SCRIPTA FAC. SCI. NAT. UNIV. PURK. BRUN.
Vol. 14 (1984), No. 6 (Geographia), p. 289—300

COMPARING DIFFERENT METHODS FOR COMPLETING THE MISSING DATA ON PRECIPITATION

R. Brázdil and T. Litschmann

Department of Geography, Faculty of Science, J. E. Purkyně University,
Kotlářská 2, Brno, Czechoslovakia

Received for publication: July 1983

SUMMARY

The paper deals with a comparison of several methods of completing the missing monthly sums of precipitation on the example of Moravian stations. The methods verified are the method of simple completion of the missing data, the method of linear regression of all values, the method of linear regression according to individual months, the quotinet method and the method of multiple linear regression. From the analysis it follows that the most suitable method from all aspects is the method of multiple linear regression.

1. INTRODUCTION

The beginnings of an extensive network of rainfall measuring stations on the territory of Bohemia and Moravia date back deep into last century, when was organized the basic network of stations. Some of them have kept their observation activity up to the present. But in many cases they have not avoided interferences which for different reasons have resulted in affecting the homogenity of the observation series (moving the stations, changes in observers, war events, etc.). Therefore stations with complete observation series in this country are relatively very few. This causes trouble in climatic studies in which it is necessary to start from long observation series (such as in studying the variations of the climate). In such cases it is necessary to homogenize the observation series by completing the missing data.

In connection with the calculation of territorial average precipitation the problem cropped up how to homogenize the precipitation series of Moravian stations with longer observation series. The objective of this contribution is to evaluate different methods applicable in completing the missing monthly sums of precipitation.

2. THE MATERIAL USED AND THE METHODS OF PROCESSION

Completing precipitation data, diurnal, monthly or annual sums, is a very complicated matter with respect to a great time and territorial variability of atmospheric precipitation. This fact can only partly be eliminated by a sufficiently dense network of rainfall measuring stations.

In Czechoslovak climatological literature these problems have been paid little attention. In completing the missing values usually the data of the nearest station are used or data following

from the maps of isohyetes. In the textbook by M. Nosek (1972) only the method of quotients (see below) is mentioned together with a note saying that for completion also regressive dependences can be used.

In this paper the following methods used for completing the missing monthly sums of precipitation have been used:

— the method of simple completion (SC)
— the quotient method (QM)
— the method of linear regression of all values (LR)
— the method of linear regression in individual months (LM)
— the method of multiple linear regression (MR).

The method of simple completion (SC) is the simplest method consisting in completing the missing data by those from the nearest station, corresponding approximately by the height above sea level and exposition parameters.

The method of quotients (QM) is based on the determination of the quotient of long-term sum of precipitation at the station to be completed and the analogon-station in the respective month. The corresponding monthly sum at the analogon-station is multiplied by the quotient obtained.

The method of linear regression (LR) of all values uses the relation

$$y = a + bx \tag{1}$$

for completing the missing data. In the relation y represents the completed sum of precipitation, x is the precipitation sum at the analogon-station; a, b are parameters determined by the method of least squares from all ordered pairs of precipitation sums by means of relations quoted by J. Anděl (1978).

The method of linear regression according to individual months (LM) starts from the same theoretical basis as the preceding one, the difference consisting in the fact that it states 12 regression equations (one for each month of the year), while in the method of linear regression of all values only one equation was stated for all months.

The method of multiple regression (MR) is employed for completing the missing data on the basis of data from surrounding stations according to the relation

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n, \tag{2}$$

Table 1. A list of stations used (the first of the stations listed in the column stations-analogons was always used for the SC, QM, LR, and LM methods, all stations for the MR method; list of stations — see Tab. 1 in R. Brázdil, 1984)

