The correction of non-homogeneous time series in preparation for hydrological forecasting

Z. BALINT  District Water Authority Nyiregháza, Hungary

Abstract. Some critical problems of the application of regression analysis for hydrological forecasting are investigated. The correction of non-homogeneous data series is made by comparing the series with a homogeneous basic series. The selection of the variables on the one hand uses dynamic programming, and on the other hand a new partial correlation coefficient. Nonlinear relations are taken into account by transforming the empirical distributions. The correction procedure is described in detail, but because of the limited length of the study only a very brief summary of the other two themes is given.

INTRODUCTION

Forecasts based on regression analysis are performed on the basis of relationships obtained from data series observed in the past. Such relationships from historical data should remain valid for the future too. To fulfill this assumption two requirements must be satisfied: (1) the data series must not contain any systematic error, and (2) all of the hydrological conditions should be constant. These are seldom satisfied in practice because of the non-homogeneity and inconsistency of the time series, as clearly illustrated by Yevjevich and Jeng (1969). That is why some correction of the series is often necessary. Sophisticated investigations can be accomplished only after first preparing series which truly reflect hydrological phenomena. This is why I give detailed description of the correction procedure and only a summary of the variate selection process and the analysis of nonlinearity. The examples concern the upper parts of the rivers Tisza and Szamos in northern Hungary.

DEFINITION OF HYDROLOGICAL HOMOGENEITY

We have no right to assume that natural phenomena are mathematically homogeneous. Thus if we make them homogeneous with a mathematical procedure we may spoil the relations existing between the hydrologically homogeneous ("true") values of the variables reflecting the phenomena free of transformations.

Any data series \( x \) may contain four elements: periodicity (an effect of climate) \( P \); continuous change or trend (e.g. changes of the river bed) \( T \); jumps (e.g. rebuilding of a gauge) \( J \); and random fluctuation \( e \); so

\[
x_t = P_t + T_t + J_t + e_t
\]

The periodicity and random fluctuation components result from the climate's natural properties, thus they must not be corrected if one wants to obtain a series.
which properly reflects the phenomenon. Thus a series is hydrologically homogeneous if:

$$x_t = P_t + \epsilon_t$$

$$T_t = 0 \text{ and } J_t = 0$$

COMPARATIVE HOMOGENEITY TEST

The substance of the comparative homogeneity test

The basic assumptions are: (1) that a minimum of one hydrologically homogeneous basic series $x_B$ can be identified from among the several series from gauges elsewhere within the basin (otherwise we can construct a fictitious series $x_{fB}$); and (2) the oscillations of the periodic components have identical wavelengths for all of the series and are proportional to the standard deviations. Under these assumptions the difference of the standardized elements of the investigated $x$ from the basic series $x_B$ produces a new series $\delta$ having no periodicity and having a random fluctuation less than that of the original investigated series if there is a correlation between the two variables. The continuous sum of the $\delta$ series forms a new series $z$, which when plotted can assist us in determining the type (continuous change or jump) and the location in time of the transforming effect. The presumed location can be more strictly and physically controlled and the mathematical tests and corrections can be better accomplished on the basis of the $\delta$ series.

The basic series

The role of the basic series is to eliminate the periodic component and reduce the random fluctuation. As can be seen from a number of past studies, e.g. Csoma and Szigyárto (1975), Kovács (1974), and Yevjevich (1972), the determination of the periodicity is a problem still not completely solved. That is why the test described here accepts the periodicity of the basic series as proportional to that of the investigated series and eliminates it by subtracting them. To justify this assumption the series must be from gauges which are not too far apart; otherwise using auxiliary stations the test can be performed step by step. If no single homogeneous series is available, a fictitious basic series must be constructed by pooling the data from several comparable series within an area known to be not too badly affected by errors.

The difference of the series

Besides the elimination of the periodicity the advantage of the difference series is that it can be proved that the random error of $\delta$ is decreasing if the correlation coefficient between $x$ and $x_B$ is greater than 0.5.

