Markov chain as a tool for forecasting daily precipitation in the vicinity of the city of Bydgoszcz, Poland

Waldemar Bojar[1], *Leszek* Knopik[1], *Jacek* Żarski[2], *Renata* Kuśmierek-Tomaszewska[2,*], and *Wojciech* Żarski[1]

[1]UTP University of Science and Technology, Faculty of Management, Engineering Management Department, ul. Fordońska 430, 85-790 Bydgoszcz, Poland
[2]UTP University of Science and Technology, Faculty of Agriculture and Biotechnology, Department of Agrometeorology, Plant Irrigation and Horticulture, ul. Bernardyńska 6, 85-029 Bydgoszcz, Poland

Abstract. The crop yield depends on numerous weather factors, but mainly on the rainfall pattern and course of air temperature during vegetation period. Investigating the dependence of yields on rainfall, apart from its amount, there also should be taken into account dry spell periods. The two-state Markov chain was considered as a precipitation pattern in the investigation, since it is generally recognized as a simple and effective model of the precipitation occurrence. Based on the daily precipitation totals from the period 1971–2013, the Markov chain was designated. The data were derived from a measuring point of the University of Science and Technology in Bydgoszcz, Poland. As one of the objectives was to determine the order of the Markov chain examined describing the change of precipitation in subsequent days. Another aim was to investigate rainfall dependencies on a month of a year. An analysis of this data leads to the conclusion that the chain is second order. This is confirmed by the two criteria used: BIC (Bayesian Information Criteria) and AIC (Akaike Information Criteria). The research regarded the precipitation volume dependence on a month of the year.

1 Introduction

Estimating and predicting precipitation is a problem of fundamental importance for agriculture, hydrology and ecology. Information on the probability of precipitation, its size and the number of days without rainfall, is necessary for designing sanitary systems for draining rainwater, or planning irrigation systems as an alternative system of growing plants in order to build a rational management of water in the soil [1]. Determining the distribution of rainfall is also necessary for planning the use of water resources on a larger scale, eg in the national economy. The management of water resources in a given area is always based on historical, current and future meteorological data at various time scales: annual, monthly and daily. The forecast of daytime rainfall is an extremely difficult since atmospheric precipitation, as a meteorological element, is a very complex phenomenon and changes with time and place [2]. It is one of the most unpredictable events even for updated climate models. The range of uncertainty varies depending on the model, the physical characteristics of the atmosphere and the complexity associated with its mathematical modeling.

Cazacioc and Cipu [3] point out three methods of precipitation forecasting: subjective forecasts based on the experience of forecasters, deterministic forecasts obtained from numerical weather forecast models as well as statistical forecasts. The two latter tools are the most objective. Statistical models contribute to a large extent to reducing the uncertainty resulting from the complexity of the phenomenon of precipitation.

Among the statistical techniques of atmospheric precipitation modeling are Markov chains, used to predict short-term rainfall. The use of Markov chains for rainfall modeling was introduced over 40 years ago [4–6]. In spite of general simplicity, the model of the first order Markov chain (based on rainfall from the previous day) remains a suitable technique for modeling rainfall data in many geographical areas [7]. Markov chain models have two advantages: forecasts are available immediately after completing the observation, because they use only local weather information as predictors and require minimal calculations after processing climatological data.

Markov chains determine the state of each day as "wet" or "dry" and allow explaining the relationship between the current day's state and the previous day. The order of the chain is the maximum number of days preceding the day on which the state of the current day depends. Most of the Markov chain models mentioned in the literature are first-order models [8–10]. Some researchers [11–13] point out the inconvenience in the use of first-order models that underestimate the length, frequency and variability of precipitation, which is why higher-order models are recommended. The two state Markov chain, as a model of rainfall, was studied by Gabriel and Neumann [5] and was generalized by Teodorovic and Woolhiser [14] and Katz [8]. Many authors use Markov chains to model the occurrence of daily rainfall [15-20]. The low order is most often

* Corresponding author: rkusmier@utp.edu.pl

preferred for two reasons: the number of parameters is kept to a minimum so as to obtain a better estimate, moreover, the latter use of a tailored model to calculate other variables, such as the probability of long dry (rain free) periods, is much simpler.

The distribution of the number of days with precipitation usually has gamma distribution characteristics. In the analysis of pluviometric conditions, two characteristic features of rainfall are usually distinguished: their occurrence and their height or intensity. In modeling, they can be considered together or separately. This paper only discusses the occurrence of precipitation, more precisely modeling daily rainfall.

