Journal of Hydroscience and Hydraulic Engineering
Vol. 12, No. 2, November, 1994, 123-130

CHECKING HOMOGENEITY OF FREQUENCY-DISTRIBUTIONS OF DAILY PRECIPITATION AND THEIR PRACTICAL GROUPING BY MULTINOMIAL DISTRIBUTION MODEL AND AIC

BY

Masato Suzuki

Department of Civil Engineering, Gifu National College of Technology,
Motosu-gun, Shinsei-cho, Gifu, Japan

and

Masashi Nagao

Department of Civil Engineering, Nagoya Institute of Technology, Nagoya, Japan

SYNOPSIS

In an ordinary stochastic analysis, hydrological engineers have usually assumed the stationariness and homogeneity of hydrological quantities. This study gives reconsideration to these assumptions, through checking the homogeneity of shape of frequency-distributions. Checking of homogeneity is done by using a multinomial-distribution model and a AIC criterion. If every distribution is not considered as homogenous, a practical grouping of those distributions is carried out by the same technique. Numerical examinations are illustrated on the data of daily precipitation in Gifu city from 1891 to 1990. It is shown that six-pieces dividing model in increasing order of a maximum daily precipitation is the most suitable grouping.

INTRODUCTION

In a stochastic analysis, hydrological engineers have usually assumed that the stationariness and homogeneity of hydrologic data. In a frequency analysis, such as an estimation of the probability of the extreme variable, the hydrologic data are customarily considered as outcomes from an only one population. However there are some exceptions, for example, the rejection of an abnormal variable. This study gives reconsideration to this assumption through the checking similarity of the shape of discrete distributions of a daily-precipitation. When one consider they are not similar, a practical grouping for those distributions is proposed by the same technique.

CHECKING THE SIMILARITY OF THE SHAPE OF DISTRIBUTIONS BY A MULTINOMIAL DISTRIBUTION MODEL

Stochastic researchers proposed the checking method of the similarity of means of the multinomial distribution model and AIC criterion (1). For a detailed discussion of this technique, see Ref.1. In this paper, we describe the important point of this method only.

Checking the similarity of the marginal distribution (2)

The multinomial distribution represents the distribution of probabilities in Bernoulli trials that have some categories of occurrence. If there are c possible incompatible results of some random event for which the separate probability p_i (i=1,2,....,c), then in n trials, the distribution of k_1 events of the first category, k_2 of the second category, ..., k_c of the cth category is

$$m(k_1,\cdots,k_c \mid p_1,\cdots,p_c) = \frac{n!}{k_1! \cdots k_c!} p_1^{k_1} \cdots p_c^{k_c} \tag{1}$$

And maximum likelihood estimation of p_i is given as follows:

$$p_i = \frac{n_i}{n} \quad (i=0,1,\cdots,c), \quad n = \sum_{i=1}^{c} n_i \tag{2}$$

where c = total number of categories; n_i = number of results of the ith category. The similarity of frequency distributions of hydrologic data are examined by the model of multinomial distribution. To illustrate the procedure, we define the following notation describing the frequency of daily hydrologic data.

n : the total number of hydrologic data (total number of days)
$n(i_1,i_2)$: the number of hydrologic data in i_2th category of i_1th year
$n(i_1)$: the number of hydrologic data in i_1th year
$p(i_2 \mid i_1)$: the probability of occurrence of i_2th category for a given i_1th-year
$\left(\sum_{i_2} p(i_2 \mid i_1) = 1\right)$

By the product of component distributions, the probability to obtain hydrologic frequency distributions data is given as follows:

$$P\left[\{n(i_1,i_2)\} \mid \{p(i_2 \mid i_1)\}\right] = \prod_{i_1=1}^{c_1} \left\{ \frac{n(i)!}{\prod_{i_2=1}^{c_2} n(i_1,i_2)!} \prod_{i_2=1}^{c_2} p(i_2 \mid i_1)^{n(i_1,i_2)} \right\} \tag{3}$$

where c_1 = the number of year; c_2 = the number of class of categories. The logarithmic likelihood of this probability is represented as follows:

$$L[p(i_2 \mid i_1)] = K_4 + \sum_{i_2} \sum_{i_1} n(i_1,i_2) \times \ln p(i_2 \mid i_1)$$

where $$K_4 = \ln \left\{ \prod_{i_1=1}^{c_1} n(i_1)! \Big/ \prod_{i_1=1}^{c_1} \prod_{i_2=1}^{c_2} n(i_1,i_2)! \right\}. \tag{4}$$

In the above equation, $p(i_2 \mid i_1)$ is considered as a parameter to be estimated. In order to show the parameter of the model, in this research, we propose the following 3types of parameter. We call an independent significant parameter as a free parameter.

