Reconstruction, prediction and simulation of multiple monthly stream-flow series

L. TORELLI

Received on April 2nd, 1976

SUMMARY. — The logarithms of monthly stream-flows are usually found to have a Normal distribution. Stream-flow series are auto-correlated up to a given time lag s. Moreover stream-flow series of the same region are cross correlated.

Consequently the vector

$$\mathbf{Y}(t^*) = (Y_1(t^*-s),\ Y_1(t^*-s+1),\ \ldots,\ Y_1(t^*+s),\ Y_2(t^*-s),\ Y_2(t^*-s+1),\ \ldots,\ Y_2(t^*+s),\ \ldots,\ Y_n(t^*-s),\ Y_n(t^*-s+1),\ \ldots,\ Y_n(t^*+s))'$$

of the values of the logarithms of the series at stations 1, 2, ..., n from time t^*-s to time t^*+s is a Normal vector containing the maximum of information on any missing data at time t^*,

$$Y_{i1}(t^*),\ Y_{i2}(t^*),\ \ldots,\ Y_{ik}(t^*),\ i1,\ i2,\ \ldots,\ ik \leqslant n$$

If a set of simultaneous observations of the series is available the covariance matrix and the expectation of $\mathbf{Y}(t^*)$ can be estimated. One can thus reconstruct the missing data by the conditional expectation of $Y_{i1}(t^*)$, $Y_{i2}(t^*)$, ..., $Y_{ik}(t^*)$, given the other observed values of vector $\mathbf{Y}(t^*)$. The theory provides also the covariance matrix of $Y_{i1}(t^*)$, $Y_{i2}(t^*)$, ..., $Y_{ik}(t^*)$ and thus the variances of the reconstruction.

Prediction and simulation can be accomplished by a similar technique.

The method has been used to reconstruct the missing data of the streams of the Emilia Romagna Region of Italy and to produce synthetic multivariate series there from.

(*) IDROTECNECO – Roma.

RIASSUNTO. — I logaritmi delle portate mensili dei corsi d'acqua hanno una distribuzione normale. I valori delle portate sono auto-correlati fino ad un certo ritardo s. Inoltre i valori di portata a sezioni vicine tra loro sono anch'essi intercorrelati.

Conseguentemente il vettore

$$\mathbf{Y}(t^*) = (Y_1(t^*-s), Y_1(t^*-s+1), \ldots, Y_1(t^*+s), Y_2(t^*-s), Y_2(t^*-s+1), \ldots, Y_2(t^*+s), \ldots, Y_n(t^*-s), Y_n(t^*-s+1), \ldots, Y_n(t^*+s))'$$

dei valori dei logaritmi dei valori di portata alle stazioni di misura 1,2, ..., n dal tempo (t^*-s) al tempo (t^*+s) è un vettore normale contenente tutta l'informazione riguardo a eventuali dati mancanti al tempo t^*,

$$Y_{i1}(t^*), Y_{i2}(t^*), \ldots, Y_{ik}(t^*), i1, i2, \ldots, ik \leqslant n$$

Se si dispone di un insieme di osservazioni contemporanee delle serie, si possono stimare la matrice della covarianza e la media di $\mathbf{Y}(t^*)$. Si possono quindi ricostruire, attraverso i metodi dell'Analisi Multivariata i dati mancanti, usando le medie di $Y_{i1}(t^*)$, $Y_{i2}(t^*)$, ..., $Y_{ik}(t^*)$ condizionate dagli altri valori osservati di $\mathbf{Y}(t^*)$.

Tali metodi forniscono anche la matrice della covarianza di $Y_{i1}(t^*)$, $Y_{i2}(t^*)$, ..., $Y_{ik}(t^*)$ e conseguentemente le varianze della ricostruzione.

La previsione e la simulazione possono ottenersi con una tecnica analoga.

Questo metodo è stato usato per ricostruire i dati mancanti di portata mensile alle stazioni di misura sui corsi d'acqua dell'Emilia Romagna e per produrre serie sintetiche (simulate), multivariate alle stesse stazioni.

1. INTRODUCTION

The present paper addresses itself to the following problems:

1 – reconstruction of missing data of monthly stream-flows at measuring stations;

2 – prediction of monthly stream-flows at several stations on the basis of past observations at the same stations;

3 – simulation of simultaneous monthly stream-flows at several stations

All the above problems are solved by the same algorithm based on certain results of Normal Multivariate Analysis. Section 2 recalls the mathematical theory used. Section 3 formulates the problems in terms of the theory. Section 4 discusses the estimation of the parameters required in the theory. Finally section 5 discusses the results obtained in a case study concerning the streams of the Emilia-Romagna Region of Italy.

