A probabilistic forecast system based on Markov chains

Eleonóra Bombiczová[1], Robin Kokot[2], Anamarija Kozina[3], Gracia Zrnić[4] | Ivan Miošić[5]

[1]Gymnázium Petra Pázmáňa s VJM, Nové Zámky
[2]Prva gimnazija Varaždin, Varaždin
[3]Gimnazija Metković, Metković
[4]Ženska opća gimnazija Družbe sestara milosrdnica s pravom javnosti, Zagreb
[5]Department of Mathematics, Faculty of Science, University of Zagreb

Within the frame of this project, the main aim was to use Markov chains and related mathematical methods in order to develop a model of the weather in West Jurong, Singapore. Historical data for the past nine years was analysed, and in doing so discovered that, for the purposes of the project, optimal accuracy and precision were achieved when measures of weather components (e.g. temperature) were predicted within predetermined intervals. Our secondary aim was to compare the resulting forecast to real-world data and conclude whether such approach can be applied reliably. Having constructed the model, we interpreted the results as correct, however with limited exactness.

Introduction

Weather forecasting is known to have an irreplaceable role for various purposes: agriculture, air traffic, anticipation of extreme conditions etc. Advanced meteorological techniques have been developed in order to take all relevant factors into account, but forecasts remain imperfect due to the chaotic nature of the atmosphere and our incomplete understanding of it, the massive computational power required to solve the equations that describe it and so on. The aim of this project was determining whether such complications can be avoided in a simplified model, and to which point in the future it would still be applicable. The means of accomplishing that consisted of statistical analysis of past data (with the goal of establishing patterns and the probabilities with which they occur) and implementing obtained numbers in a Markov-chain based model.
The initial assumption was that weather as a process has the Markov property - that the probability of each event depended only on the state attained in the previous one. Though undeniably incorrect, this was close enough for superficial analysis, and allowed for a design of a simplified models. Two versions of tests were performed: one dealing with predicting daily, and another predicting monthly temperatures.

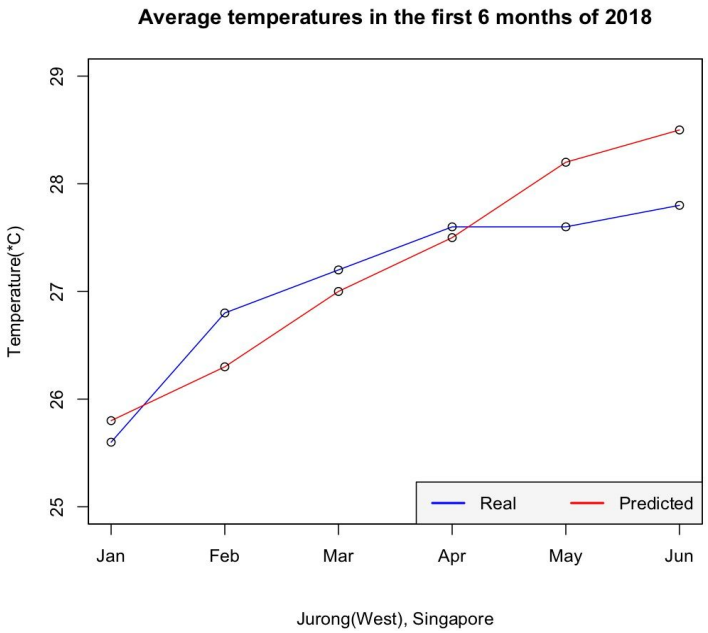
Results

When processed as described in Methods, statistical data yielded results laid out in Table 1.

Minimum temperature /°C		Maximum temperature /°C	
Our prediction	Actual data	Our prediction	Actual data
	27		30
	27		31
	28		31
<26.1, 28]	27	<29.75, 31.5]	31
<26.1, 28]	28	<29.75, 31.5]	31
<26.1, 28]	27	<29.75, 31.5]	31
<26.1, 28]	28	<29.75, 31.5]	31
<26.1, 28]	27	<29.75, 31.5]	31
<26.1, 28]	28	<29.75, 31.5]	31
<26.1, 28]	27	<29.75, 31.5]	31
<26.1, 28]		<29.75, 31.5]	
<26.1, 28]		<29.75, 31.5]	

Table 1. Predicted intervals and measured temperatures for maximum and minimum temperatures

As for predictions during a longer timespan, Graph 1 shows the predicted and actual (measured) average values for temperature for the first 6 months of 2018. The graph shows a strong correlation between predicted values and actual values from which was concluded that the level of precision and accuracy regarding our measurements was satisfactory at that scale. Comparison between the two is shown in Table 2 as well.