MODEL(1) : $p(i_2 \mid i_1) = \theta(i_2)$ — number of free parameter : c_2-1
MODEL(M) : $p(i_2 \mid i_1) = \theta(i_2 \mid m_j)$ $(j=1,2,\cdots,M)$ number of free parameter : $M \times (c_2-1)$
MODEL(c_1) : $p(i_2 \mid i_1) = \theta(i_2 \mid i_1)$ — number of free parameter : $c_1 \times (c_2-1)$

$$\theta(i_2 \mid m_j) = \sum_{i_1 \in m_j} n(i_1,i_2) \Big/ \sum_{i_1 \in m_j} n(i_1), \quad m_j\text{:dividing group} \tag{5}$$

In MODEL(1), we assume the probability of i_2th category is common for every year. In MODEL(M), we assume that the distribution of each year is represented by the M kind of distributions. On the other hand, MODEL(c_1) assumes that the distribution of each year is represented by the different distribution respectively. In the case of MODEL(M), the logarithmic likelihood, $L[p(i_2,i_1)]$, is essentially given by Eq.4. Because the term K_4 in Eq.4 is common for all models, we can discuss the likelihood by the second term of Eq.4. So, we call $\sum_{i_2}\sum_{i_1} n(i_1,i_2)\times\ln p(i_2|i_1)$ as a logarithmic likelihood. We use the AIC (Akaike's Information Criterion) as the degree of fitness of the model to the observed data.

$$AIC=-2\times L\,[p(i_2,i_1)]+2\times[\text{number of free parameters}] \tag{6}$$

After the calculation of AIC for every model, the model that has the minimum AIC is adopted as an optimal one. If MODEL(1) is adopted, the distributions are concluded to be similar for all years. On the other hand, if MODEL(M) is adopted, the distributions are concluded not to be similar. In this analysis the corresponding the number of groups and the division domain are simultaneously obtained.

Checking the similarity of the joint distribution (3)

When you discuss the distribution of daily precipitation, the serial continuity of the occurrence might not be disregarded. Then, we consider the checking of similarity on a sequential frequency distribution. This sequential character is usually described by a serial correlation. We use the following notations:

$n(i_1,i_2,i_3)$: the frequency of sequential occurrence on daily precipitation of i_2th and i_3th category in i_1th year. ($i_1=1,2\cdots,c_1,$; $i_2, i_3=1,2,\cdots c_2$)

$p(i_2,i_3\,|\,i_1)$: the conditional probability of occurrence on daily precipitation of i_2th and i_3th category with a given condition of i_1th year, where $\sum_{i_2}\sum_{i_3} p(i_2,i_3\,|\,i_1)=1$

$p(i_1,i_2,i_3)$: the joint probability of occurrence on daily precipitation of i_2th and i_3th category in i_1th year, where $\sum_{i_1}\sum_{i_2}\sum_{i_3} p(i_1,i_2,i_3)=1$

The probability, $P\left[\{n(i_1,i_2,i_3)\}\,|\,\{p(i_1,i_2,i_3)\}\right]$, for any data set, $\{n(i_1,i_2,i_3)\}$, is represented by the following multinomial distribution:

$$P\left[\{n(i_1,i_2,i_3)\}|\{p(i_1,i_2,i_3)\}\right]=\left[n!\Big/\prod_{i_1,i_2,i_3} n(i_1,i_2,i_3)!\right]\prod_{i_1,i_2,i_3} p(i_1,i_2,i_3)^{n(i_1,i_2,i_3)} \tag{7}$$

From Eq.7 , logarithmic likelihood is represented as follows,

$$L[p(i_2,i_3|i_1)]=K_6+\sum_{i_2}\sum_{i_1}\sum_{i_3} n(i_1,i_2,i_3)\times\ln p(i_2,i_3|i_1)$$

where $$K_6=\ln\left\{\frac{\prod_{i_1=1}^{c_1} n(i_1)\,!}{\prod_{i_1=1}^{c_1}\prod_{i_2=1}^{c_2}\prod_{i_3=1}^{c_3} n(i_1,i_2,i_3)\,!}\right\}-\ln n!-\sum_{i_1=1}^{c_1}\ln n(i_1)!+\sum_{i_1=1}^{c_1} n(i_1)\times\ln p(i_1) \tag{8}$$