2. The theory

One makes use of the following theory: Given a Normal random vector $\mathbf{Y}$ with expectation m and covariance matrix Σ if one partitions $\mathbf{Y}$ as follows:

$$\mathbf{Y} = (Y'_1, Y'_2)' \qquad [1]$$

such partition induces the following partitions in m and Σ

$$m = (m'_1, m'_2)' = (E Y'_1, E Y'_2)'$$

$$\Sigma = \begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} \qquad [2]$$

where Σ_{11} and Σ_{22} are the covariance matrices of Y_1 and Y_2 and $\Sigma_{12} = \Sigma'_{21}$ is the cross covariance matrix of the vectors Y_1 and Y_2. ($'$ means transposed).

The following result holds (Anderson 1958): the conditional distribution of Y_1 given $Y_2 = y_2$, is Normal with the following parameters:

$$E(Y_1 | Y_2 = y_2) = m_1 + \Sigma_{12} \Sigma_{22}^{-1} y_2$$

$$CM(Y_1 | Y_2 = y_2) = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \qquad [3]$$

Where CM means covariance matrix.

$\hat{\mathbf{Y}}_1 = E(Y_1 | Y_2 = y_2)$ is an optimal predictor for Y_1 in the usual sense. In fact if $Y_1(i)$, $\hat{Y}_1(i)$ are the i^{th} elements of Y_1, $\hat{Y}_1$ respectively, one gets

$$E(Y_1(i) - \hat{Y}_1(i)) = 0$$

$$E(Y_1(i) - \hat{Y}_1(i))^2 = \min_{CGR} E(Y_1(i) - c)^2 \qquad [4]$$

It is also clear that in Bayesian terms the above result completely specifies the distribution of the states of nature so that an optimal predictor, which is optimal with respect to a general loss function, may be determined.

3. Reduction of the problem to the theory

Assume that one deals with the series $\{X_i(t)\}$ $i = 1, 2, \ldots, n$ of monthly stream-flows at stations 1, 2, ..., n. The Kolmogorov-Smirnov tests accept the Normality of $\{Y_i(t)\} = \{\log X_i(t)\}$.

The series $\{Y_i(t)\}$ are both auto-correlated and cross correlated up to a given maximum time lag, s. Let us consider the Normal random vector:

$$\begin{aligned}\mathbf{Y}(t^*) = (&Y_1(t^*-s),\ Y_1(t^*-s+1),\ \ldots,\ Y_1(t^*+s)\\ &Y_2(t^*-s),\ Y_2(t^*-s+1),\ \ldots,\ Y_2(t^*+s),\ \ldots,\\ &Y_n(t^*-s),\ Y_n(t^*-s+1),\ \ldots,\ Y_n(t^*+s))'\end{aligned} \qquad [5]$$

Apparently the vector $\mathbf{Y}(t^*)$ contains all the information stored in the series $\{Y_i(t)\}$, $i=1, 2, \ldots, n$, about missing data at time t^*. Thus if the parameters, i.e. the expectation and the covariance matrix of $\mathbf{Y}(t^*)$, are known, an optimal reconstruction of the missing data can be obtained, according to the theory of section 2.

The estimation of the parameters is discussed in section 4. Let us note, now, that the same theory can be used to predict at time t^* the values of the multiple series at time t^*+1. In this case vector $Y(t^*)$ will be:

$$\begin{aligned}\mathbf{Y}(t^*) = (&Y_1(t^*-s),\ Y_1(t^*-s+1),\ \ldots,\ Y_1(t^*+1),\\ &Y_2(t^*-s),\ Y_2(t^*-s+1),\ \ldots,\ Y_2(t^*+1),\ \ldots,\\ &Y_n(t^*-s),\ Y_n(t^*-s+1),\ \ldots,\ Y_n(t^*+1))\end{aligned} \qquad [6]$$

and one will seek the conditional expectations of

$$Y_1(t^*+1),\ Y_2(t^*+1),\ \ldots,\ Y_n(t^*+1)$$

As to simulation, this is obtained by making, at any time step, a prediction on the basis of values already simulated and adding to the predicted values a vector ε of randomly generated residuals with mean 0 and a covariance matrix specified by formula [3]. Using a random number generator one obtains a vector γ with the identity as covariance matrix. An ε having a desired covariance matrix $\mathbf{B}$ is obtained as follows: $\varepsilon = \mathbf{P}\gamma$ where $\mathbf{P}$ is such that $\mathbf{PP}' = \mathbf{B}$.