Evaluation of the continuous sum

The continuous sum graph is descriptive because it accumulates the systematic error. Figure 1 shows a series $x$ having a periodic and a random component and two jumps $U_1$ and $U_2$. The calculated $\delta$ series has no periodic component but contains two standardized jumps $\delta^*_1$ and $\delta^*_2$. The $z$ series fluctuates around a line made up of three straight limbs. The positions of the breaks mark the locations of the jumps. This line can be determined by, for example, least squares.

In the case of Fig. 2 the best fitting line to the $z$ series is a curve which indicates that there is a continuous change or trend in the series. If the $z$ graph is composed of parallel straight lines as in the initial portion of the rising limb of Fig. 4, there are insufficient data on the nature of the shift. The outcome of the analysis must be interpreted, first of all, by looking for possible physical sources for the effect. If we
can find no physical cause for the transformation we can test our hypothesis and correct the series mathematically as follows.

**t-test**
Student's t-test can be applied if the hypothesized changes take the form of jumps. The so-called null hypothesis in this case is $H_0: M(\delta) = 0$, i.e., we judge whether the difference of the mean of the $\delta$ series, $M(\delta)$ from zero could occur by chance or not. The test must be performed separately for each of the sections of the $\delta$ series.

**F-test**
Fisher's $F$-test can be applied for testing either jumps or continuous change in the following form: $H_0: D(\delta) = D(\eta)$ where $\eta$ is the fluctuation of the $\delta$ variable around the hypothesized regularity, and $D$ is the standard deviation. $D(\eta)$ should be expressed as a single value for the whole series.

**Stripe-test**
Both the $t$ and the $F$-test assume that the distribution function of $\delta$ is normal (Gaussian). This is seldom satisfied in practice. Both tests investigate one property
of the section, either $M(\delta)$ or $D(\delta)$, but we should test the hypothesis preferably on
the basis of the best fitting line. This is done by the stripe-test which is applicable
for discrete random variables having any kind of distribution function.

The hypothesis to be tested is: the elements of the investigated section fluctuate
around a straight line within a stripe of width $B$. The test judges: what is
the probability of forming the above hypothesis if the sample elements are drawn
randomly and independently. A straight line is determined by two points, i.e. the
probability of fitting perfectly a line to the first two points is 1.0. Thus

$$p_3 = P(x_3 \in B_3 | x_1 \in B_1 \cap x_2 \in B_2)$$

$$p_k = P(x_k \in B_k \cap x_{k-1} \in B_{k-1} \cap \ldots \cap x_3 \in B_3 | x_1 \in B_1 \cap x_2 \in B_2)$$

where $p_k$ is the probability of the first $k$ elements of the section being allocated
within the stripe, and $B_t$ indicates the place and width of the stripe at the $ith$
element. The $p_{Bi} = P(x_t \in B_i)$ probability can be determined as shown in Figs 1 and
2, by determining the distribution function of the series and subtracting the
probabilities contained within the limits of the stripe at the $ith$
element.

Provided the sample elements are independent, the $p_k$ conditional probability is

$$p_k = p_{B3} \cdot p_{B4} \cdot \ldots \cdot p_{Bk-1} \cdot p_{Bk} = \prod_{i=3}^{k} p_{Bi}$$

If $p_{Bi} = p_B = \text{constant}$ (Fig. 1) then $p_k = p_B^{k-2}$.

The criterion of the decision for accepting or rejecting the series as a homogeneous
one can be the same as used in the other tests.

**Correction of the series**

Having shown that the series are non-homogeneous the correction can be made on the
basis of the $\delta$ series by calculating the means of the sections. This provides the
standardized jump $U^*$ or trend $\Delta^*$. The actual values are $U = U^* \sigma_x$ and $\Delta = \Delta^* \sigma_x$. A
repeat investigation may be necessary after completing the procedure because the
correction alters the distributions too.

**EVALUATION AND PRACTICAL EXPERIENCE**

The comparative homogeneity test proved to be very useful in the case when the
cause of the non-homogeneity is the change of the mean. The results of our fore­
casting model based on multivariate regression improved more because of the
correction than because of taking into account more than three variables or non­
linear relations.