$i_1=1.....c_1,\ i_2=1.....c_2,\ i_3=1.....c_3$

In these equations, $p(i_2,i_3\,|\,i_1)$ is considered as a parameter. As well as the case of marginal distribution, in our model building, the type of distribution can be expressed by these parameters. The expression of the parameter is depend on the occurrence probability of the continuous daily precipitation, and by the sequential event is considered as independent or dependent each other. A model for a joint distribution can be described as well as the case of marginal distribution. Firstly we assume the parameters of the model is independent of each other. Then joint distribution is given by the product of marginal distribution. The probability of sequential occurrence of i_1 and i_2 is represented as follows,

$$p(i_2,i_3|i_1)=p(i_2|i_1)\times p(i_3|i_1)$$
$$=\theta(i_2|m_j)\times\theta(i_3|m_j) \qquad (9)$$

where $\theta(i_2|m_j)=\sum_{i_1\in m_j}\sum_{i_3=1}^{c_3} n(i_1,i_2,i_3)\Big/\sum_{i_1\in m_j} n(i_1)$,number of free parameter: $M\times(c_2-1)$.

On the contrary, if we assume the parameter of the model is dependent on each other, the probability is rewritten as follows,

$$p(i_2,i_3|i_1)=\theta(i_2,i_3|m_j)\ , \qquad (10)$$
$$\theta(i_2,i_3|m_j)=\sum_{i_1\in m_j} n(i_1,i_2,i_3)\Big/\sum_{i_1\in m_j} n(i_1)$$,number of free parameter: $M\times(c_2\times c_2-1)$

As well as the marginal distribution, because of the constant term in logarithmic likelihood is common in all models, K_6 may be excluded in the numerical calculation. The model which will minimize AIC is adopted as the optimal one. It is clear from Eq.9 and Eq.10, that the number of the free parameters of dependent model is larger than the independent model. So this analysis has a tendency not to select the dependent model.

Example of the checking of the similarity

Now we show the numerical example of the checking of the similarity about marginal distributions by using the AIC criterion. We use the three examples of frequency distributions shown in Table 1. The type A is the frequency distribution of

Table 1 Frequency distributions for example of calculation of AIC

class	A		B		C	
	frequency (relative frequency)					
1	330	(0.904)	321	(0.879)	1650	(0.904)
2	19	(0.052)	27	(0.074)	95	(0.052)
3	5	(0.014)	7	(0.019)	25	(0.014)
4	6	(0.016)	2	(0.005)	30	(0.016)
5	4	(0.011)	1	(0.003)	20	(0.011)
6	0	(0.)	1	(0.003)	0	(0.)
7	1	(0.003)	3	(0.008)	5	(0.003)
8	0	(0.)	1	(0.003)	0	(0.)
9	0	(0.)	2	(0.005)	0	(0.)

the daily precipitation data at Gifu city in 1986, and we use the 15mm class width for the representation of daily precipitation. Type B is the daily precipitation in 1971. And the type C is the artificial distribution generated from type A by multiply the occurrence number of each class by five. Therefore, a relative frequency of type C distribution is the same as type A.

First of all, we check the similarity of frequency distribution A and B. We assume that those two distributions are similar. Then the logarithmic likelihood can be obtained from Eq.4 as follows,

$LL=(330+321)\times ln\{(330+321)/730\}+(330+321)\times ln\{(19+27)/730\}+\ldots+(330+321)\times ln\{(0+2)/730\}=-357.862$

In this case, the number of the class is 9, and the number of a free parameter is 8(=9-1). Finally we obtain AIC≒732 for the MODEL(1). On the other hand, in the case of the model that assumed two distributions are not similar, logarithmic likelihood can be expressed as follows,

LL=330×ln{330/365}+19×ln{19/365}+...+321×ln{321/365}+...+2×ln{2/365}= -351.627
In this case, the number of free parameters is 16, which is two times bigger than that of MODEL(1). Then AIC of MODEL(2) is 735. Comparing these models, MODEL(1) should be adopted because AIC(1) is smaller than AIC(2). Therefore, it is concluded that the shape of distributions are similar, because of the similarity of their parameters.