Completed station	Stations-analogons
Hodonín	Dubňany-Jarohněvice, Prušánky, Břeclav
Horní Bečva	Hutisko-Solanec, Rožnov p. Radhoštěm, Valašská Bystřice
Horní Bečva	Rožnov p. R., Hutisko-Solanec, Valašská Bystřice
Hostýn	Bystřice p. Hostýnem, Holešov, Rajnochovice
Kladeruby n. Osl.	Hrotovice, Náměšť nad Oslavou, Ivančice
Kladeruby n. Osl.	Náměšť nad Oslavou, Hrotovice, Ivančice
Kněževes	Velké Meziříčí, Bohdalov, Skřinářov
Lipník n. Bečvou	Hranice, Přerov
Lysá hora-Mountain	Čeladná-Podolánky, Ostravice, Velké Karlovice-Javorníky
Lysá hora-Mountain	Ostravice, Čeladná-Podolánky, Velké Karlovice-Javorníky
Pavlovice u Přerova	Lipník n. Bečvou, Přerov, Hranice
Pavlovice u Přerova	Přerov, Lipník n. Bečvou, Hranice
Prušánky	Břeclav, Hodonín
Prušánky	Dubňany-Jarohněvice, Břeclav
Rajnochovice	Hošťálková, Kelč, Valašské Meziříčí
Rajnochovice	Kelč, Hošťálková, Valašské Meziříčí
Skřinářov	Bohdalov, Velké Meziříčí, Kněževes
Valtice	Břeclav, Mikulov, Prušánky
Valtice	Mikulov, Břeclav, Prušánky
Zlaté Hory	Zlaté Hory-Rejvíz, Jindřichov, Mikulovice

where y is the completed precipitation sum, x_i the precipitation sum at the i-th analogon. The estimates of parameters a_i are again made by the method of least squares from all ordered n-tuples of precipitation sums. In detail this method is described in the paper Proceedings of the Roving Seminar on the Use of Computers in Hydrology and Water Resources Planning (1980).

A practical application of the above methods, with the exception of SC, depends on the possibility of utilizing the means of modern computation technology which enable the procession of extensive data sets and thus also the employment of such statistical methods as were formerly impossible to perform because of their exacting and time-consuming character. Therefore for all the above methods programs were elaborated in the language FORTRAN IV and they were tested in the computer EC 1033 of the Institute of Computation, J. E. Purkyně University (the programs are available at the Department of Geography).

For verifying the suitability of the methods used stations were chosen from regions representing orographically and from the point of view of precipitation different parts of Moravia. They are listed in Tab. 1. For comparing actually measured values with those completed, decades were chosen in which the stations to be completed as well as the analogons had complete series of observations. The remaining measured data from the period of 1881—1980 were used for determining the regression dependencies. The decades chosen for the individual stations were different from the point of view of time selection, which, in view of the variability of precipitation, makes the comparison of the methods employed more objective.

3. RESULTS OF PROCESSION

To judge the applicability of the individual methods for completing the monthly sums of precipitation attention was paid to their suitability for completing the individual data as well as for determining a long-term (ten year) average.

Differences between the calculated sums of precipitation and those actually measured (in the following text only differences) for all stations and all methods

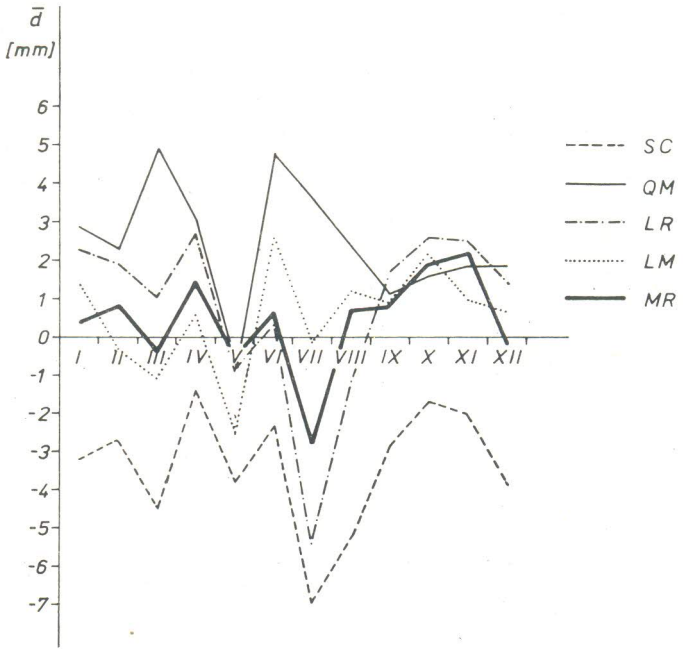

Fig. 1. Annual variation of average values of differences according to individual methods

