Forecasting the long-term monthly variations of major floods

MARIO LEFEBVRE
Department of Mathematics and Industrial Engineering
Polytechnique Montréal
2500, chemin de Polytechnique, Montréal (Québec) H3T 1J4
CANADA

Abstract: The monthly variations of major floods are modelled as a discrete-time Markov chain. Based on this stochastic process, it is possible, with the help of real-life data, to forecast the future variations of these events. We are interested in the duration of the floods and in the area affected. By dividing the data set into two equal parts, we can try to determine whether there are signs of the effects of climate change or global warming.

Key-Words: Markov chains, limiting probabilities, climate change.

Received: April 13, 2021. Revised: March 10, 2022. Accepted: April 13, 2022. Published: May 5, 2022.

1 Introduction

In [4], the author modelled the monthly variations of major floods worldwide as a discrete-time Markov chain having three possible states. Similarly, this stochastic process was used as a model for the monthly or yearly variations of earthquakes. In both cases, using real-life data, it was found that the models that were proposed were indeed appropriate to describe the evolution of these events.

Moreover, the limiting probabilities of the Markov chains were computed, in order to forecast the long-term behaviour of the processes. Rather surprisingly, the author concluded that the major floods were seemingly occurring almost at random and did not show signs of increase due to climate change, whereas earthquakes, and especially major ones, were trending upwards.

Markov chains have been used by other authors as models in various applications. In hydrology, Avilés *et al.* [1] forecast drought events based on these stochastic processes, while Matis *et al.* [5] used them to forecast cotton yields. Drton *et al.* [2] proposed a Markov chain to model tornadic activity.

Similarly, Markov or semi-Markov processes often served as models to forecast earthquakes; see, for instance, Sadeghian [8] and Panorias *et al.* [6].

Now, in [4], in the case of major floods, the variable of interest was their number per month. There are however other variables that can be considered. In the current paper, we will study two such variables, namely the total duration of the floods and the total area affected.

As in [4], the data set used will also be divided into two equal parts to determine whether there have been some significant changes in the variations of major floods during the period considered.

In the next section, the mathematical background will be presented. The model will then be implemented for the total duration of the floods and the total area affected in Sections 3 and 4, respectively.

2 Mathematical background

We will briefly recall the mathematical results needed to carry out our study. See also [3] or [4].

A (time-homogeneous) discrete-time Markov chain is a stochastic process $\{X_n, n = 0, 1, 2, \dots\}$ such that

$$P[X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]$$
$$= P[X_{n+1} = j \mid X_n = i] := p_{i,j}$$

for all states $i_0, \dots, i_{n-1}, i, j$ in the state space S and for any n. In this paper, we will assume that the state space of the process is the finite set $S = \{0, 1, 2\}$. Hence, we assume that for any n, X_n is equal to one of the numbers 0, 1 or 2, which are actually a coding system. The matrix $\boldsymbol{P}$ of the various $p_{i,j}$'s is the transition matrix of the Markov chain.

In the case of a discrete-time Markov chain, the states i and j can be the same in $p_{i,j}$. If we denote by K_i the number of time units that the chain spends in state i before moving to a different state, then (by independence) we can write that

$$P[K_i = k] = (p_{i,i})^{k-1}(1 - p_{i,i})$$

for $k = 1, 2, 3, \dots$ That is, the random variable K_i has a *geometric distribution* with parameter $p :=$

$1 - p_{i,i}$. Notice that the above probability is strictly decreasing with k.

Next, we define the limiting probability that the Markov chain will be in state i when it is in equilibrium:

$$\pi_i = \lim_{n\to\infty} P[X_n = i].$$

Under some conditions that will clearly be fulfilled in our case, we can show (see, for instance, [7]) that the limiting probabilities exist and can be obtained by solving the following system of linear equations:

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}, \tag{1}$$

where $\boldsymbol{\pi} := (\pi_0, \pi_1, \pi_2)$, subject to the condition

$$\sum_{i=0}^{2} \pi_i = 1. \tag{2}$$

In the next section, the total duration of the major floods that occurred during a given month will be considered.