Secondly, the similarity of the shape of type B and C will be examined. According to the same procedure, in the case we assumed that the shape of distributions are similar, AIC(1)≒2021 and LL(1)≒ -1002. In the case of MODEL(2), where the shape of distributions are assumed to be dissimilar, LL(2)≒-975 and AIC(2)≒1982. In these cases MODEL(2) should be adopted because AIC(2) is smaller than AIC(1), and it is concluded that the shape of type B and C are not similar. So these result mean that the similarity depend on the number of frequency of the target distributions, even though the relative frequency of type A and C are equal. In the comparison of logarithmic likelihood in type A and B, MODEL(2) is larger. Then we can consider that MODEL(2) is a suitable model. However, the number of free parameters in MODEL(2) becomes two times of that in MODEL(1). Because the difference of the number of free parameters is larger than the difference of logarithmic likelihood. So MODEL(1) is adopted according to AIC.

On the other hand, in the comparison of type B and C, the logarithmic likelihood of MODEL(2) is larger than that of MODEL(1). The number of free parameters of MODEL(2) is larger than the MODEL(1) too. But, the difference of logarithmic likelihood is larger than differences of the number of free parameters, because the frequency of distribution is large. As a result, when AIC is adopted as a criterion of the checking, MODEL(2) should be adopted.

In general, when the frequency increase, the absolute value of logarithmic likelihood also increases. The ratio of logarithmic likelihood occupies in AIC increases, but the number of free parameters is constant. As a result, the difference of the shape of the distribution affects the checking of the similarity. Even though there is some difference in the shape of the distributions, when the frequency is few, the difference of logarithmic likelihood becomes small. As a result, the shape of distributions is likely to be concluded similar. Consequently, the similarity of the shape of distribution can be determined by the combination of the difference of shape of distribution and the frequency.

CASE STUDY

Method of analysis

Daily precipitation in Gifu city from 1891 to 1990 (100 years) are used as data to apply the procedures presented. We will check the similarity of the frequency distribution of daily precipitation in each year. The number of the class of frequency distribution is selected as 18, by using the Sturges' formula about all obtained data. The class width is assumed to be 15mm from the maximum and the minimum value. As described previously, frequency distributions of precipitation must be divided into some groups in order to check the similarity of distribution. There are about 20,000 cases for 100 years of data to divide randomly into only two pairs, and the amount of the checking calculation is extra ordinarily huge. Then, for the practical convenience, we use the basic statistics as indices that show the characteristic of frequency distribution in each year.

All the first frequency distributions of each year are ordered according to their index values. And distributions are divided into two groups and three groups according to the order. There is a 99 kinds of trials when you use the dividing into two groups, as follows,

(1*th*) *and* (2-100*th*), (1-2*th*) *and*(3-100*th*), ···, (1-99*th*) *and* (100*th*)

where *ith* means the *ith* order statistic for an index variable. Next there are 4851 ways of dividing 100 years distribution into three groups. The dividing method into eight is carried out by the 2×2×2 dividing. The best division method in the ordered data can be determined by comparing AIC for each division model. Finally, the optimal method of ordering and grouping are obtained by comparing AIC for the best division model in each index.

Results and consideration

In the case of two division, we use **average, variance, coefficient of variance, coefficient of skewness, number of dry days,** and **annual maximum precipitation** as indices. Referring to the result of two division, in the case of three division, we use **average, variance, annual maximum precipitation** as indices. AIC of the most optimal division model for each index is shown in Table 2. In the table, regardless of the

Table 2 AIC of dividing model using each index

	mean	variance	coe. of var.	maximum	dry days	chronological
1-divide	36808.34	*	*	*	*	*
2-divide	36744.78	36742.25	36752.88	36731.93	36807.88	36813.64
3-divide	36728.65	36719.78	36724.37	36677.84	-	-
4-divide	36729.16	36713.78	36724.92	36652.72	36831.09	-
6-divide(3-2)	36759.00	36737.83	36745.25	36656.58	-	-
6-divide(2-3)	36754.33	36733.12	36741.23	36642.67	-	-
8-divide	36787.95	36770.18	36781.85	36654.42	-	-
100-divide	39155.62	*	*	*	*	*