Table 2. Average values of differences $\bar{d}$ and their probable errors c in mm (probable error $c = 0.6745 \cdot s$, where s is the standard deviation)

Method		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
SC	$\bar{d}$	−3.2	−2.7	−4.5	−1.4	−3.8	−2.3	−6.9	−5.2	−2.8	−1.7	−2.0	−3.8
	c	7.1	6.1	7.7	4.3	7.4	9.4	17.6	10.0	4.6	5.4	5.4	9.1
QM	$\bar{d}$	2.9	2.3	4.9	3.0	−0.7	4.8	3.6	2.4	1.1	1.6	1.9	1.9
	c	4.0	3.2	3.8	5.9	8.8	8.1	10.8	8.6	5.1	5.1	6.1	4.2
LR	$\bar{d}$	2.3	1.9	1.0	2.7	−0.8	0.3	−5.4	−1.2	1.7	2.6	2.5	1.4
	c	3.8	3.6	3.6	4.4	7.1	6.9	11.3	8.2	5.1	3.6	4.2	5.1
LM	$\bar{d}$	1.4	−0.3	−1.1	0.6	−2.5	2.6	−0.1	1.2	0.9	2.2	1.0	0.7
	c	3.5	3.6	3.8	4.0	6.7	5.0	9.5	6.2	4.0	3.2	4.2	3.6
MR	$\bar{d}$	0.4	0.7	−0.4	1.4	−0.6	0.6	−2.8	0.7	0.8	1.9	2.2	−0.2
	c	2.9	2.3	3.6	4.0	4.9	4.0	9.1	4.1	2.6	1.8	2.5	3.3

Table 3. The average of differences expressed as percentage of the measured ten year average ($\bar{d}$) and their probable errors c (as percentage)

Method		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
SC	$\bar{d}$	−2.4	−3.8	−5.0	−1.7	−2.7	−0.4	−0.9	−3.5	−3.9	−1.8	−2.6	−3.1
	c	13.2	10.3	12.8	6.8	7.4	7.5	10.8	8.8	8.9	10.3	8.3	14.0
QM	$\bar{d}$	8.1	4.3	3.9	5.9	1.1	3.8	5.6	2.3	2.4	2.2	3.1	4.6
	c	10.1	6.0	8.8	11.6	10.5	8.6	15.6	10.2	11.2	12.0	10.7	9.4
LR	$\bar{d}$	7.5	5.0	6.4	3.9	0.1	−0.6	−2.0	−0.6	4.2	5.6	3.8	4.4
	c	9.1	6.4	7.4	7.8	8.6	6.7	7.7	8.6	11.1	8.2	7.1	8.3
LM	$\bar{d}$	3.3	−2.3	−1.3	1.4	−1.3	3.3	3.0	2.7	1.9	4.4	2.0	−0.3
	c	7.0	7.0	7.8	9.7	7.7	6.5	8.1	7.7	8.5	7.0	7.7	8.0
MR	$\bar{d}$	0.5	0.5	0.3	1.6	−0.8	−0.2	−1.0	0.2	0.7	3.7	2.8	0.1
	c	7.5	5.6	7.4	7.3	6.6	6.0	5.9	5.5	5.4	4.3	4.1	5.8

Table 4. Average standard deviations of differences s and their probable errors c (in mm)

Method		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
SC	s	10.8	10.7	9.5	11.2	15.8	21.7	30.0	21.7	13.7	10.0	11.1	11.7
	c	4.4	5.4	4.0	3.8	4.3	6.4	17.2	7.4	8.6	3.7	5.3	6.3
QM	s	12.0	10.2	9.8	12.4	16.8	24.5	30.4	21.3	14.2	10.9	12.0	11.7
	c	5.6	4.8	3.6	4.4	4.5	8.7	19.2	6.1	8.3	3.8	5.2	5.7
LR	s	10.4	10.7	9.9	11.4	15.8	21.3	27.9	22.0	15.3	10.2	12.0	11.3
	c	5.3	5.4	3.5	3.5	5.0	6.6	17.4	8.1	7.3	3.4	4.7	6.5
LM	s	11.4	10.1	9.7	12.2	16.8	19.8	28.1	21.4	14.1	9.3	11.0	10.6
	c	4.3	4.4	3.6	3.6	4.6	6.1	18.0	6.8	7.9	3.2	4.8	5.5
MR	s	9.1	9.5	7.5	9.1	14.0	20.0	23.7	19.4	12.1	9.1	8.6	9.6
	c	3.5	5.9	2.6	3.4	4.7	7.5	12.2	6.1	5.9	5.2	4.0	6.0