The objective of the paper was basic statistical analysis of daily rainfall and, in particular, determining the order of the analyzed Markov for modeling everyday rainfall phenomena in the region of the city of Bydgoszcz, Poland. In addition, the dependencies of precipitation totals on the month of the year were examined.

2 Material and research methods

The paper presents a statistical analysis of data on daily rainfall totals recorded in the January-December period in the years 1971–2013 at the meteorological station of the Faculty of Agriculture and Biotechnology of the University of Science and Technology in Bydgoszcz at the Research Center located in the agricultural area of Mochle, about 17 km away from Bydgoszcz.

The basic methods of statistical analysis used in the work are methods related to estimation of Markov chain parameters. In particular, the matrix of transition probabilities for the Markov chain of precipitation in the selected years was estimated and the stationary probabilities were determined. The basic statistical operation was to determine the order of the Markov chain. In order to achieve this, two criteria of determining the chain order: the BIC (Bayesian Information Criteria) [21] and the AIC (Akaike Information Criteria) [22]. Both are based on the log-likelihood functions for transition probability of the Markov chain constructed on certain data series.

3 Markov chain

The simplest kinds of discrete variable is that which has binary values (1 / 0) corresponding to two state in which it can exist. For daily precipitation, those are their occurrence or non-occurrence. A sequence of daily observations from meteorological station constitutes time series of that discrete variable.

For the first order Markov chain, the transition probability to future state depends only on its current state. Knowing that at day i the variable X is either in state 0 (no precipitation occurs and X(i) = 0), or state 1 (precipitation occurs and X(i) = 1). It may be assumed that

$$P\{X(n+1) = x_{n+1} \mid X(n) = x_n, X(n-1) = x_{n-1}, \ldots, X(0) = x_0\} \quad (1)$$

$$= P\{X(n+1) = x_{n+1} \mid X(n) = x_n\} \text{ for any } x_0, x_1, \ldots, x_n, x_{n+1} \in \{0, 1\}. \quad (2)$$

Stochastic process X(n), n = 0, 1, 2, … is called Markov chain. Conditional transition probability at day i + 1 by one step is defined as

$$p_{00} = P\{X(i+1) = 0 \mid X(i) = 0\} \quad (3)$$

$$p_{01} = P\{X(i+1) = 1 \mid X(i) = 0\} \quad (4)$$

$$p_{10} = P\{X(i+1) = 0 \mid X(i) = 1\} \quad (5)$$

$$p_{11} = P\{X(i+1) = 1 \mid X(i) = 1\} \quad (6)$$

It is easy see that $p_{00} + p_{01} = 1$ and $p_{10} + p_{11} = 1$. Matrix P of transition probabilities defined as:

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} \quad (7)$$

is called transition matrix in one step.

The model is fully defined by two transition probabilities: p_{01} (the probability that precipitation will occur tomorrow if precipitation did not occur today) and p_{11} (the probability that precipitation will occur tomorrow if precipitation occurred today). These probabilities can easily be computed from observed precipitation occurrence time series. Their maximum-likelihood estimates p_{01} and p_{11}, are given by

$$p_{01} = n_{01} / (n_{00} + n_{01}) \quad (8)$$

$$p_{11} = n_{11} / (n_{10} + n_{11}) \quad (9)$$

where n_{01} is the historical count of wet days that followed dry days, n_{00} is the historical count of dry days that the followed dry days and so on.

For a Markov chain describing the daily occurrence or non-occurrence of precipitation, the stationary probability π_1 for precipitation, corresponds to the unconditional probability of precipitation is given by formula:

$$\pi_1 = p_{01} / (1 + p_{01} - p_{11}) \quad (10)$$

Analogously for unconditional probability π_0 is

$$\pi_0 = (1 - p_{11}) / (1 + p_{01} - p_{11}) \quad (11)$$

3.1. Estimation of matrix P for 1971 – 2013

Matrix for years 1971 – 1980 is as follows:

$$P_1 = \begin{bmatrix} 0.746 & 0.253 \\ 0.492 & 0.508 \end{bmatrix} \quad (12)$$

For years 1981– 1999:

$$P_2 = \begin{bmatrix} 0.732 & 0.268 \\ 0.494 & 0.506 \end{bmatrix} \quad (13)$$