*: the same as left
-: no calculation

index, the "1-divide" means MODEL(1). The "100-divide" means MODEL(100). For example, "6-divide(3-2)" means that data is one divided into three and each group is divided into two again. And, "6-divide(2-3)" means that each group divided into two is divided into three. Minimum AIC model among all dividing models is the 6 dividing model using annual maximum precipitation as an index. Based on this result, it can be concluded that modeling by six kinds of distributions is more appropriate than the modeling that the daily precipitation frequency distribution of each year is similar. Moreover, the AIC of two divided model arranged in a chronological order is larger than that of MODEL(1). Therefore, the monotonous change like a trend may not be admitted for the shape of frequency distribution of daily precipitation.

Fig.1 shows the procedure of division into eight for annual maximum precipitation as an index. At first, all data sample is divided into two parts. The division is

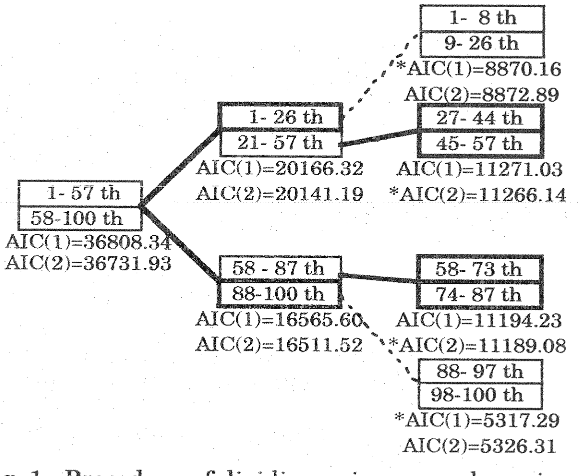

Fig. 1 Procedure of dividing using annual maximum precipitation as an index (eight-divide)

thecombination of 1-57th and 58-100th. The optimal AIC(2) is 36732. When this AIC(2) is compared with AIC(1) (≒36808) of MODEL(1), AIC(2) is smaller. Therefore, we proceed the following division into four and eight At these steps, the method of division is checked by comparing AIC(1) and AIC(2) by the similar manner as the first steps. And, the most suitable divide in these procedures are obtained. We select the smaller AIC in these processes of the division step goes 4 to 8 divide, and the summing results means the AIC of this division model. The order enclosed with the thick line in figure is a result of the final division.

By the way, the optimal division method and AIC obtained according to this procedure correspond to the result shown in Table 2. The relative frequency distribution and frequency distribution of each divided group are shown in Table 3. In Table 3,

Table 3 Relative frequency distribution of most suitable division using annual maximum precipitation as an index

	88-100th		74-87th		58-73th		45-57th		27-44th		1-26th	
1	0.88	(4161)	0.89	(4541)	0.89	(5194)	0.89	(4193)	0.89	(5840)	0.89	(8497)
2	0.07	(312)	0.05	(277)	0.06	(354)	0.06	(305)	0.07	(433)	0.06	(537)
3	0.02	(116)	0.03	(145)	0.03	(147)	0.03	(131)	0.02	(157)	0.03	(256)
4	0.02	(80)	0.01	(61)	0.01	(60)	0.01	(58)	0.01	(69)	0.01	(115)
5	0.01	(31)	0.01	(39)	0.01	(32)	0.00	(23)	0.01	(36)	0.00	(60)
6	0.00	(10)	0.00	(15)	0.00	(20)	0.00	(18)	0.00	(16)	0.00	(34)
7	0.00	(10)	0.00	(12)	0.00	(14)	0.00	(4)	0.00	(22)	0	(0)
8	0.00	(7)	0.00	(1)	0.00	(7)	0.00	(15)	0	(0)	0	(0)
9	0.00	(1)	0.00	(5)	0.00	(18)	0	(0)	0	(0)	0	(0)
10	0	(0)	0.00	(10)	0	(0)	0	(0)	0	(0)	0	(0)
11	0	(0)	0.00	(5)	0	(0)	0	(0)	0	(0)	0	(0)
12	0.00	(6)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
13	0.00	(3)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
14	0.00	(5)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
15	0.00	(2)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
16	0.00	(1)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
17	0.00	(2)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
18	0.00	(2)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)

although the maximum class decreases gradually, but a clear change is not found in a relative frequency of other classes. This means that the maximum value of the class of frequency distribution plays an important role in the checking of similarity of frequency distribution.