Figures 3 and 4 show two examples, one for the River Tisza and one for a main
tributary of the Tisza, the Szamos. The duration of record from which the data
(maximum flood stages) were collected is about 25 years. The series for Huszt on
the River Tisza includes three jumps. All of the empirical distributions $(x, x_B, \delta)$
are given in Fig. 3. The series for Szatmârnémeti on the River Szamos contains only
one jump but it also contains three deficient data. These stages originated from icy
floods. Precisely at the time of the jump the river gauge was reconstructed but the
change of its level was not given precisely. Both empirical distributions in Fig. 4
concern the $\delta$ series differing only in the scale.

Table 1 contains the results of the different mathematical tests concerning the
$\delta$ series for Huszt. $|t|$ and $F^*$ are calculated from the standard deviation derived
from the entire series, while $|t_j|$ and $F_{jt}^*$ are based on the standard deviation around
the hypothetical jump, calculated separately for the different sections. $H_0$ means the
$H_0$ hypothesis is true (the series can be accepted as homogeneous) while $H_1$
means the jump can be accepted. $t_{95}$ and $F_{95}$ are the critical values.
The correction of non-homogeneous time series

FIGURE 3. Comparative homogeneity test of Huszt (x) - Tiszabecs (x_B), River Tisza.

FIGURE 4. Comparative homogeneity test of Szatmárnémeti (x) - Csenger (x_B), River Szamos.
TABLE 1. The results of the mathematical tests for Huszt on the River Tisza.

<table>
<thead>
<tr>
<th>part</th>
<th>averages of the parts of ( d )</th>
<th>standard deviations of the parts of ( d )</th>
<th>( t )-test</th>
<th>( F )-test</th>
<th>String-test</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. ( k = 9 )</td>
<td>( \bar{d} = -0.267 )</td>
<td>( d_1 = 0.707 )</td>
<td>( 1.68 )</td>
<td>( H_0 )</td>
<td>( F^* = 1.79 )</td>
<td>( H_0 )</td>
</tr>
<tr>
<td>II. ( k = 26 )</td>
<td></td>
<td>( d_1 = 0.594 )</td>
<td>( 1.60 )</td>
<td>( H_1 )</td>
<td>( F^* = 1.61 )</td>
<td>( H_0 )</td>
</tr>
<tr>
<td>III. ( k = 11 )</td>
<td></td>
<td>( d_1 = 0.570 )</td>
<td>( 1.59 )</td>
<td>( H_1 )</td>
<td>( F^* = 1.75 )</td>
<td>( H_1 )</td>
</tr>
<tr>
<td>IV. ( k = 26 )</td>
<td></td>
<td>( d_1 = 0.378 )</td>
<td>( 1.59 )</td>
<td>( H_2 )</td>
<td>( F^* = 1.85 )</td>
<td>( H_2 )</td>
</tr>
<tr>
<td>total series ( n = 72 )</td>
<td></td>
<td>( d_1 = 0.46 )</td>
<td>( 1.59 )</td>
<td>( H_1 )</td>
<td>( F^* = 1.84 )</td>
<td>( H_1 )</td>
</tr>
</tbody>
</table>

**SELECTION OF THE INDEPENDENT VARIABLES**

**Application to dynamic programming**

The number of the independent variables is continuously extended. The decision on the optimum policy is made on the basis of the multiple correlation coefficient. The reliabilities of the regression coefficients are taken into account as limits.

**Partial correlation coefficients**

To avoid some problems emerging when using the partial correlation coefficients a new coefficient is introduced:

\[
r_{X_1Y} \cdot x_1, x_2 \ldots x_n = \left( 1 - \frac{\sigma_e^2}{\sigma_{X_1}^2 + \sigma_e^2} \right)^{1/2}
\]

where

- \( \sigma_e \) = standard deviation of the error of the forecast,
- \( \sigma_{X_1} \) = standard deviation of \( X_1 \) independent variable,
- \( a_i \) = regression coefficient in equation:

\[
Y_{if} = a_0 + a_1X_1 + \ldots + a_mX_m
\]

\( Y_{if} \) = forecasted dependent values.

**TRANSFORMATION OF NONLINEAR SYSTEMS**

Contrary to the generally used transformations this method does not transform the direct relations but the empirical distribution functions of the variables. It is proved that the relationships can be considered approximately linear if all the variables have normal (Gaussian) distribution functions.
REFERENCES