for the decade selected. From sets of differences thus originating the following statistical characteristics were determined:

— the average difference (in mm) — Tab. 2, Fig. 1

— the average of differences expressed as percentage of the measured ten year average — Tab. 3, Fig. 2

— standard deviation — Tab. 4, Fig. 3

— the average of absolute values of differences — Tab. 5, Fig. 4

— the value of the testing criterion of the t-test for the paired values — Tab. 6, Fig. 5.

From Tabs. 2—6 and Figs. 1—5 it is evident that the best results are those obtained by the MR method. The other methods are approximately equivalent. This is best evident from Fig. 4, where from the point of view of absolute values

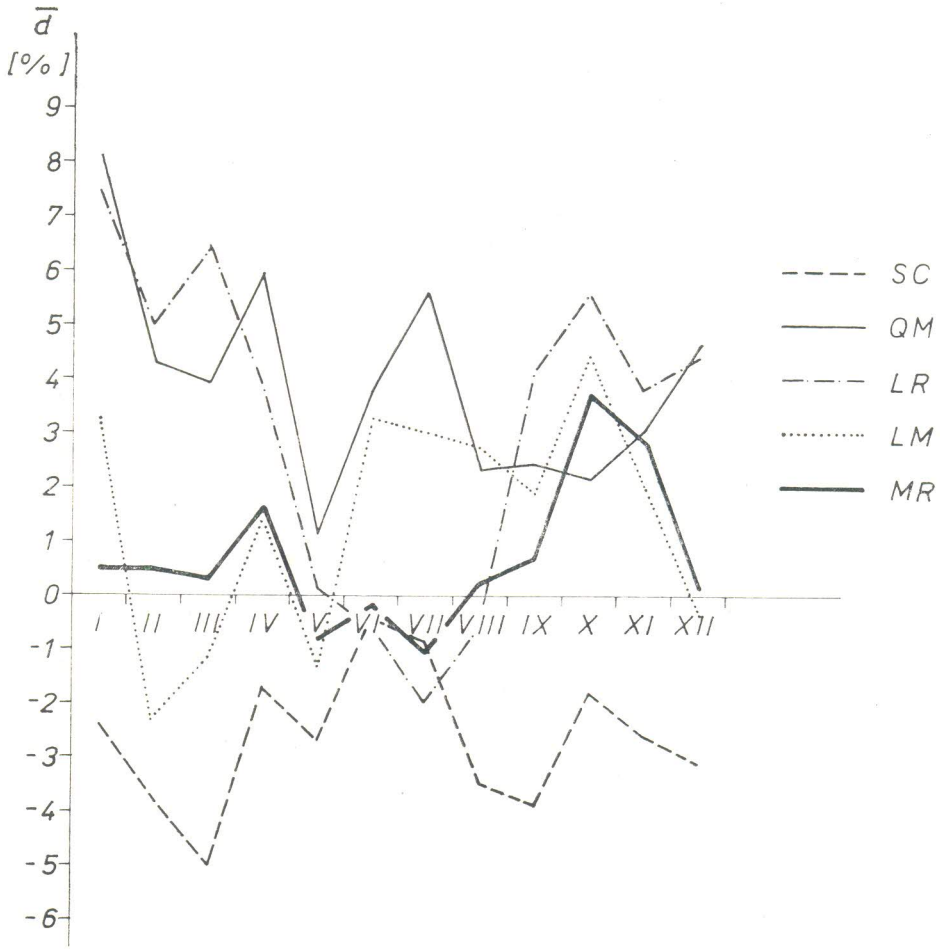

Fig. 2. Annual variation of averages of differences expressed as percentage of the measured ten year average according to individual methods

of differences the MR method yields substantially lower differences than the other methods. As can be expected, from the figure the annual variation of absolute values of differences can well be observed, with maximum values in the summer half of the year (maximum in July), and the lowest values in the cold half of the year (minimum in October). These differences in the annual variation are in connection with a different character of the precipitation activity in the two parts of the year. Also in the relative representation of differences (Fig. 2) the suitability