Graph 1. Plot of predicted and measured values over a 6-month timespan

	Predicted values	Real values
January	25.8	25.6
February	26.3	26.8

March	27.0	27.2
April	27.5	27.6
May	28.2	27.6
June	28.5	27.8

Table 2. Predicted and measured values over a 6-month timespan

Methods

Markov chains

The following is a description of modelling temperature changes via Markov chains. Having been given past data in Excel sheets, and having used it to calculate minimum and maximum values for both minimum and maximum temperatures for each day, we found mean values for all categories (i.e. we then had mean minimum minimum temperature, mean maximum minimum temperature, mean minimum maximum temperature, and mean maximum maximum temperature). The first pair was used as downer and upper limit for the interval within which future minimum temperatures was to be predicted, and the second for the same purpose regarding maximum temperatures. During past years, those intervals appeared relatively unchanging, and year 2017, which was used for testing, also indicated that they were appropriate. Nonetheless, the intention was to adjust them intuitively if the starting data differed noticeably from corresponding periods in previous years, especially if they were out of range. Such slight adjustments were made in one place.

Thus, we had two intervals, and they were divided into three equal subintervals, representing low, medium, and high minimum/maximum temperature (a separate Markov chain was made for those two). They were the three states of the process. The same division into low, medium and high could be done for each month that we inspected. By counting how many times out of total each subinterval lasted or was replaced by another, calculation of the transition probabilities could be done. Upon being presented with the starting data, finding the initial distributions was a matter of ascertaining which subintervals the data fit in. Consequently, two 3x3 matrices complete with transition probabilities were obtained, as well as their corresponding initial distributions. Using an online tool *WolframAlpha*, they were multiplied as many times as there were days. Conversion of resulting probabilities into predictions is elaborated in *Discussion*.

R

A useful tool in our data processing was R, a software environment for statistical computing and graphics. One purpose of it was to graphically simplify the behaviour of weather in order to make reappearing patterns more obvious; it acted as a guide in discerning which occurrences were indeed connected, as opposed to their short-term concurring being a coincidence. Using functions such as *plot(), cor.test()*, and others, a better understanding of how various components of weather were interdependent was gained.

Apart from predicting future states from present ones, it can be convenient in weather analysis to be able to make conclusions about a present value of one weather component, from the present value of another. An addition to the Markov chain model was therefore a series of graphs and correlation coefficients obtained through R - this was its other purpose.

Among notable observations was the expected correlation between maximum and minimum (and consequently, mean) temperature, and the more surprising correlation between mean temperature and maximum wind speed.
Contrary to previous expectations, daily rainfall didn't correlate meaningfully with temperature, or any other given values.
There were no notable tendencies to negative correlation.

Temperature forecast over a multiple-month period

In order to calculate the probability of temperature reaching a certain value in a given month, the behaviour of the temperatures during the given time interval (2009 - 2016) had to be analysed. This was done by taking the maximum and minimum values for temperature for each month of every year, and since the goal was to predict the temperature values for the first 6 months of 2018, these temperatures were divided according to months and the average temperature for each month during the aforementioned time interval was calculated (Table 3).