3 Total duration of the floods

Let F_n be the number of floods during month n. In [4], the author defined the following three states for the variable X_n:

$$\begin{array}{ll} 0: & \text{if } F_n - F_{n-1} < -2, \\ 1: & \text{if } -2 \leq F_n - F_{n-1} \leq 2, \\ 2: & \text{if } F_n - F_{n-1} > 2. \end{array}$$

Making use of the data set found on the site flood-observatory.colorado.edu, which gives a list of large flood events worldwide from 1985, it was found that the stochastic process $\{X_n, n = 2, 3, \ldots\}$ can be considered as a Markov chain.

The data for the years 2000 to 2016 were used in the study. There are 2825 floods in the data set for this period, so that the average number of floods per month is 13,85.

For each flood, the data set provides the dates when it began and ended, its magnitude, the number of dead, the area affected, etc. The magnitude of a flood is a number defined by

$$M = \text{Log}(\text{Duration} \times \text{Severity} \times \text{Affected Area}),$$

in which the Duration is in days, the Affected Area is in square kilometres and the Severity is equal to 1, 1,5 or 2 for large, very large and extreme events, respectively. For the definition of the variable Severity, see the site http://floodobservatory.colorado.edu/Archives/ArchiveNotes.html. A flood having an M greater than 4 (respectively 6) is considered as *severe* (respectively *very severe*). The vast majority of the floods in the data set are at least severe.

The estimated transition matrix was found to be

$$\boldsymbol{P} = \begin{pmatrix} 1/6 & 19/66 & 6/11 \\ 9/34 & 27/68 & 23/68 \\ 37/68 & 23/68 & 2/17 \end{pmatrix},$$

from which we obtain the following limiting probabilities:

$$\pi_0 = 0{,}3257, \quad \pi_1 = 0{,}3420, \quad \pi_2 = 0{,}3324.$$

As mentioned above, we must therefore conclude rather surprisingly that, in the long run, the three states of the Markov chain are almost equally likely. Furthermore, we find that the average value of the differences $F_n - F_{n-1}$ is 0,0345. Thus, the monthly variations of the number of major floods do not show any trend during the period 2000-2016. This conclusion is strengthened when we divide the data set into two parts (from 2000 to 2007, and from 2008 to 2016) and we calculate the corresponding limiting probabilities; see Table I.

Table I: Limiting probabilities calculated for the periods 2000-2007 and 2008-2016.

Period	π_0	π_1	π_2
2000-2007	0,3368	0,3263	0,3368
2008-2016	0,3149	0,3575	0,3275

Indeed, the π_i's did not change much between the two time periods, and are consequently close to the values obtained for the whole period. Actually, we see that there are less variations during the period 2008-2016, because state 1 then has the largest limiting probability. This is confirmed by the fact that the standard deviation of the monthly variations decreased from 7,54 (in 2000-2007) to 5,90 (in 2008-2016). Finally, the mean also decreased, from 0,116 to $-0{,}037$.

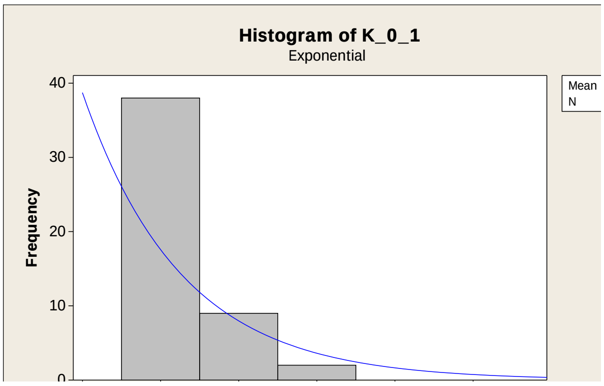
Now, although the *number* of monthly major floods appears to be quite stable, there are other variables related to floods that are important. In this section, we consider the total duration of the floods that started during a given month.

Let M_n be the total duration of the floods that started during month n. As in the case of the number of floods, we define three states for the stochastic process $\{X_n, n = 2, 3, \ldots\}$. We write that X_n is equal to

$$\begin{array}{ll} 0: & \text{if } M_n - M_{n-1} < -50, \\ 1: & \text{if } -50 \leq M_n - M_{n-1} \leq 50, \\ 2: & \text{if } M_n - M_{n-1} > 50. \end{array}$$

Using the data for the whole time period 2000-2016, we first obtain the histograms for the variables K_0, K_1 and K_2 defined above. These histograms are shown in Figures 1 to 3, respectively.