For years 1981 – 1999:

$$P_3 = \begin{bmatrix} 0.712 & 0.288 \\ 0.477 & 0.523 \end{bmatrix} \quad (14)$$

For years 2001 – 2010:

$$P_4 = \begin{bmatrix} 0.737 & 0.263 \\ 0.583 & 0.417 \end{bmatrix} \quad (15)$$

For years 2011 – 2013

$$P_5 = \begin{bmatrix} 0.741 & 0.263 \\ 0.590 & 0.411 \end{bmatrix} \quad (16)$$

The analysis of the matrices P_1, P_2, P_3, P_4 and P_5 of transfer probabilities shows respectively differences between the values of probabilities. The determined matrices are characterized by the succession of days without precipitation and days with precipitation. Table 1 contains the values of stationary probabilities π_1 corresponding to the above matrices

Table 1. Stationary probabilities π_1 determined for five of the analyzed.

Years	π_1
1971 – 1980	0.69
1981 – 1990	0.65
1991 – 2000	0.62
2001 – 2010	0.69
2011 – 2013	0.69

The lowest border probability has been observed for the decade of 1991 – 2000. The following problem arises: has this fact been confirmed by other studies.

4 Higher Order Markov Chains

Let X(i), n = 1, 2, …, n is observed the time series, n_0 number 0's, n_1 number 1's in this sequences, and $p_0 = n_0 / n$, $p_1 = n_1 / n$.

First, consider for instance a second – order Markov chain. Then transition probability for second-order Markov chain depends on the state i – 1, i, i + 1.
Transition probability of second order Markov chain can be defined as

$$p_{krs} = P\{X(i+1) = s \mid X(i) = r, X(i-1) = k\}, \text{ for } k, r, s \in \{0, 1\} \quad (17)$$

Transition probabilities p_{krs} is estimated by formula

$$p_{krs} = n_{krs} / n_{kr\bullet} \quad (18)$$

where

$$n_{krs\bullet} = \sum n_{krst}. \quad (19)$$

is number of realization of transition between states

$$\{x(i-1) = k\} \rightarrow \{x(i) = r\} \rightarrow \{x(i+1) = s\} \quad (20)$$

in considered time series.

Accordingly, probabilities p_{krst}. are thus defined

$$p_{krst} = P\{X(i+1) = t \mid X(i) = s, X(i-1) = r, X(i-2) = k \quad (21)$$

for k, r, s, t ∈ {0, 1}.

By n_{krst} is denoted number of realization of transition states

$$\{x(i-2) = k\} \rightarrow \{x(i-1) = r\} \rightarrow \{x(i) = s\} \rightarrow \{x(i+1) = t\}. \quad (22)$$

Transition probabilities p_{krst} are estimated by formula

$$p_{krst} = n_{krst} / n_{krs\bullet} \quad (23)$$

These log-likelihood functions depend on the transition count and the estimated transition probabilities. The log-likelihood functions for Markov chain of the order 0, 1, 2 and 3 are

$$L_0 = \sum_{j=0}^{s-1} n_j \ln(\hat{p}_j) \quad (24)$$

$$L_1 = \sum_{i=0}^{s-1}\sum_{j=0}^{s-1} n_{ij} \ln(\hat{p}_{ij}) \quad (25)$$

$$L_2 = \sum_{h=0}^{s-1}\sum_{i=0}^{s-1}\sum_{j=0}^{s-1} n_{hij} \ln(\hat{p}_{hij}) \quad (26)$$

$$L_3 = \sum_{g=0}^{s-1}\sum_{h=0}^{s-1}\sum_{i=0}^{s-1}\sum_{j=0}^{s-1} n_{ghij} \ln(\hat{p}_{ghij}) \quad (27)$$

Summations in this formula are performed over all the state s of the Markov chain. Statistics of criteria AIC(m) and BIC(m) can be written as

$$AIC(m) = -2\, L_m + 2\, s^m (s-1) \quad (28)$$

$$BIC(m) = -2\, L_m + s^m (\ln(m)) \quad (29)$$

The order m is chosen as appropriate that minimizes the functions (29) and (30).

Letting s = 2, the value of statistics AIC(m) and BIC(m) is calculated in Table 2.

Table 2. Order choices based on AIC and BIC.

Order	**AIC**	**BIC**
0	1125.945	1123.945
1	1107.179	1103.179
2	1088.614	1083.387
3	1094.694	1087.487

An analysis of Table 2 allows one to conclude that both AIC and BIC indicate the choice of 2nd order for the examined Markov chain.