	Jan	Feb	Mar	Apr	May	June	July	August	September	October	November	December
2009	N/A	N/A	27.1	27.5	28.2	28.5	27.7	27.7	27.8	27.5	26.3	26.2
2010	N/A	N/A	N/A	N/A	28.4	27.6	27.1	27.2	27.3	27.8	26.8	26.1
2011	25.2	26.3	26.4	27.2	27.9	27.9	28.3	27.5	27.5	26.9	26.4	25.8
2012	26.3	26.6	26.6	27.4	27.9	28.7	27.6	27.6	27.7	27.3	26.6	25.9
2013	26.5	26	27.7	27.8	28	28.9	27.8	27.7	27.3	27.2	26.7	25.9
2014	25.7	26.8	27.2	27.7	28	29	N/A	N/A	N/A	N/A	N/A	N/A
2015	N/A	26.6	27.3	27.6	28.2	28.3	28.7	28	28.1	28.1	26.9	27
2016	31.6	28.1	28.2	28.7	28.5	27.9	27.85	28.6	27.9	27.7	26.9	27

Table 3. showing the average temperature for each month of every year in the given time interval

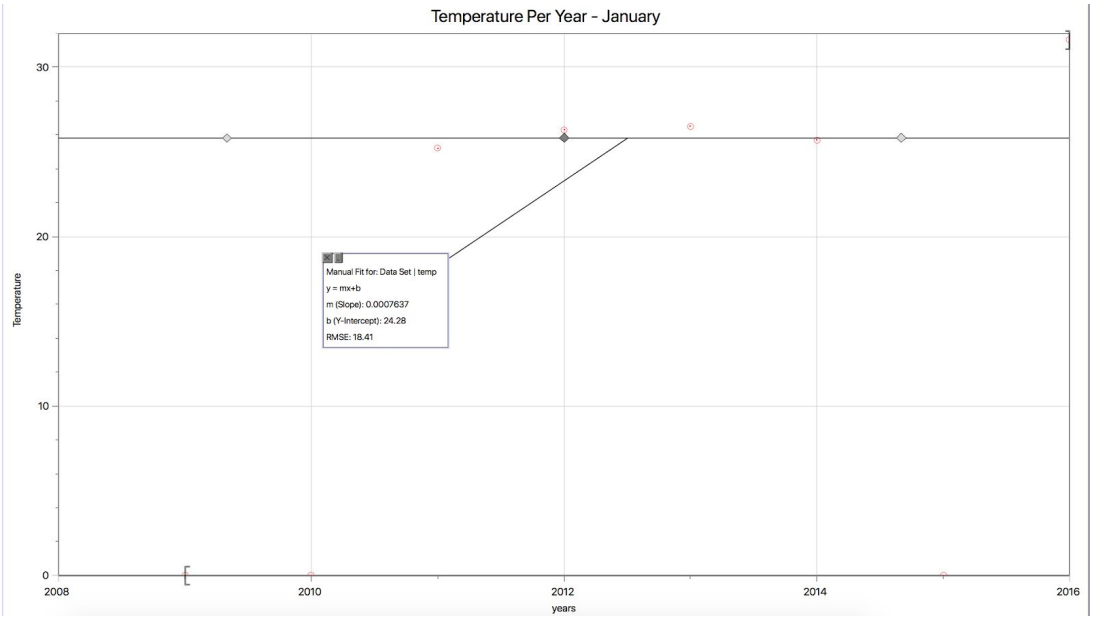
These values were then taken and obtained an average which was later used in categorizing the values into intervals. Intervals were used in order to see what part of the data shows the biggest grouping around certain temperature values (the goal was to see which temperatures occur most often and in which months).
Before the data was organised an interval size which would be the most beneficial for this type of measurement had to be established and the conclusion was that in order to assure relatively high precision and accuracy temperature values had to be organised into intervals of 3,5 ℃. If we took smaller intervals we would have more blank intervals and we couldn't use them in representation of our data, and if we took bigger intervals we would sacrifice the level of precision with which we can predict the temperature values.
The table attached below (Table 4) shows the probabilities of the temperature in a certain month being within a given interval, these probabilities were calculated by analysing the values of average temperatures on every day of each month and were then organized into intervals which were later averaged out to give out the needed probabilities.

	[20-23,5]	[23,5-27]	[27-30,5]	[30,5-34]
Jan	0,02	0,71	0,27	0
Feb	0	0,64	0,36	0
Mar	0	0,3	0,7	0
Apr	0	0,28	0,72	0
May	0	0,11	0,89	0
Jun	0	0,15	0,84	0,01
Jul	0	0,26	0,74	0
Aug	0	0,25	0,75	0
Sep	0	0,29	0,71	0
Oct	0	0,31	0,69	0
Nov	0	0,64	0,36	0
Dec	0	0,76	0,24	0

Table 4. showing the probabilities that the temperature in each month will be in a certain interval based on the statistical data spanning over 7 years

Extrapolation from a graph

Complementary method used to determine the temperature for future years from the given data was that searching the graph of previous years and individual values for temperature in a certain month and extrapolating the value for upcoming years (in this case 2018.) by using the extrapolation method. The first step was to find the specific average temperatures for a certain month over the course of a couple of years. Then the years versus the temperature values were plotted (Graph 2).