As we can see, the histograms present approximately the exponential decrease that should be observed if the random variable K_i, for $i = 0, 1, 2$, has a



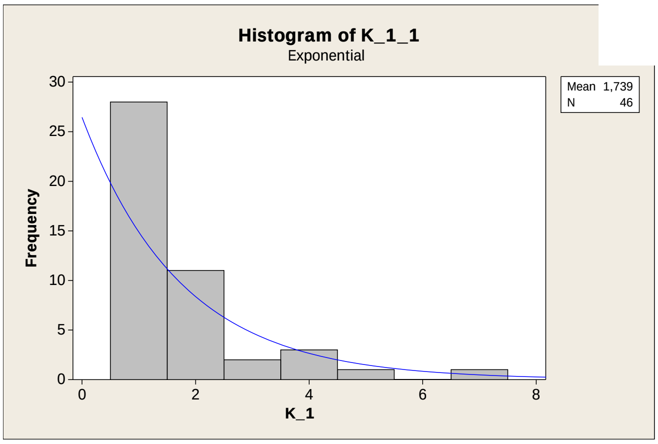


Figure 2: Histogram of the variable K_1 in the case of the total duration of the floods.

geometric distribution. Therefore, we may conclude that assuming that $\{X_n, n = 2, 3, \ldots\}$ is a Markov chain is realistic.

Next, we can easily estimate the transition probabilities $p_{i,j}$, for $i, j \in \{0, 1, 2\}$. We find the following estimated transition matrix:

$$\boldsymbol{P} = \begin{pmatrix} 13/64 & 26/64 & 25/64 \\ 19/79 & 34/79 & 26/79 \\ 31/59 & 20/59 & 8/59 \end{pmatrix}.$$

Then, solving the system (1), (2), we obtain the limiting probabilities:

$$\pi_0 = 0{,}3120, \quad \pi_1 = 0{,}3962, \quad \pi_2 = 0{,}2918.$$

Moreover, we have the following descriptive statistics of the 203 differences $M_n - M_{n-1}$:

$$\bar{x} = -0{,}276 \quad \text{and} \quad s = 160{,}4.$$

Hence, as in the case of the number of major floods, we must come to the conclusion that the duration of the floods is quite stable. There is in fact a very slight decrease in the average total duration of the floods, and π_0 is larger than π_2.

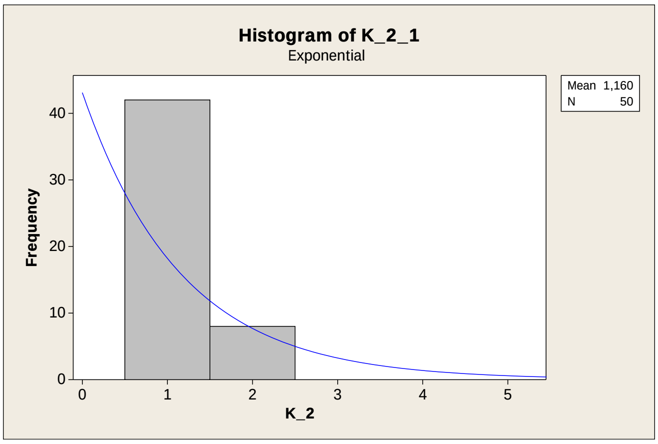


Figure 3: Histogram of the variable K_2 in the case of the total duration of the floods.

To complete this section, we compute the limiting probabilities for two equal subsets of the data set: first from January 2000 to June 2008, and then from July 2008 to December 2016. The results are presented in Table II.

Table II: Limiting probabilities for the total duration of the floods calculated for the periods I: January 2000 to June 2008, and II: July 2008 to December 2016.

Period	π_0	π_1	π_2
I	0,2803	0,4098	0,3099
II	0,3431	0,3824	0,2745

Since the value of π_0 has increased very significantly in the second time period considered, it is possible to state that, not only the total duration of the major floods shows no sign of increase, it actually seems to be decreasing. However, the descriptive statistics of the differences are

$$\bar{x}_I = -0{,}833 \quad \text{and} \quad s_I = 192{,}2,$$

and

$$\bar{x}_{II} = 0{,}287 \quad \text{and} \quad s_{II} = 121{,}2.$$

Thus, the average difference increased slightly (about 1,12 days, or 26,88 hours), but the standard deviation is much smaller during the second part of the period considered.

