

Generation of monthly synthetic streamflow series based on the method of fragments

A. T. Silva & M. M. Portela

Department of Civil Engineering, IST, Lisbon, Portugal

Abstract

Synthetic time series generation has long been an important tool for the planning and management of water resources systems. This technique allows for a significant reduction of the uncertainty associated with hydrological phenomena. In this article a procedure is proposed for generating synthetic series of annual and monthly flows that combines two models, a probabilistic one, applied at an annual level, and at a monthly level, a deterministic disaggregation model. The modeling of the annual flow series is based on the random sampling of the log-Pearson III law of probability. The disaggregation of annual flows into monthly flows uses the method of fragments. For the application of this method, a new procedure was developed and tested for the automatic definition of the classes of fragments, reducing the need for intervention of the modeler, resulting in a more general and robust approach. The combination of the two models was tested on a data set of 54 streamflow samples from gauging stations geographically spread over Mainland Portugal. For each gauging station, 1200 synthetic series were generated, with a length equal to that of the corresponding sample. The quality of the generated series was evaluated by their capacity to preserve the most significant statistical characteristics of the samples of annual and monthly flows, namely the mean, standard deviation, and skewness coefficient. Confidence intervals were established for this evaluation, and the results show that, generally, the statistics of the samples are contained in these intervals. Thereupon it was concluded that the developed procedure is adequate.

Keywords: generation of synthetic series, disaggregation models, method of fragments, definition of classes of fragments, log-Pearson III law.



1 Introduction

The utilization of synthetic streamflow series is an important and useful tool for practitioners and researchers on water resources management, as it enables the assessment of the uncertainty associated with natural hydrological phenomena in general and with the flow regime in particular. In the understanding that a sample of observed streamflows comprises a single event, namely the historical one, synthetic series obtained from that sample are considered to be alternate events with a probability of occurrence that is admittedly equal to that of the observed event. By means of statistical analysis, it is possible to assign probabilistic criteria to the performance of the hydraulic systems in whose analysis synthetic flow series were applied.

The generation of synthetic series can be made simultaneously at different time levels using a disaggregation technique. This technique encompasses a model for generating values at a given time level, namely the year, combined with a model to disaggregate these values into a lower level, namely the month, while achieving a preservation of the main statistical characteristics of the samples at both time levels, such as the means, the standard deviations, and the skewness coefficients.

The objective of the research presented in this paper was to develop and test a methodology for generating annual and monthly synthetic streamflow series, integrating a probabilistic generation model at the annual level, and a disaggregation model – the method of fragments – at the monthly level. To enable the application of that methodology to a large number of observed streamflow samples, with different lengths and statistical characteristics, a procedure was developed and tested for the automatic definition of the classes of fragments, resulting in a much more general and robust approach, thus reducing the need of intervention of the modeler.

The data set utilized in the research consisted of 54 monthly streamflow samples from gauging stations geographically spread over Mainland Portugal – Figure 1.

2 Proposed methodology

2.1 General considerations

The proposed methodology to generate synthetic monthly streamflow series consisted of a two-level approach, comprehending a generation model for the annual synthetic series, and a disaggregation model to obtain the corresponding monthly flows.

According to Quintela [1], the annual streamflow series of Portuguese rivers is an independent time series if the hydrologic year (starting on October 1st) is adopted as a time-step, provided that an analysis of the correlation structure of the series does not invalidate this assumption. Hence, the modeling of the annual streamflow series used a non-deterministic probabilistic model, based on the random sampling of the log-Pearson III law.



The disaggregation of the generated annual flows into monthly flows used the method of fragments, proposed by Svanidze in 1961 (according to Santos [2, p.19]). This method considers that the within-year distribution of streamflows is identical in years with similar values of the annual streamflow. The method of fragments is capable of generating monthly streamflow series while preserving monthly and annual statistic.

It should be noted that the research underlying this study was carried out in the scope of a larger study on the development of criteria for the design of the storage capacity of artificial reservoirs. Such framework conditioned some of the criteria applied to the generation of synthetic flow series, namely, the number of synthetic series, M , to generate from each N -year sample, which was fixed as 1200 in accordance with what Guimarães [3, p.175] indicates concerning reservoir storage design studies.

2.2 Generation of annual streamflows

The model for generating annual flows consists of a random sampling of the log-Pearson III distribution, which is the Pearson III distribution applied to the logarithms of the flow, that is, if $W = \ln(X)$ follows the Pearson III distribution, then X follows the log-Pearson III distribution.

Because that distribution can only take values between 0 and ∞ , the generation of negative flows is naturally avoided.

The log-Pearson III law has a flexible distribution, assuming a number of different shapes depending on the mean, variance and skewness of the sample, hence the treatment of skewed data is assured [4, p.375].