We examine the model divided into until eight for joint frequency distributions of daily precipitation for each year. The annual maximum precipitation is useful to make a marginal distribution model. And a coefficient of autocorrelation is essential index of the order. As described previously, there is two kinds of model for joint distribution. That is the model classified to consider the autocorrelation or not. We examine four kinds of model. Those are combination of division and the introducing the autocorrelation or not.

When coefficient of autocorrelation is assumed as an index, three division model is the most suitable. On the other hand, when annual maximum precipitation is assumed as an index, eight division model is suitable. The AIC using annual maximum precipitation as an index is smaller than that using a coefficient of autocorrelation. The dependent model is adopted than the independent one. In the other case, independent models are adopted. When you divide a data into many group, frequency of each distribution decreases. So, under the AIC criterion, the increase of the number of free parameters is larger than the increase of logarithmic likelihood. As a result, the dependent model is not adopted. The result of this division is compared with the one of the division of marginal distribution. The order of the boundary of division is almost similar though there is a difference in a final number of division.

The division by an index of annual maximum precipitation is suitable for both marginal distribution and joint distribution. Then, an analogous application is done to the marginal distribution of not yearly but specific period which annual maximum precipitation occurs. The date when the annual maximum value had occurred is examined for daily precipitation during 100 years. The period which was recorded from beginning day to ending day is assumed to be a rainy period. We select that rainy period as the duration 237 days from the 16 February to 21 October. We use the three kinds of the index, there are **annual maximum precipitation, variance,** and **average**. The optimal model finally obtained is the division into five using annual maximum precipitation as an index. This result is compared with the result of marginal distribution when each year is assumed to be a unit. In the case of rainy period, the group from 45th to 72th is divided into two, i.e. from 45th to 57th and from 58th to 72th. Additionally, another case leads to the same results. Checking similarity of both each year and the rainy periods is influenced by the maximum class in the model.

SUMMARY

In this research, a daily precipitation at a observation point was examined from the viewpoint of the similarity of the shape of frequency distribution. As a result, it has been clarified that it is effective to use the annual maximum precipitation as an index divide both marginal distribution and joint distribution. We adopt the multinomial distribution model that have a large number of parameters. Therefore, the model in consideration of autocorrelation has a tendency not to be adopted for the joint distribution. We plan the practical development of this technique in the checking similarity of frequency distribution for a number of observation points.

REFERENCE

1. Sakamoto,Y., M. Ishiguro and G. Kitagawa: Statistical Theory Information Statistics, Kyoritsu Shuppan Corp., pp.74-77, 1983(in Japanese).
2. Suzuki, M. and M. Nagao, : Checking similarity of the shape of hydrologic frequency distribution, Proc. of the annual conference of the JSCE, II, pp.670-671, 1992(in Japanese)
3. Sakamoto,Y., M. Ishiguro and G. Kitagawa: Statistical Theory Information Statistics, Kyoritsu Shuppan Corp., pp.96-105, 1983(in Japanese)

APPENDIX-NOTATION

The following symbols are used in this paper.

n = total number of hydrologic data (total number of days);
c_1 = number of year;
c_2 = number of class of frequency distribution ;
$n(i_1,i_2)$ = frequency of hydrologic data of every-year(i_1:i_1th-year,i_2:class of frequency distribution);
$n(i_1)$ = number of hydrologic data in i_1th year;
$p(i_2 \mid i_1)$ = probability of occurrence of i_2th category with conditional in i_1th-year;
$n(i_1,i_2,i_3)$ = frequency of sequential occurrence of daily precipitation of i_2 and i_3 category in i_1 year, where $i_1=1,2,...,c_1$, $i_2=1,2,...,c_2$;
$p(i_2,i_3 \mid i_1)$ = probability of occurrence of daily precipitation of i_1 and i_2 category with the condition of i_1 year, where $\sum_{i_2}\sum_{i_3} p\ (i_2,i_3|i_1)=1$ holds;
$p(i_1,i_2,i_3)$ = probability of occurrence of daily precipitation of i_1 and i_2 category in i_1 year, where $\sum_{i_1}\sum_{i_2}\sum_{i_3} p\ (i_1,i_2,i_3)=1$ holds;
AIC = AKAIKE Information Criterion; and
LL = logarithmic likelihood.