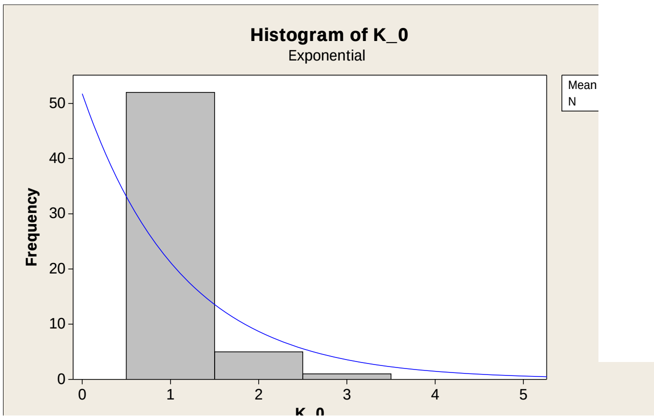
In the next section, we will turn to the total area affected by the floods.

4 Total area affected by the floods

Let A_n be the total area (in 10^6 square kilometres) affected by the floods during month n. We define the following states for the random variable X_n:

$$\begin{array}{ll} 0: & \text{if } A_n - A_{n-1} < -5, \\ 1: & \text{if } -5 \leq A_n - A_{n-1} \leq 5, \\ 2: & \text{if } A_n - A_{n-1} > 5. \end{array}$$

The histograms for the variables K_0, K_1 and K_2 obtained for the time period 2000-2016 are presented in Figures 4 to 6, respectively. Since the three ran-



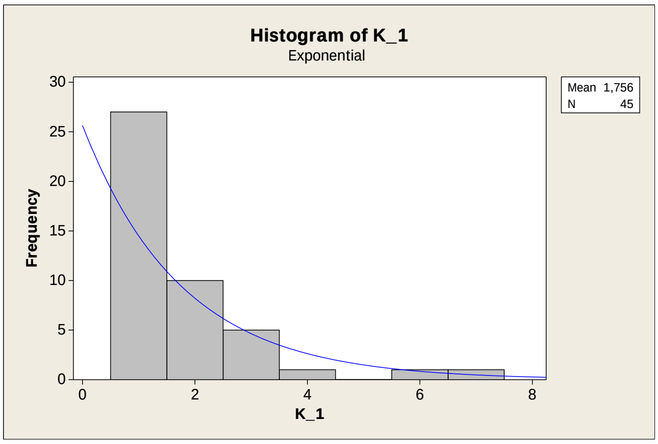


Figure 5: Histogram of the variable K_1 in the case of the total area affected by the floods.

We find that the estimated transition matrix $\boldsymbol{P}$ is

$$\boldsymbol{P} = \begin{pmatrix} 7/62 & 32/62 & 23/62 \\ 17/82 & 37/82 & 28/82 \\ 38/58 & 12/58 & 8/58 \end{pmatrix},$$

from which we estimate the limiting probabilities:

$$\pi_0 = 0{,}3086, \quad \pi_1 = 0{,}4001, \quad \pi_2 = 0{,}2913.$$

We see that the limiting probabilities are very close to the ones computed in the previous section. Therefore, we must again conclude that there is no sign of upward or downward trend for the monthly total area affected by the floods. When we divide the data set into two equal parts, we obtain the values in Table III.
Table III: Limiting probabilities for the total area affected by the floods calculated for the periods I: January 2000 to June 2008, and II: July 2008 to December 2016.

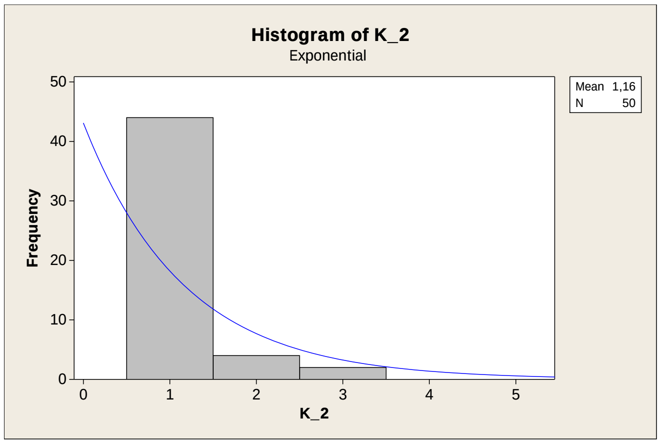


Figure 6: Histogram of the variable K_2 in the case of the total area affected by the floods.