5 Probability distribution of daily precipitation

The results of the basic statistical analysis of daily rainfall recorded in the years 1971 - 2013 are presented below. During this period, out of the 15,706 analyzed days, n = 5,367 days with atmospheric precipitation. Days with precipitation accounted for 5,367/1,5706 = 34.2% of all analyzed cases. The first to be examined was the dependence of precipitation on the month of the year. The length of the series for each month is n = 42. The graph below (Fig. 1) presents the dependences of the mean value of precipitation on the month with the confidence interval, for the mean value with the confidence level 1 - α = 0.95.

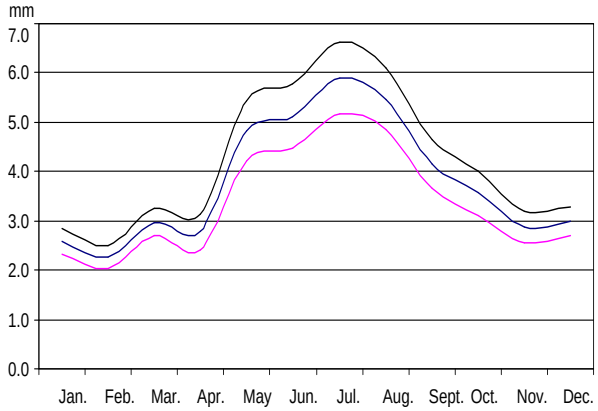

Fig. 1. Mean values of precipitation for each month including confidence intervals.

Source data in Fig. 1 as well as the results of calculations of basic statistics are presented in Table 3. It includes the values for basic statistics for each month of the year:

- mean value – M
- left side confidence interval – LSCI
- right side confidence interval – RSCI
- standard deviation – SD
- coefficient variable – CV
- maximum value – MAX
- sample size – N

The analysis of the graph (Fig. 1) as well as the results presented in Table 3 show that the highest rainfall occurs in the months: May, June, July, August and September, while the lowest in February. The longest confidence interval and the highest mean values are characteristic for the period of May, June, July and August. Whereas for December, January, February, March and April, the intervals of confidence are the smallest, the highest variability of daily rainfall was found in June and July. Similarly, the most rainy days were recorded in these months. The confidence interval for the average value was determined on the basis of a statistical sample of a large number, but a relatively large spread of daily totals of precipitation (all coefficients of variation are greater than or equal to 1) implies a relatively large width of this interval.

Table 3. Values of basic statistics for daily precipitation.

Month	M	LSCI	RSCI	SD	CV	MAX	N
Jan.	2.58	2.32	2.83	2.81	1.09	25.20	450
Feb.	2.27	2.03	2.51	2.39	1.06	12.60	375
Mar.	2.97	2.69	3.26	2.94	0.99	22.40	403
Apr.	2.78	2.41	3.15	3.77	1.36	36.20	400
May	4.82	4.22	5.42	6.36	1.32	56.50	430
Jun.	5.12	4.47	5.78	7.37	1.44	84.60	490
Jul.	5.90	5.18	6.62	8.53	1.45	78.00	539
Aug.	5.47	4.84	6.09	6.72	1.23	52.20	448
Sept	4.16	3.65	4.66	5.48	1.32	36.50	451
Oct.	3.56	3.11	4.01	4.72	1.33	32.00	422
Nov.	2.87	2.56	3.18	3.48	1.21	23.00	482
Dec.	2.99	2.70	3.27	3.18	1.06	22.10	477

6 Probability distribution of daily precipitation

For each month the dependence of the volume of daily rainfall on the year was examined. For all of the 12 months no statistically significant dependence has been confirmed. This means that for the years 1971-2013 the mean total of all daily rainfall neither rose, nor fell.

7 Concluding remarks

A high variability of rainfall over time can lead to the occurrence of atmospheric drought, and when the off-season occurs in the growing season, it takes the form of agricultural droughts, which result in low yields [23]. Analyses of the variability of precipitation summed up with the use of Markov chains, not only in the theoretical context, but also applied in various fields of the economy were conducted by many authors. The obtained rainfall models included various time scales [24–27]. Daily models have gained widespread use as suitable for use in a detailed water balance as well as agricultural and environmental models [26]. For daytime precipitation modeling at one point, Stern and Coe [28] used a second order Markov chain to describe precipitation and gamma distribution to forecast the amount of rainfall. The results of the analysis of the Markov chains may be implemented in the forecast of rainfall in a given vegetative season [29, 30] or for irrigation scheduling [31]. On the basis of climate reports and future climate change scenarios, one may expect a demand to monitor and forecast rainfall on the basis on simple statistic tools.