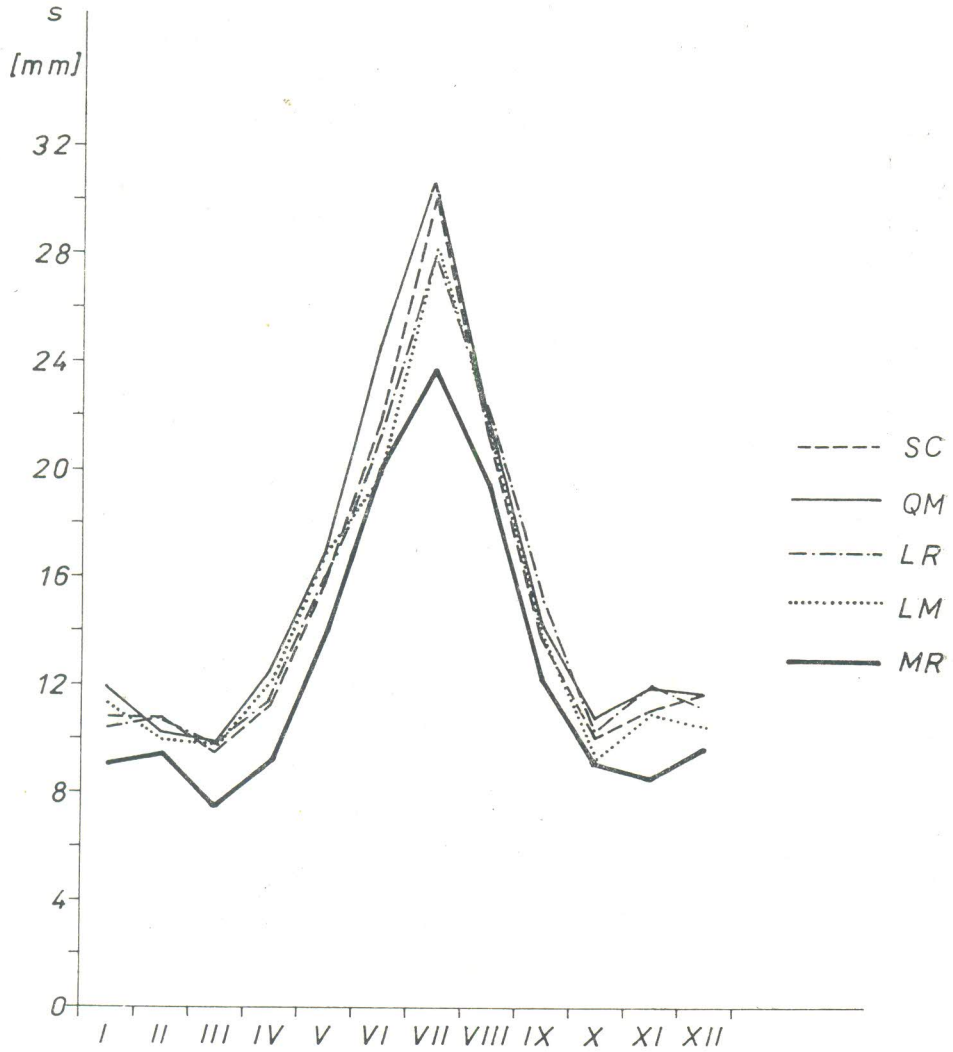

Fig. 3. Annual variation of the average standard deviations of differences according to individual methods

of the MR method is evident, in which the values mostly vary within the interval from −1 to +1 %, only in October and November acquiring about 3 %.