Period	π_0	π_1	π_2
I	0,2934	0,4171	0,2893
II	0,3235	0,3824	0,2941

We observe an increase (respectively a decrease) of π_0 (respectively π_1), and a slight increase of π_2, which implies that there are more variations in the second part of the time period considered than in the first one. However, the limiting probabilities are rather stable.

The main descriptive statistics of the differences $A_n - A_{n-1}$ are presented in Table IV.

Table IV: Descriptive statistics of the monthly differences $A_n - A_{n-1}$.

Period	$\bar{x}$	s
01/01 – 12/16	0,0791	15,63
01/01 – 06/08	0,1380	15,87
07/08 – 12/16	0,0207	15,46

We see that the average and the standard deviation of the monthly variations are quite stable, with a small decrease in each case.

5 Conclusion

In this paper, we continued the study of the monthly variations of the major floods worldwide that was started in [4]. In the previous paper, it was found that the number of major floods does not show any sign of upward trend during the period 2000-2016. In the current paper, we considered two important characteristics of the floods, namely their duration and the area affected. In both cases, the conclusion was the same as in [4]. Indeed, we observe a slight increase in the monthly variations, but the most likely state of the Markov chain remains the one that corresponds to small variations of the variable of interest. This conclusion is strengthened when we compute the descriptive statistics of the two variables. We find that the mean of the observations is close to zero.

We also considered the number of people who died because of the floods. This variable is more volatile, because it depends in particular on the countries that were affected by the floods. At any rate, if we denote by D_n the total number of dead during month n, we find that the average of the differences $D_n - D_{n-1}$ decreased during the period 2000-2016: it went from 2,35 between January 2000 to June 2008, to -10,60 between July 2008 to December 2016. Thus, again we see no sign of upward trend.

Acknowledgments. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References:

[1] Avilés, A., Célleri, R., Solera, A. and Paredes, J., Probabilistic forecasting of drought events using Markov chain- and Bayesian network-based models: A case study of an Andean regulated river basin, *Water*, Vol. 8, No. 2, 2016, 16 pages. DOI: 10.3390/w8020037

[2] Drton, M., Marzban, C., Guttorp, P. and Schaefer, J. T., A Markov chain model of tornadic activity, *Monthly Weather Review*, Vol. 131, 2003, pp. 2941–2953. DOI: 10.1175/1520-0493(2003)131<2941:AMCMOT>2.0.CO;2

[3] Lefebvre, M., Modelling and forecasting temperature and precipitation in Italy, *Atti della Accademia Peloritana dei Pericolanti - Classe di Scienze Fisiche, Matematiche e Naturali*, Vol. 97, No. 2, A2, 2019, 9 pages. DOI: 10.1478/AAPP.972A2

[4] Lefebvre, M., A Markov chain model for floods and earthquakes, *Proceedings of the 56th ESReDA Seminar*, Linz, Austria, May 23-24, 2019, pp. 46–55. (Available online)

[5] Matis, J. H., Birkett, T. and Boudreaux, D., An application of the Markov chain approach to forecasting cotton yields from surveys, *Agricultural Systems*, Vol. 29, No. 4, 1989, pp. 357–370. DOI: 10.1016/0308-521X(89)90097-8

[6] Panorias, C., Papadopoulou, A. and Tsapanos, T., On the earthquake occurrences in Japan and the surrounding area via semi Markov modeling, *Bulletin of the Geological Society of Greece*, Vol. 50, No. 3, 2016, pp. 1535–1542. DOI: 10.12681/bgsg.11866

[7] Ross, S. M., *Introduction to Probability Models*, 12th Edition, Amsterdam, Elsevier/Academic Press, 2019. DOI: 10.1016/C2017-0-01324-1

[8] Sadeghian, R., Forecasting time and place of earthquakes using a semi-Markov model: With case study in Tehran province, *Journal of Industrial Engineering International*, Vol. 8, 2012, pp. 1–7. DOI: 10.1186/2251-712X-8-20