For each sample, the probabilistic model generates logarithms of annual flows, \hat{W}_i ($i=1,2,\dots,N$), where N is the length of such sample. The synthetic annual flows, \hat{X}_i , are then obtained by inverting the logarithmic transformation.

The model is carried out M times until M synthetic annual flow series are generated, M being fixed on 1200, as previously justified [3, p.175].

2.3 Monthly disaggregation of streamflows

The method of fragments considers that the within-year distribution of streamflows is identical in years with close annual flows. In this method, for a given year k the observed monthly flows are divided by the corresponding annual flow, X_k , the resulting set of twelve non-dimensional monthly flows constitutes the fragment pertaining to the year k , ϕ_k :

$$\phi_k = \frac{X_{k,j}}{X_k} = \left[\frac{X_{k,1}}{X_k} \quad \frac{X_{k,2}}{X_k} \quad \dots \quad \frac{X_{k,11}}{X_k} \quad \frac{X_{k,12}}{X_k} \right] \quad (1)$$

where $X_{k,j}$ are the monthly flows ($j = 1,2,\dots,12$).



The application of the method of fragments based on a given sample requires that the fragments are constituted and assembled into classes beforehand. For the constitution of the array of fragments, $[\phi]$, the annual flows must previously be sorted from the smallest to the largest:

$$[\phi] = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_k \\ \vdots \\ \phi_N \end{bmatrix} = \begin{bmatrix} \frac{X_{1,1}^*}{X_1^*} & \dots & \frac{X_{1,12}^*}{X_1^*} \\ \vdots & \ddots & \vdots \\ \frac{X_{k,1}^*}{X_k^*} & \dots & \frac{X_{k,12}^*}{X_k^*} \\ \vdots & \ddots & \vdots \\ \frac{X_{N,1}^*}{X_N^*} & \dots & \frac{X_{N,12}^*}{X_N^*} \end{bmatrix} \quad (2)$$

where X_k^* represents the k^{th} annual flow such that $X_{k-1}^* \leq X_k^* \leq X_{k+1}^*$ and $X_{k,j}^*$ the monthly flows of that year.

The classes of fragments are defined by the clustering of annual flows with similar values, and they may or may not have a fixed amplitude. In the state of the art involving the method of fragments, there is no known rule for the number and size of classes to adopt, which constitutes a source of ambiguity concerning the utilization of the method [5, p.84].

A number of alternate approaches for defining the classes of fragments was devised and compared, in terms of robustness regarding the preservation of the statistical parameters of the samples and its ability to be automated. The one that proved best considers the classes of fragments defined as probability intervals.

The procedure applied is as follows:

1. Consideration of a series of nine equally spaced non-exceedance probabilities, F_m ($m = 1, \dots, 9$), with an increment of 10%, that is: $F_1 = 10\%$; $F_2 = 20\%$; ...; $F_9 = 90\%$.
2. Estimation of the annual flows, \widehat{X}_m , for the aforementioned non-exceedance probabilities by inverting the probability distribution function of the log-Pearson III law. This results in ten equally probable classes with limits represented by $[0; \widehat{X}_1]$, $[\widehat{X}_1; \widehat{X}_2]$, ..., $[\widehat{X}_8; \widehat{X}_9]$, $[\widehat{X}_9; +\infty]$.
3. Distribution of the fragments among the successive classes, while simultaneously verifying if there are empty classes. In case each class has at least one fragment, the procedure for the definition of the classes is complete and one can proceed to the monthly disaggregation model; otherwise one needs to take the next additional step.
4. Redefinition of empty classes:

- If the first class is empty, it is included in the next class, by changing its upper limit to the value of $[0; \widehat{X}_2[$. If it remains empty, it is included in the next class and so on and so forth.
- If the last class is empty, it is included in the previous class, by changing the lower limit to the value of $[\widehat{X}_8; +\infty[$. If it remains empty it is included in the previous class and so on and so forth.
- If an intermediate class, $[\widehat{X}_m; \widehat{X}_{m+1}[$, is empty, half of the probability step that defines this class is attributed to the two bordering classes.

After the classes of fragments are defined, and having previously generated the synthetic series of annual flows, the method for generating the monthly flows proceeds with the identification, for each year i of the synthetic annual flow series, \widehat{X}_i , of the class to which that annual flow belongs – class $m+1$ such that $\widehat{X}_i \in [\widehat{X}_m; \widehat{X}_{m+1}[$.

The next step is to select the fragment ϕ_i to be used to disaggregate the annual flow \widehat{X}_i . The fragments are randomly selected, without replacement until the class runs out of fragments. When this happens, the class is refilled with the respective fragments.

Each generated annual flow, \widehat{X}_i , is disaggregated into twelve generated monthly flows in accordance with the following system:

$$\widehat{X}_{i,j} = \phi_i \widehat{X}_i \quad (3)$$

The disaggregation procedure is repeated until M N -year synthetic series of annual and monthly streamflows are generated.