4. The estimation of the parameters

The parameters one needs to know are

$$\{E\ Y_i(t),\ i = 1,2,\ \ldots,\ n,\ t \in I\}$$

and $\{\text{cov}\ (Y_i(t_1),\ Y_j(t_2));\ i \text{ and } j = 1, 2, \ldots, n;\ t_1 \text{ and } t_2 \in I\}$. If a set of simultaneous observations at the measuring stations exists, these parameters can be estimated on the basis of certain assumptions concerning the expectation and covariance structure of the series.

The expectation of each series can be assumed to have annual periodicity so that the parameters are optimally estimated by the 12 monthly sample means.

As to the covariance structure, in general,

$$\text{cov } (Y_i(t),\ Y_j(t+r)) = \gamma\ (i,\ j,\ t,\ r) \qquad [7]$$

An assumption of stationarity:

$$\text{cov } (Y_i(t),\ Y_j(t+r)) = \gamma(i,\ j,\ r) \qquad [8]$$

is not realistic since it has been shown (Torelli and Chow 1972, Torelli 1973) that it fails when $i=j$ i.e. in the univariate case. In this case Torelli and Chow have shown that the stationarity hypothesis holds for the standardized series

$$Z_i(t) = (Y_i(t) - E\ Y_i(t)) \ /\ \text{var}^{\frac{1}{2}}\ (Y_i(t)) \qquad [9]$$

where $E\ Y_i(t)$ and var $(Y_i(t))$ have annual periodicity and are estimated by the sample means and variances (*).

In other terms the above transformation eliminates seasonal effects in the covariance structure of the univariate series. It has been assumed by this writer that the above transformation eliminates seasonal effects also in the cross covariances of the multivariate series as well, so that:

$$\text{cov } (Z_i\ (t),\ Z_j\ (t+r)) = \gamma\ (i,\ j,\ r) \qquad [10]$$

It will be noted that in general

$$\gamma\ (i,\ j,\ r) \neq \gamma\ (i,\ j,\ -\ r) \qquad [11]$$

and that

$$\gamma\ (i,\ j,\ r) = \gamma\ (j,\ i,\ -r) \qquad [12]$$

Given the advantages of the standardization, this transformation has been adopted. The $\gamma\ (i,\ j,\ r)$ are then consistently estimated by the proper sample covariances. It is then possible to reconstruct, making use of the results of section 2, the value of $Z_i\ (t^*)$. The reconstructed values of $Z_i(t^*)$ are easily transformed in the reconstructed values of $Y_i(t^*)$.

(*) It must be noted the same results obtained by Torelli and Chow on the Sangamon R. at Monticello, Ill., have been found by Torelli using the same tests on a number of streams in Italy. (See for instance Torelli, in press).

5. A case study: Emilia Romagna

The method described in the previous sections has been tested in a study of the streams of the Emilia Romagna Region of Italy. In fact, this study has been carried out in the framework of Emilia-Romagna Regional Water Plan. It is based on the stream-flow data measured at the gauging stations of the Servizio Idrografico Italiano. These stations are 83 in number. The period of measurement at the various stations vary from 1 year to 48 years. By the method presented in this paper it is possible to obtain for all the sections a complete record from 1921 up to date. Alternatively simulated (synthetic) multiple series may be produced, which display the same statistical properties, expectation and covariance structure, of the measured series. The reconstructed series or the simulated series constitute the documentation on which water resource planning and hydraulic design may be based.

Reported here are the reconstructions of the data at stations 1, 2,4 (see Fig. 1), based on observations at station 3, in various periods.

