditions. This model thus incorporates the dependence of tomorrow's snow depth on today's observed snow depth to inform the probability of observing a significant snow depth tomorrow.

Under these assumptions, a simple first-order Markov model can be constructed for each weather station. Note that the station subscript i corresponds to the eight included weather stations respectively. Let,

$$a_i = P(\ X(t) = 0 \mid X(t-1) = 0\ ) = P(\ SNWD < 50 \text{ mm Tomorrow} \mid SNWD < 50 \text{ mm Today}\ )$$
$$b_i = P(\ X(t) = 0 \mid X(t-1) = 1\ ) = P(\ SNWD < 50 \text{ mm Tomorrow} \mid SNWD \geq 50 \text{ mm Today}\ )$$
$$c_i = P(\ X(t) = 1 \mid X(t-1) = 0\ ) = P(\ SNWD \geq 50 \text{ mm Tomorrow} \mid SNWD < 50 \text{ mm Today}\ )$$
$$d_i = P(\ X(t) = 1 \mid X(t-1) = 1\ ) = P(\ SNWD \geq 50 \text{ mm Tomorrow} \mid SNWD \geq 50 \text{ mm Today}\ )$$

These values can be arranged in the 2×2 transition matrix, $\mathbf{P_i}$:

$\mathbf{P_i} =$

	SNWD < 50 mm Tomorrow	SNWD ≥ 50 mm Tomorrow
SNWD < 50 mm Today	a_i	b_i
SNWD ≥ 50 mm Today	c_i	d_i

In this matrix representation, the first row (and column) denotes the absence of a significant snow depth, while the second row (and column) denotes the presence of a significant snow depth. For example, the parameter a_i represents the probability of observing an *in*significant snow depth tomorrow (i.e., a 'Green Day' tomorrow), provided an *in*significant snow depth was observed today (i.e., a 'Green Day' today). Also note that as these values are probabilities, $a_i + b_i = 1$ and $c_i + d_i = 1$. That is, a(n) (in)significant snow depth today must be followed by either the occurrence or absence of a significant snow depth tomorrow. Finally, due to these dependencies, the entire model is determined by estimation of a single probability in each row of P_i.

Estimates of these probabilities can be obtained from maximum likelihood estimation. In this case, each $\hat{a}_i$ is estimated from the number of two-day sequences where $X(t) = 0$ and $X(t-1) = 0$, and $\hat{d}_i$ from the number of consecutive day-pairs where $X(t) = 1$ and $X(t-1) = 1$. For example, $\hat{a}_i$ is estimated by counting the number of days where $SNWD <$ 50 mm on two consecutive days divided by the total number of two-day sequences where $SNWD < 50$ mm on the first day. These sequences are obtained from the historical weather patterns between December 17th and December 31st, in order to minimize bias due to the seasonality of weather patterns. An estimate of $\hat{d}_i$ is obtained in a similar manner. This process is then repeated for each of the eight weather stations, producing the eight sample matrices $\hat{\mathbf{P}}_i$. Additional details related to the estimation of Markov chain parameters are available in Anderson and Goodman (1957) and Billingsley (1961).

4. Results and Classroom Uses

Estimates of the eight respective transition matrices $\hat{P}_i$ are presented in Table 2. Note that in the majority of cases, observance of a significant snow depth on the current day suggests that a 'Snow Day' tomorrow will likely occur, as the probability of this event ranges from 77 % to 99 %. However, the presence of a significant snow depth (accumulation) on the current day in Central Park (New York) has an approximately 1 in 5 chance of melting before the next day, while in Medford WI, a 'Snow Day', today almost certainly ensures a 'Snow Day' tomorrow.

Table 2. Station Characteristics and Results

Station Identifier	Location, State	Transition Matrix $\hat{P}$
USC00305801	New York, NY	$\begin{pmatrix} 0.964 & 0.036 \\ 0.224 & 0.776 \end{pmatrix}$
USC00457507	Sedro Wooley, WA	$\begin{pmatrix} 0.986 & 0.014 \\ 0.192 & 0.808 \end{pmatrix}$
USC00243581	Glendive, MT	$\begin{pmatrix} 0.949 & 0.051 \\ 0.033 & 0.967 \end{pmatrix}$
USC00329445	Willow City, ND	$\begin{pmatrix} 0.933 & 0.067 \\ 0.012 & 0.988 \end{pmatrix}$
USC00052184	Del Norte, CO	$\begin{pmatrix} 0.974 & 0.026 \\ 0.040 & 0.960 \end{pmatrix}$
USC00475255	Medford, WI	$\begin{pmatrix} 0.861 & 0.139 \\ 0.011 & 0.989 \end{pmatrix}$
USC00111436	Charleston, IL	$\begin{pmatrix} 0.966 & 0.034 \\ 0.045 & 0.955 \end{pmatrix}$
USC00190736	Blue Hill, MA	$\begin{pmatrix} 0.929 & 0.071 \\ 0.102 & 0.898 \end{pmatrix}$

Notes:
• Station Identifiers refer to the GHCN (2010) location codes.