8 Summary

The conducted analysis confirmed that daily rainfall in the area of Bydgoszcz is characterized by high variability in individual months and years. This model of daily rainfall was found to adhere to the second-order Markov chain, which was confirmed by two applied criteria: BIC (Bayesian Information Criteria) and AIC

(Akaike Information Criteria). However, such an extensive statistical material failed to confirm the existence of the dependence of precipitation on the particular year. On the other hand, a strong dependence of the amount of daily rainfall on the individual month has been confirmed.

Research was made within the framework of the FACCE JPI – MACSUR project titled: Modelling European Agriculture with Climate Change for Food Security Acronym FACCE MACSUR 2 realized between 01/06/2015 and 31/05/2017

References

1. K. M. Barkotulla, Pakistan J. Stat. Operat. Res. **6**(1) (2010)
2. J. Żarski, S. Dudek, R. Kuśmierek-Tomaszewska, R. Rolbiecki, S. Rolbiecki, Ann. Set Envir. Protect. **15** (2013)
3. L. Cazacioc, E. C. Cipu, *Mathematics in Engineering and Numerical Physics* (Bucharest, Romania, 2004)
4. E. H. Chin, Water Resour. Res. **13**(6) (1977)
5. R. Gabriel, I. Neuman, Q J of Royal Meteor. Soc. **88** (1962)
6. R.W. Katz, J Appl. Meteorol. **16** (1977)
7. J.T. Schoof, S.C. Pryor, J. Appl. Meteor. Climatol. **47** (2008)
8. R. W. Katz, J. Appl. Probab. **14**(3) (1977)
9. C.W. Richardson, Water Resour. Res. **17**(1) (1981)
10. D. S. Wilks, Clim. Change **22** (1992)
11. R. W. Katz, M. Parlange, J. Climatol. **11 (**1998)
12. D.S. Wilks, Agric. Forest Meteorol. **96** (1999)
13. H. Wan, X. Zhang, E. M. Barrow, Atmosph.Ocean **43**(1) (2005)
14. P.Todorovic, D.A .Woolhiser, J. Appl. Meteorol. **14 (**1975)
15. T. Haan, M. Allen, Water Resour. Res. **12** (1976)
16. B. Getachew, M. Teshome, J. Degrade. Min. Land Manage. **5**(2) (2018)
17. M. Raheem, W. B. Yahya, K. Obisesan, J. Environ. Stat. **7** (2015)
18. J.B. Álvaro, P.M. Luana, Manage. Environ. Qual. Int. J. **28** (2017)
19. M. Tettey, F.T. Oduro, D. Adedia, Earth Perspectives **4(**6) (2017)
20. C.W. Richardson, D.A. Wright, *WGEN: a model for generating daily weather variables* (U.S. Dept. Agric., Agric. Res. Svc, USA, 1984)
21. G. Schwarz, Ann. Stat. **6** (1978)
22. H. Akaike, IEEE Trans. Automat. Contr. **19** (1974)
23. R. Ahmed, *An investigation of drought risk in Bangladesh during the pre-monsoon season* (Ninth Conference on Applied Climatology, Dallas, Texas, U.S.A., American Meteorological Society, Boston, Massachusetts, USA, 1995)
24. P. Banik, A. Mandal, M. S. Rahman, Discrete Dynamics in Nature and Society **7** (2002)
25. F. Sigrist, H. R. Kunsch, W. A. Stahel, Ann. Appl. Stat. **6**(4) (2012)
26. P. R. Dash, Int. J. Adv. Comput. Math. Sci. **3**(4) (2012)
27. K. Yildirak, Z. Kalaylıoglu, A. Mermer, Environ. Ecol. Stat. **22** (2015)
28. R. D. Stern, R. Coe, J. R. Stat. Soc. A **147**(1) (1984)
29. J. Hansen, S. J. Mason, L. Sun, A. Tall, Exp. Agric. **47(2)** (2011)
30. A.V. Ines, J.W. Hansen, A. W. Robertson, Int. J. Climatol. **31**(2011)
31. L. Kuchar, S. Iwański,. Infr. Ecol. Rur. Area **5** (2011)