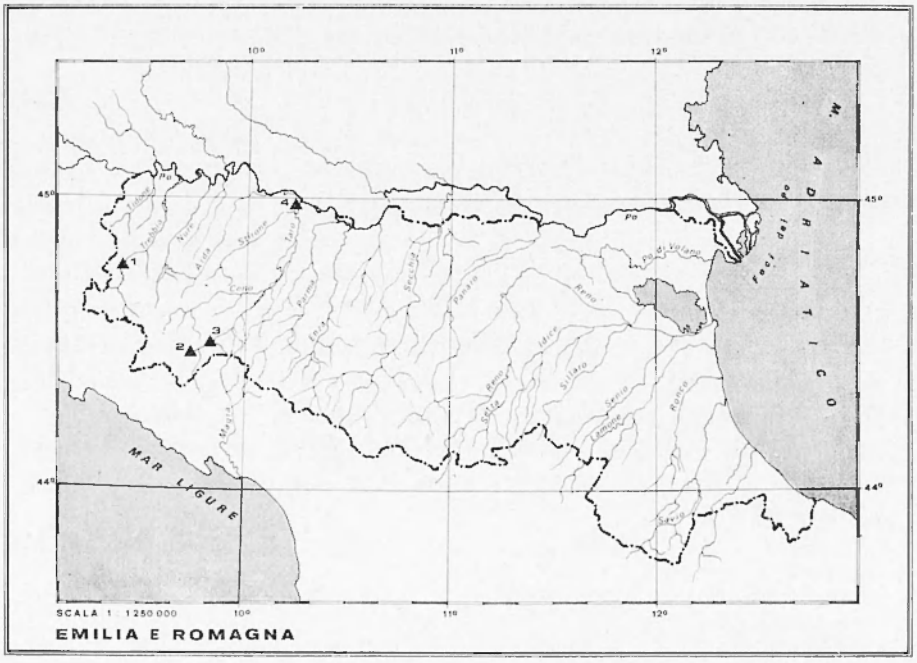

Fig. 1 – The Emilia Romagna Region.

The results of the reconstruction are shown in Fig. 2, 3 and 4. The solid line refers to the observed values, the broken line to the reconstructed values and the dotted line to the standard deviation of the reconstruction. The values at time t^* are reconstructed, using the observations at time t^*-1, t^*, t^*+1 at the guide station. Fig. 5 reports the duration curves of the observed and reconstructed series. To use the observations at time t^*-2 and t^*+2 as well does not seem to bring much improvement, given the low value of lag 2 covariances. The mean square error of the reconstruction at stations 1, 2 and 4 are 14.03,0.51 and 136.48, respectively. If one considers that the mean square deviations from the monthly mean are 63.09, 1.56, and 679.37, respectively, one may conclude that about 75% of the variance of the processes is accounted for.

Fig. 6 shows the predicted values at station 2 based on past observations at the stations 1, 2, 3, 4. The solid, and broken lines have the same meaning as before. The poor quality of the prediction is due to the poor lag 1 correlation of the series. In fact γ $(i, j, 1)$ is about 0.3. This is due to the nature prevalently impermeable of the ground. However the predictor works considerably better of the monthly means (dotted line). The covariances at lags —1, 0, and 1 between the standardized value at stations 1, 2, 3 and 4, and the standardized value at station 3, are shown in Table 1.

TABLE 1

Covariances between the standardized values at Stations 1, 2, 3 and 4 and the standardized values at station 3

STATION 3

	Lag —1	Lag 0	Lag 1
Station 1	0.308	0.834	0.320
Station 2	0.170	0.718	0.257
Station 3	0.393	1.000	0.393
Station 4	0.297	0.877	0.366

AKNOWLEDGEMENT

The Author wishes to thank IDRO. Ser. and IDROTECNECO for the permission to publish this article.