The described data may be included in a preliminary introduction or review of Markov chains in an elementary course in stochastic analysis or applied probability. A less developed form of this example was presented at a review session illustrating the use and properties of Markov chains for a second-year course in Applied Probability, for which the author was a teaching assistant. Even in its less developed state, the example was well-received as the presentation of actual data using local (Ottawa, Canada) weather patterns increased student attentiveness and comprehension of the material.

Application of these transition matrices may involve standard Markov chain analysis ques-

tions. For example, upon presentation of the transition matrix (P), the student could be asked to determine characteristics of the Markov process, such as the limiting probabilities (Ross 2003, Theorem 4.1). These correspond to the long-term proportion of time where one would expect a 'Snow Day' or 'Green Day' during winter break (December 17 and December 30) respectively (Ross 2003, Examples 4.1 and 4.17).

Alternatively, holiday-related questions, such as estimating the probability of observing a 'White Christmas' given a significant snow depth is observed Christmas Eve; or, calculating the probability of observing adequate ski conditions on December 28th, given that significant snow depth is observed on Christmas Day can be introduced. A sample worksheet of potential questions is presented in Appendix A.

5. Conclusions

The Daily Global Historical Climate Network (GHCN 2010) is an extensive repository of freely-available weather data. In light of this resource, it would be straightforward to create other alternative weather examples for most local environments using the accompanying source code. Available variables include temperature, precipitation, snowfall and snow depth for over 40,000 weather stations worldwide, suggesting a range of potential weather applications for virtually any country or geographical region.

Although the primary goal of the paper was to present appropriate transition matrices for an entertaining Markov chain example, the available data allows for alternative analyses, such as Markov chain modeling of precipitation data. In addition, continuous outcomes including daily temperature records are available and may be modeled using time series techniques. This database may also supplement the monthly temperature data described by Jacobson et al. (2009).

Acknowledgments

The author would like to thank Nooshin Khobzi for her invaluable discussion regarding the presentation of the final draft as well as the Associate Editor and two anonymous referees for their helpful comments and suggestions. The author's work is supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about markov chains. *The Annals of Mathematical Statistics*, 28:89–110.

Billingsley, P. (1961). Statistical methods in markov chains. *The Annals of Mathematical Statistics*, 32:12–40.

Chin, E. H. (1977). Modeling daily precipitation occurrence process with markov chain. *Water Resources Research*, 13:949–956.

Gabriel, K. R. and Neumann, J. (1962). A markov chain model for daily rainfall occurrence at tel aviv. *Quarterly Journal of the Royal Meteorological Society*, 88:90–95.

GHCN (2010). Global historical climate network. Available at: http://www1.ncdc.noaa.gov/pub/data/ghcn/daily.

Jacobson, T., James, J., and Schwertman, N. C. (2009). An example of using linear regression of seasonal weather patterns to enhance undergraduate learning. *Journal of Statistics Education*, 17(2). Available online at http://www.amstat.org/publications/jse/v17n2/jacobson.pdf.

Katz, R. W. (1977). Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16:671–676.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ross, S. M. (2003). *Introduction to Probability Models*. Academic Press, San Diego, CA.

Schoof, J. T. and Pryor, S. C. (2008). On the proper order of markov chain model for daily precipitation occurrence in the contiguous united states. *Journal of Applied Meteorology and Climatology*, 47:2477–2486.

Wikipedia (2010). Examples of markov chains. Available at: http://en.wikipedia.org/wiki/Examples_of_Markov_chains.

A. Sample Worksheet

$\mathbf{P}_{\text{Central Park, NY}} =$

	'Green Day' Tomorrow	'Snow Day' Tomorrow
'Green Day' Today	0.964	0.036
'Snow Day' Today	0.224	0.776

$\mathbf{P}_{\text{Medford, WI}} =$

	'Green Day' Tomorrow	'Snow Day' Tomorrow
'Green Day' Today	0.861	0.139
'Snow Day' Today	0.011	0.989

1. If a significant snow depth ('Snow Day') is observed on Christmas Eve, what is the probability of observing a 'White Christmas' ('Snow Day') in Central Park, NY?
2. What is the long-term probability of observing a 'Snow Day' in Medford, WI?
3. Suppose there is a significant snow depth on Christmas Eve, what is the probability of having adequate ski conditions in Central Park on December 27th?

B. The Central Park Data

To illustrate the included functions, the data cleaning and estimation procedures for the Central Park (New York) weather station are presented in detail. Note that these calculations represent only one of the eight included stations, as represented by the subscript i. All analyses and data cleaning were performed in the R environment. The eight described data sets and accompanying R functions are also available for download from the *JSE* site.