For comparing the individual methods graphically the order of agreement was used which was obtained from ten year averages of absolute values of differences for all stations used and for individual months. As an example differences for the month of July the difference can be used in the pair of stations Prušánky (the station to be completed) — Břeclav (analogon) (Tab. 7).

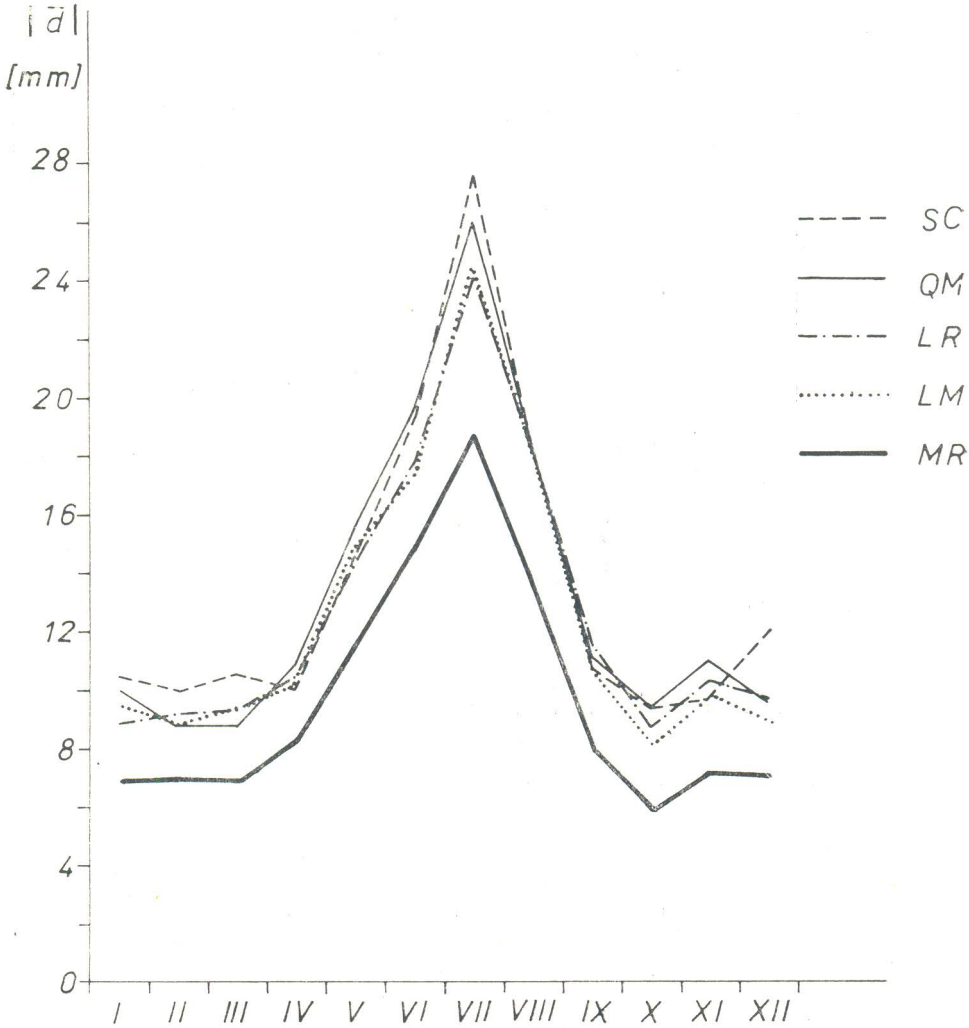

Fig. 4. Annual variation of averages of absolute values of differences according to individual methods

Table 5. Averages of absolute values of differences $|\bar{d}|$ and their probable errors c (in mm)