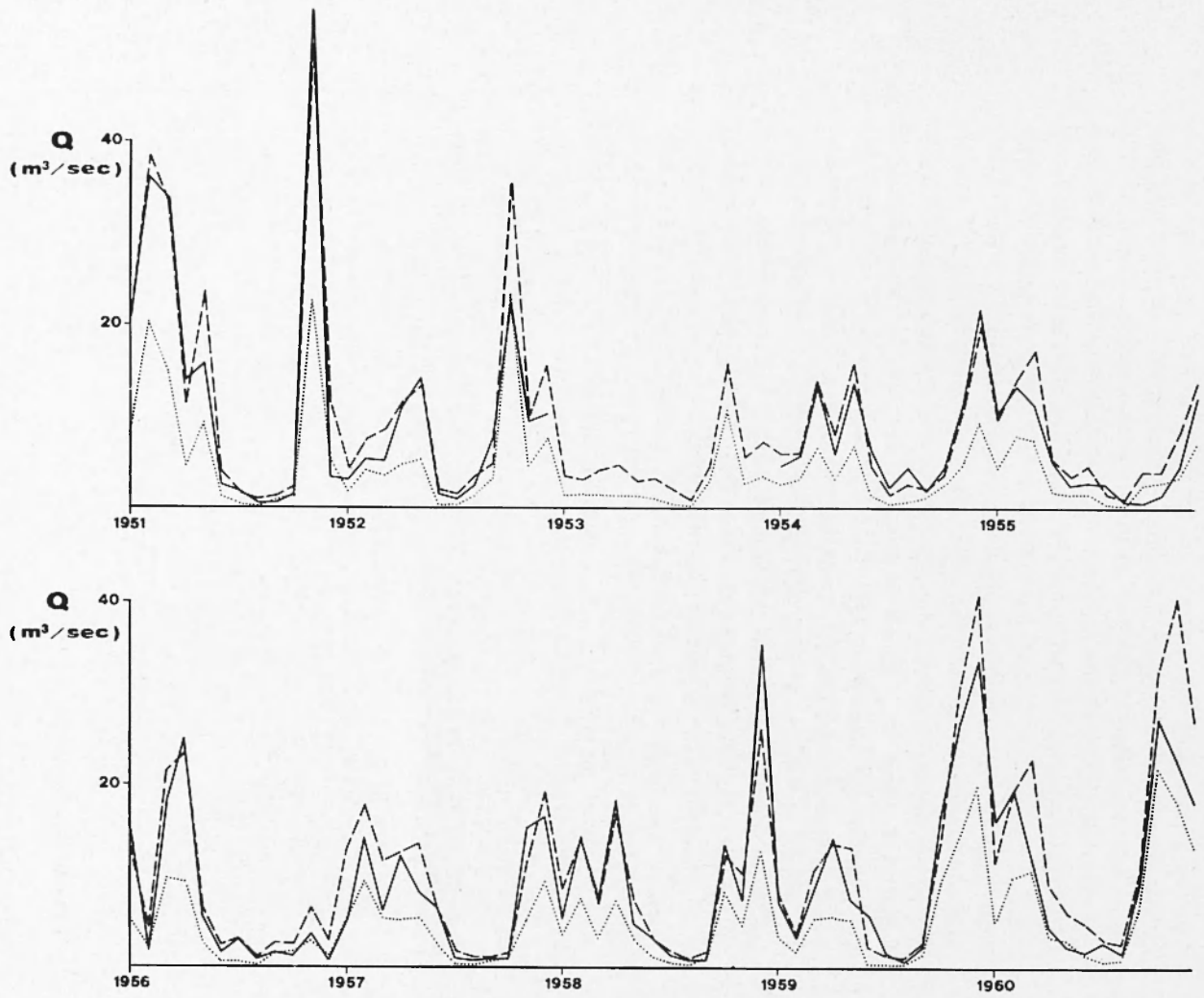

Fig. 2 – Observed values (solid line), reconstructed values (broken line) and standard deviation of the reconstruction (dotted line) at station 1. The guide station is station 3.