You may access these data files by clicking on the name of the file in this table. Note that the data files for the eight included weather stations are quite large. Also note that if you download the files directly from the GHCN home page they will have a .dly extension while if you download the data files from the *JSE* site they will have a .dat extension. The R code to read in the Central Park data is written for downloading files from the *JSE* site. The only change needed in the code if you download files from the GHCN home page is to change the .dat extension to .dly.

The R script for the analysis of the Central Park data is available at http://www.amstat.org/publications/jse/v18n3/CentralParkRCommands.txt. The documentation file that describes the Central Park data is available at http://www.amstat.org/publications/jse/v18n3/CentralPark.txt.

B.1 Data Cleaning

1. Locate the file CentralPark.dat in the accompanying online supplement. Alternatively this data set is available from the GHCN web site, http://www1.ncdc.noaa.gov/pub/data/ghcn/daily/hcn/, as USC00305801.dly. Note that files downloaded from the *JSE* online supplement were saved using the .dat extension for consistency with recently published articles.

2. Using the *CentralParkRCommands.R* code, specify the correct file name (and download location) and import the data into R. Specification of the correct path or working directory is necessary for the included code to function as designed. Note that the raw data set is represented by a matrix of 10, 354 rows and 128 columns. (Note that it is convenient to put the data file in the same location as the R working directory. To find out your R working directory, use getwd( ).)

3. The provided function locationCleaning() is used to select the desired range of days to include in the analysis (firstDay=17 and lastDay=31), month (December=12) and weather condition (SNWD) of interest, while summarizing the data in an accessible format. Note that these examples can be generalized for use with any weather condition (TMAX, TMIN, PRCP, SNOW or SNWD), as well as month or day range of interest. The use of this function is presented below:

```
clean <- locationCleaning(data=centralPark,
weather="SNWD", month = 12, firstDay = 17, lastDay = 31)
```

In the interest of brevity, only the first (and last) 5 rows of the resulting matrix are presented on the following page. Note that the first column denotes the year of interest, while the subsequent columns represent the observed snow depth (in mm) on the corresponding day in December of that year. At this stage, the resulting matrix has 93 rows and 16 columns, as SNWD records are unavailable from 1999–2002.

Year	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1912	0	0	0	0	0	0	0	305	203	152	102	76	51	0	0
1913	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1914	0	0	0	0	0	0	0	152	102	51	25	0	0	0	0
1915	25	0	0	0	0	0	0	0	0	25	25	0	0	0	0
1916	1778	1270	1016	686	305	0	0	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2004	0	0	0	0	0	0	0	0	0	0	51	51	0	0	0
2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2008	0	0	0	76	76	76	76	0	0	0	0	0	0	0	0
2009	0	0	0	254	178	178	127	102	51	0	0	0	0	0	0

4. The final data cleaning step requires the relabeling of missing values and the specification of a significant snow depth at which to dichotomize the observed snow depth quantities. Recall that a significant snow depth (50 mm) on the jth day is represented as $X(t) = 1$, while $X(t) = 0$ indicates otherwise. This threshold value of SNWD is specified by SNOWCrit=50 (mm). The dichotomization at this level is accomplished through use of the makeBinary() function:

```
SNOWBinary <- makeBinary(SNOW=clean, SNOWCrit=50)
```

This creates the following (abbreviated) data set for Central Park:

Year	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1912	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
1913	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1914	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
1915	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1916	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2004	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
2005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2006	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2008	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
2009	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0

B.2 Estimation of the Transition Matrix

Upon completion of the data cleaning procedures, the data are represented by a 93×16 matrix composed of the value 1 if at least 50 mm of snow was observed on that day and 0 otherwise. All missing values were also changed from the original -9999 to 'NA'.

The estimation function, makeP() provides the maximum likelihood estimates of the elements of P_1:

```
Probs <- makeP(SNOW01=SNOWBinary)
```

which results in:

$$\hat{\mathbf{P}}_{\text{Central Park, NY}} = \begin{pmatrix} \hat{a_1} & \hat{b_1} \\ \hat{c_1} & \hat{d_1} \end{pmatrix} = \begin{pmatrix} 0.964 & 0.036 \\ 0.224 & 0.776 \end{pmatrix}$$

Recall that in this example, $\hat{b}_1$ denotes the estimated probability of observing a significant snow depth tomorrow given that no significant snow depth was observed today, and $\hat{d}_1$ denotes the estimated probability of observing a 'Snow Day' tomorrow given a significant snow depth was observed today. In addition, the makeP() function provides an overall estimate of the probability of observing a significant snow depth on any day ($p = 0.128$). This value may be used to demonstrate the increase in model power through the incorporation of today's snow depth in the prediction of tomorrow's snow depth outcome.

Michael A. Rotondi
mrotondi@uwo.ca
Department of Epidemiology and Biostatistics
The University of Western Ontario
Room K201
London, Ontario, N6A 5C1
CANADA