Method		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
SC	$\|\bar{d}\|$	10.5	10.0	10.6	10.1	14.6	19.3	27.6	18.2	10.8	9.3	9.8	12.0
	c	5.6	5.2	5.8	3.4	5.2	6.0	16.6	7.2	4.7	3.8	4.4	7.6
QM	$\|\bar{d}\|$	10.0	8.9	8.9	11.9	15.6	19.5	25.9	18.3	11.1	9.4	10.9	9.5
	c	4.4	4.3	3.2	4.2	5.0	5.3	14.2	6.2	4.5	3.8	5.3	4.5
LR	$\|\bar{d}\|$	8.9	9.2	9.3	10.4	14.3	17.7	24.1	18.4	11.5	8.7	10.3	9.7
	c	3.1	4.4	3.5	3.8	4.2	4.7	13.4	6.8	5.5	3.0	4.5	4.8
LM	$\|\bar{d}\|$	9.5	8.9	9.3	10.3	14.9	17.4	24.4	18.1	10.7	8.1	9.8	8.9
	c	4.2	4.0	3.4	3.0	4.0	4.3	13.0	5.7	4.6	2.8	4.0	4.0
MR	$\|\bar{d}\|$	6.9	7.0	6.9	8.2	11.8	14.9	18.7	13.6	8.0	5.8	7.1	7.0
	c	2.2	3.7	2.5	3.0	3.5	3.2	10.1	2.9	3.5	2.0	3.0	3.9

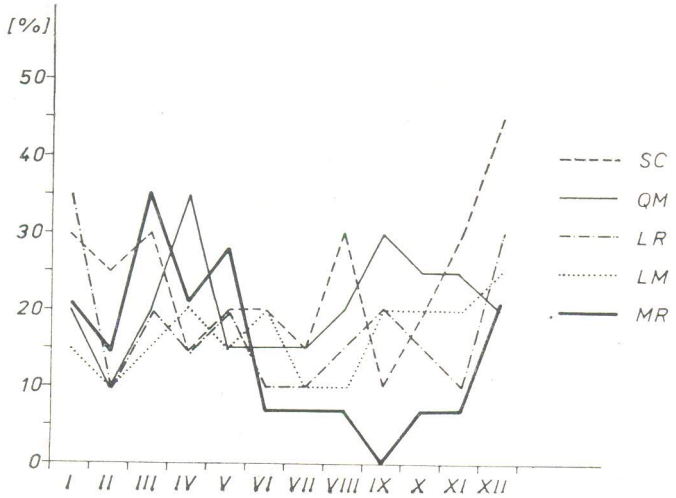

Fig. 5. Annual variation of probabilities of the occurrence of significant difference according to individual methods

Altogether 1,200 orders of agreement were obtained. The relative frequences of the occurrence of the individual orders of agreement are given in Tab. 8, from which it clearly follows that the values calculated by means of the MR method are closest to the values actually measured.

In evaluating the closeness of the individual methods according to the first order of agreement only a rough initial orientation can be obtained, since higher orders of agreement are not considered, which assert themselves in determining the order of methods showing less agreement in theoretical and empirical values. Therefore index *I* was introduced considering also relative frequencies of higher orders of agreement in such a way that the frequencies of the higher orders of

agreement reduce the suitability of the given method for completing the missing data. The index I is determined from the relation:

$$I = F_1 + 2F_2 + \ldots + nF_n, \tag{3}$$

where $n = 1, 2, .., 5$ and F_n is relative frequency in percentage of the n-th order.

The calculated values of I were, for the sake of better orientation, converted from the relation 100 (the method has only the first order) — 500 (the method giving only the 5th order) to the relation 0 (the most suitable method) — 100 (the worst method) by means of the relation $I_k = \frac{I - 100}{4}$. The values of index I_k are given in Table 9 from which again the best closeness of the MR method to the values measured can be observed. From this view the worst results are those obtai-

Table 6. The probability of occurrence of the significant difference (percentage) of calculated and measured sums of precipitation at the level of significance 0.05

Method	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Mean
SC	30	25	30	15	20	20	15	30	10	20	30	45	24
QM	20	10	20	35	15	15	15	20	30	25	25	20	21
LR	35	10	20	15	20	10	10	15	20	15	10	30	18
LM	15	10	15	20	15	20	10	10	20	20	20	25	17
MR	21	14	35	21	28	7	7	7	0	7	7	21	15

Table 7. The order of agreement for differences between the calculated and the measured sums of precipitation of the month of July for stations Prušánky (completed) — Břeclav (analogon)

	Methods				
	SC	QM	LR	LM	MR
Difference (in mm)	15.6	14.5	14.2	14.4	8.4
Order of agreement	5.	4.	2.	3.	1.