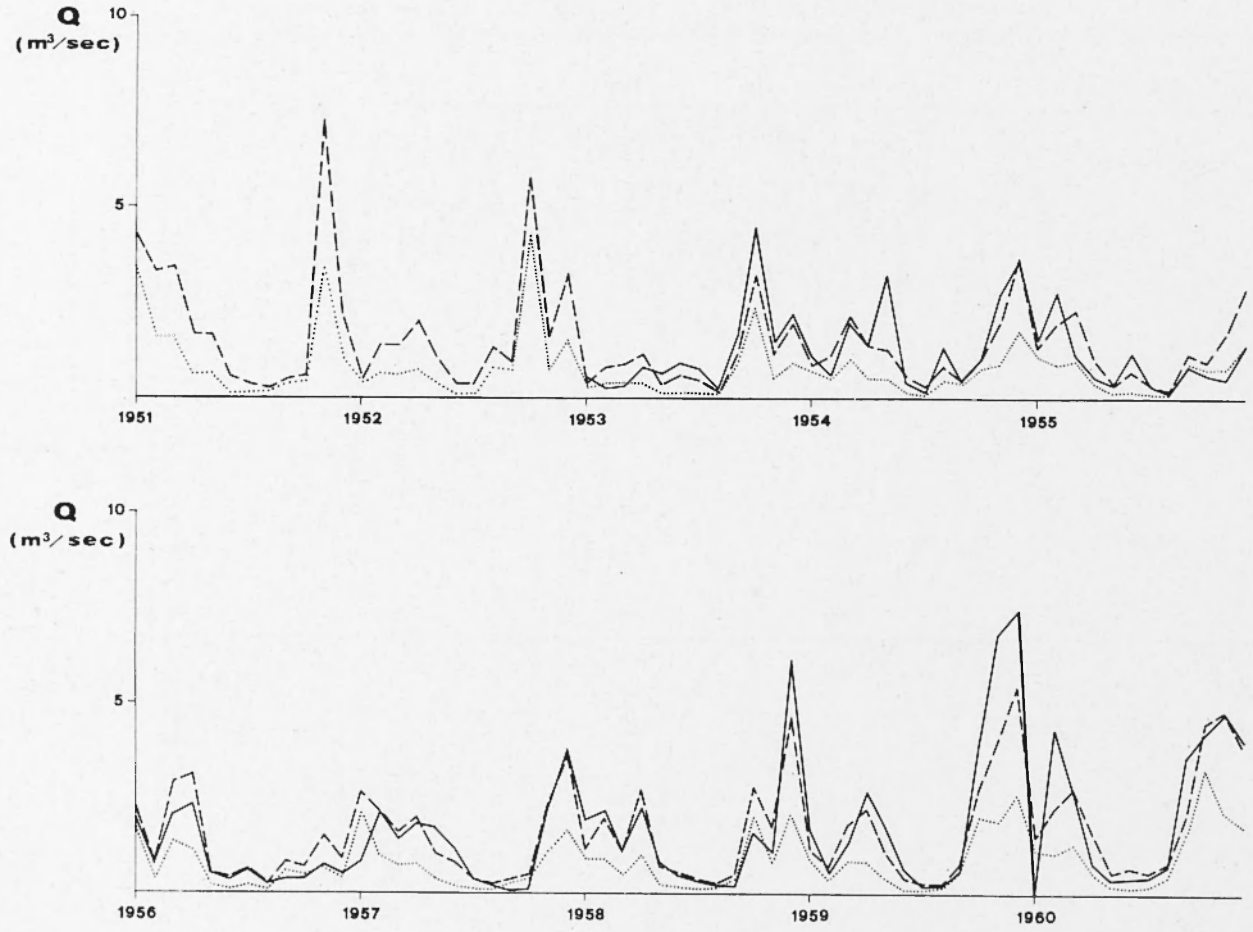

Fig. 3 – Observed values (solid line), reconstructed values (broken line) and standard deviation of the reconstruction (dotted line) at station 2. The guide station is station 3.

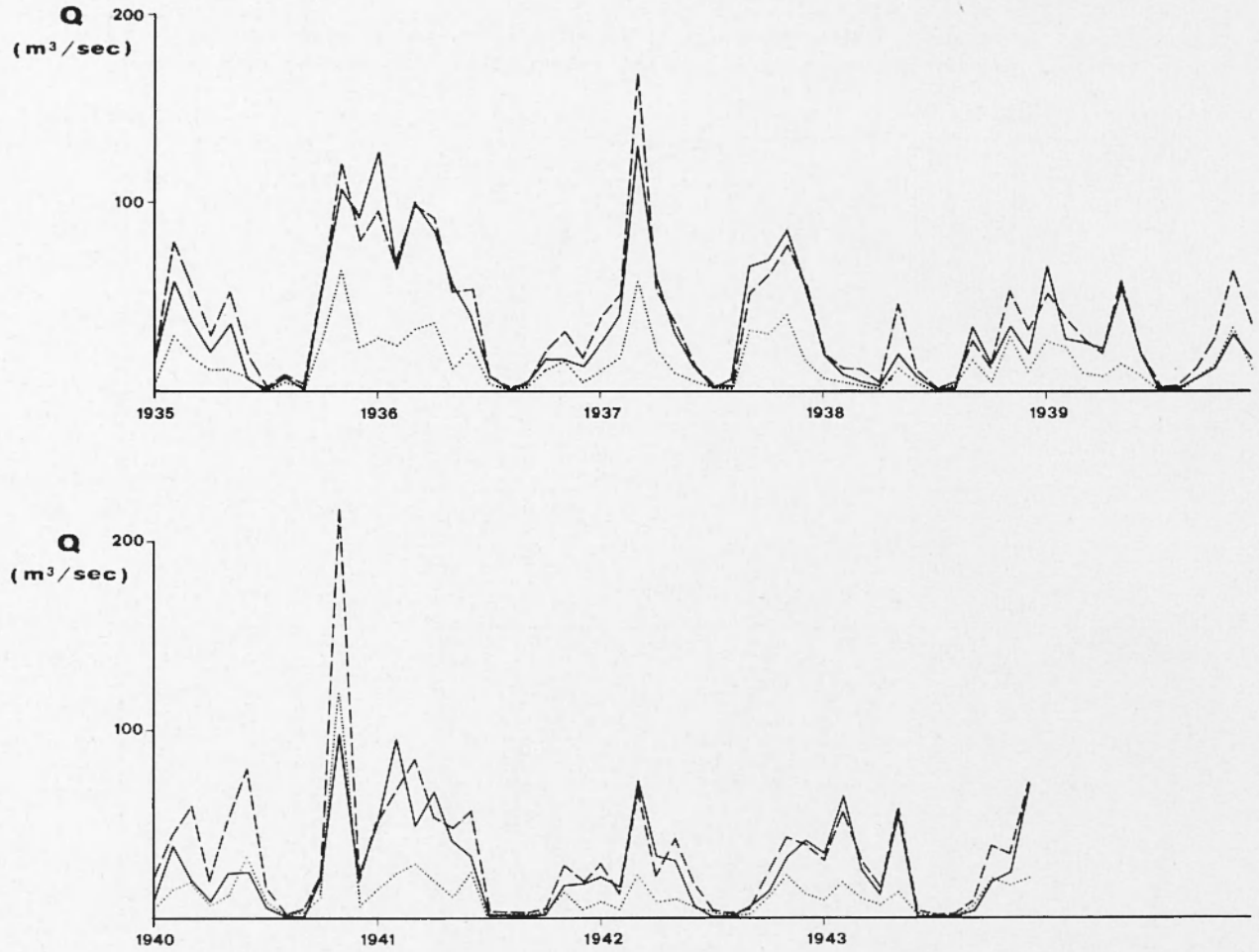

Fig. 4 – Observed values (solid line), reconstructed values (broken line) and standard deviation of the reconstruction (dotted line) at station 4. The guide station is station 3.