Graph 2. showing the graph plotted of average annual temperatures for each month

After the graph was plotted the software was used to determine the best fit line and the equation of that line in the form y=mx+b ,where X was the variable (in this case the year) which could be put in to get the wanted temperature value. When the graphs were plotted for all the months the data was organised in the table and the value for the temperatures in 2018 were determined (Table 5).

	m	b	x	avg. Temp in 2018
january	0,0007637	24,28	2018	25,82
february	0,017	-8,272	2018	26,34
march	0,00603	14,85	2018	27,02
april	0,003049	21,37	2018	27,52
may	0,00687	14,37	2018	28,23
june	0,007335	13,78	2018	28,19
july	0,033	-38,9	2018	27,69
august	0,001617	24,77	2018	28,30
september	0,01776	-8,109	2018	27,73
october	0,038	-49,36	2018	27,32
november	0,01544	-4,653	2018	26,50
december	0.0083	9.5	2018	26.25

Table 5.

Discussion

As was visible from Table 1, measured temperatures perfectly matched their corresponding predicted intervals. A conclusion can be drawn that the method excelled at its purpose, however it is advisable to note that

1) Precision was sacrificed for accuracy and practicality. Predictions would presumably turn out less reliable had it not been so.
2) Temperatures in Jurong, Singapore were not the ideal testing data for the purpose of this experiment. Matrices 1 and 2 show transition probabilities for minimum and maximum temperatures. Because of highest probability in place 3,3 it is evident that temperatures tend to remain high once they get that way. A location with extremely stable high temperatures wasn't challenging enough for the model to prove its worth, especially with the starting temperatures being high as well. Choosing it is the admitted weakness of the method.

$$\begin{bmatrix} .16 & .42 & .42 \\ .13 & .34 & .53 \\ .1 & .24 & .67 \end{bmatrix}$$ Matrix 1.

$$\begin{bmatrix} .33 & .22 & .45 \\ .07 & .44 & .49 \\ .24 & .24 & .52 \end{bmatrix}$$ Matrix 2.

The aforementioned slight adjustments that were made in one place were not perceived to be indicative of any major flaw in the procedure, although one that could be fully automatized (one which would not require human insight in order to be optimized) would be preferable.

Initially, the idea was to link multiple weather components in a Markov chain, and find transition probabilities between states that were combinations of different categories of them (e.g. one would represent the transition between low rainfall and high temperature, to medium rainfall and medium temperature). This was conceptually feasible, but impractical due to its low precision and the amount of work required for computation.

The case is similar in analysis on a greater scale: although the method yielded relatively accurate and precise results there are multiple factors which affected

the calculations, and if eliminated could improve the results in future analysis of such systems.

The first of this numerous factors is, again, the geographical location. The weather does not change much over the years and all the factors tend to remain constant over the months.

Another improvement could be gathering data over a longer period - more complex weather patterns could be observed with higher precision.

Overall, the investigation was a success in a sense that accurate results for the temperatures were obtained, which was the primary goal. The initial premise that weather could be statistically analysed and its parameters probabilistically calculated by looking only at a narrow set of data can be considered to be true in the circumstances of our experiment.

References

WolframAlpha, Widgets, Matrix Calculator. ©2018 Wolfram Alpha LLC http://www.wolframalpha.com/widgets/view.jsp?id=abec16f483abb4f1810ca029aadf8446 (Accessed 2018-08-18)

The R Project for Statistical Computing, c/o Institute for Statistics and Mathematics. © The R Foundation https://www.r-project.org/ (Accessed 2018-08-14)

The Meteorological Service Singapore (MSS) http://www.weather.gov.sg/climate-historical-daily (Accessed 2018-07-30)