Table 8. Relative frequency of occurrence of the order of agreement (percentage) for the individual methods employed

Order of agreement	Methods				
	SC	QM	LR	LM	MR
1.	13.3	9.6	9.2	10.8	63.8
2.	21.7	23.3	23.8	25.0	14.6
3.	25.4	15.8	28.3	27.1	8.8
4.	19.6	28.8	25.8	19.2	8.3
5.	20.0	22.5	12.9	17.9	4.5

Table 9. Values of index I_k (%)

	Methods				
	SC	QM	LR	LM	MR
I_k	52.9	57.8	52.4	52.1	18.9

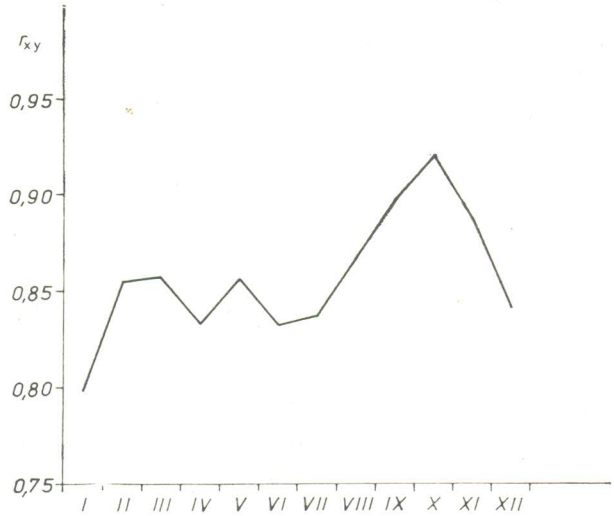

Fig. 6. Annual variation of correlation coefficients r_{xy} of the LM method

ned by means of the QM method, the remaining three methods being practically equivalent with each other.

The analysis performed can be completed by the values of the correlation coefficients obtained in determining linear regression dependences according to the measured data of the stations to be completed and the corresponding analogons (MR, LM, LR methods). The mean correlation coefficients for these methods were as follows: MR 0.927, LM 0.858, LR 0.886, Fig. 6 gives an idea of the annual variation of the correlation coefficient of the LM method. The highest value is reached in October, which is in good agreement with the minimum of averages of absolute values of differences (see Fig. 4). Unexpectedly low values of correlation coefficients in December and, above all, in January are evidently connected with problems of measuring solid precipitation.

4. CONCLUSIONS

The conclusions of the procession can be summarized into the following points:
a) From the method used for completing the monthly sums of precipitation the method of multiple linear regression (MR) appears to be the most suitable. This follows from its very essence, since several analogons are taken for one station to

be completed, thus increasing the probability of recording locally limited shower precipitations. The selection of stations-analogons is recommended on the basis of the size of correlation coefficients between the station to be completed and the stations in its surroundings.

b) From the procession it further follows that the other methods employed are practically equivalent, even though in a particular case some of them can yield a better result than the MR method.

c) The often used method of simple completion (SC) gives relatively good results only in completing a small number of data missing. In a large number of completed data there can be an undesirable distortion of the long-term average, particularly in stations with different heights above sea level in the mountains.

d) The highest correlation coefficients between the monthly sums of precipitation at stations completed and the analogons are obtained in the MR method (0.927), whereas the LR and LM methods yield substantially lower values.

REFERENCES

Anděl J. (1978): Matematická statistika. SNTL, Praha, 352 p.

Brázdil R. (1984): Maximum monthly sums of precipitation in Moravia. Scripta Fac. Sci. Nat. Univ. Purk. Brun., Vol. 14, No. 6 (Geographia), p. 259—284.

Kašpárek L. (1982): Analýza korelačních vztahů mezi průtokovými a srážkovými řadami. In: Práce a studie, seš. 2, ČHMÚ Praha, 38 p.

Nosek M. (1972): Metody v klimatologii. Academia, Praha, 434 p.

Proceedings of the Roving Seminar on the Use of Computers in Hydrology and Water Resources Planning. Water resource series No 52, United Nations, New York, 1980, 356 p.