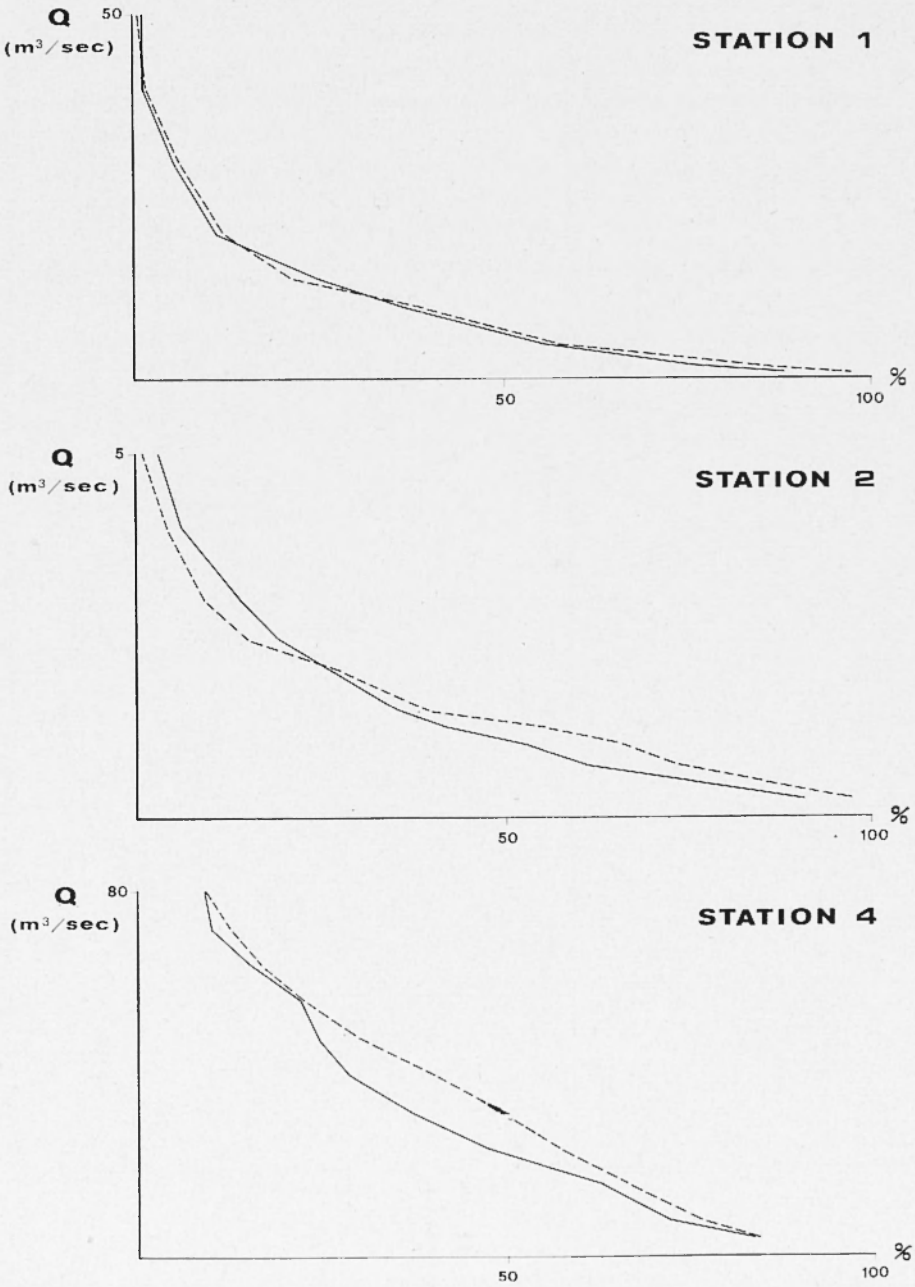

Fig. 5 – Duration curves at stations 1, 2 and 4; observed series (solid line) and reconstructed series (broken line).

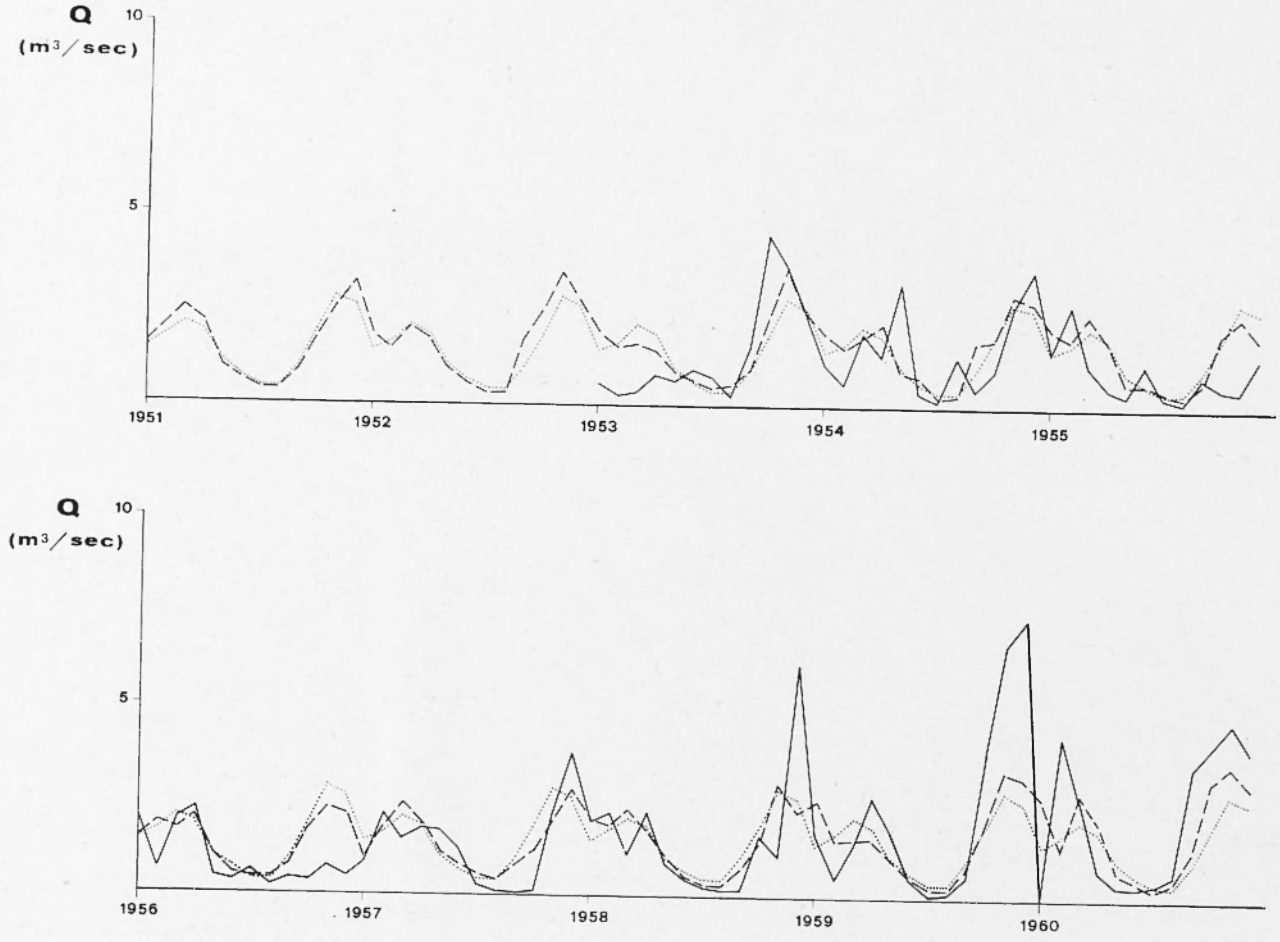

Fig. 6 – Observed values (solid line), predicted values (broken line) and monthly means (dotted line) at station 2.

REFERENCES

ANDERSON T. W., 1958. – *Introduction to multivariate statistical analysis.* "Wiley", New York.

TORELLI L., and CHOW W. T., 1972. – *Tests of stationarity of Hydrologic time series.* "Int. Symp. on Uncertainties in Hydrologic and Water Resource Systems", Tucson, Arizona, 11-14 Dic.

TORELLI L., 1973. – *The analysis of monthly hydrologic time series.* Ph. D. thesis, University of Illinois.

TORELLI L. and TOMASI P. – *Previsione e simulazione delle portate medie mensili del fiume Tevere a Ripetta.* "Idrotecnica", in print.