2.4 Assessment of the quality of the synthetic series

The results of the generation model at both temporal levels were evaluated in terms of the capacity of the generated synthetic flow series to preserve the main statistical characteristics of the corresponding samples. This assessment is made by comparing the synthetic series with the samples in terms of the main historical statistics. The synthetic series are considered to be valid alternative sequences to the sample series if the sample's statistics are preserved, thus enabling the use of the generated series for the purpose for which they were needed [3, p.81].

If M synthetic series are generated, the comparison of a given sample series' statistic, θ (such as the mean, the standard deviation or the skewness coefficient), with the corresponding synthetic series' statistics $\hat{\theta}^{(m)}$ ($m=1,2,\dots,M$), can be made by obtaining the mean ($\bar{\theta}$) and standard deviation ($s_{\hat{\theta}}$) of $\hat{\theta}^{(m)}$ and by establishing a confidence interval of $(1-\alpha)$ for θ ,



$$\left] \bar{\theta} - z_{1-\alpha/2} s_{\bar{\theta}}; \bar{\theta} + z_{1-\alpha/2} s_{\bar{\theta}} \left[\quad (4)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Normal standard distribution, α being the significance level. If the statistic θ of the sample series belongs to that interval, it is considered to be preserved in the synthetic series.

3 Streamflow data

The streamflow data set used in the analysis consists of 54 continuous monthly and corresponding annual samples of historical streamflows from gauging stations geographically distributed over Mainland Portugal – Figure 1. The average length of the samples is close to 35 years, while the shortest and the longest samples respectively have the lengths of 11 years and 73 years. The catchment area varies from 4 to 2718 km². Concerning the mean annual flow depth, the data set displays a wide range of values, between 90.4 mm and 2214.1 mm.

The correlation structure of the annual streamflow samples was analyzed by computing the lag-one and lag-two serial correlation coefficients, as proposed by

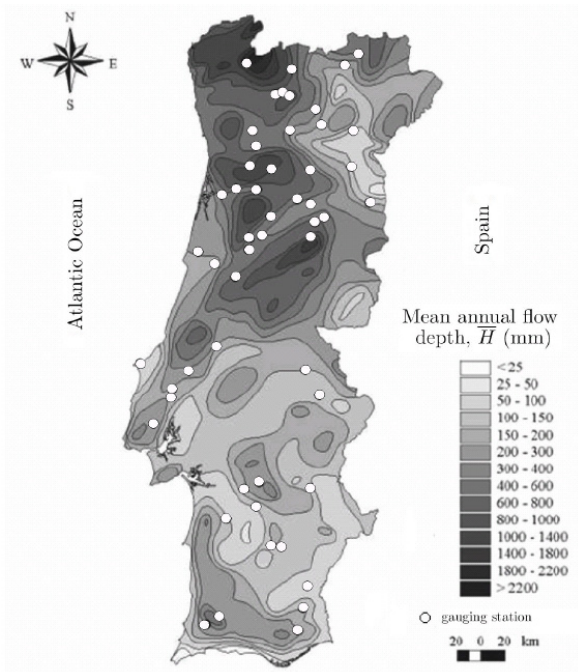


Figure 1: Mainland Portugal. Schematic location of the gauging stations used in the study. Contour map of the mean annual flow depth, after Portela and Quintela [6].

Salas *et al.*, [7, pp.48–49]. The results of that analysis proved that the historical samples of annual streamflows constitute independent time-series, provided that the hydrological year (starting on October 1st) is adopted as a time-step. Therefore the log-Pearson III law was adopted at the annual level.

4 Results and discussion

4.1 Previous considerations

As described in Item 2, for each of the 54 streamflow samples of the data set, a two-level streamflow generation model was applied to generate 1200 synthetic annual and monthly streamflow series with an average length close to 35 years. The results of the models at both temporal levels were evaluated in terms of the capacity of the generated synthetic series to preserve the main statistical characteristics of the corresponding samples. This assessment is made by comparing the synthetic series with the samples in terms of the main historical statistics as described in Item 2.4. The comparison utilized confidence intervals, computed by eqn (4). A confidence level of $1 - \alpha = 95\%$ was adopted for the confidence intervals.

4.2 Results at the annual level

The generation of the annual streamflows takes place in the domain of the logarithmic transforms of the random variable (annual streamflows), hence, the confidence intervals and the samples' statistics to be used when appraising the quality of the model refer to the logarithms, namely the means, standard deviations and skewness coefficients of the logarithms of the annual streamflows.

Figure 2 displays the representation of those confidence intervals, together with the values of the samples' statistics. In the figure, each sample is identified in the horizontal axis by a number (from 1 to 54). The samples were ranked by increasing mean annual flow depth.

The analysis of Figure 2 reveals that, for every sample, the historical values of the statistics are contained within the corresponding confidence intervals.

An additional analysis was made by obtaining similar diagrams but in the domain of the random variable itself (the annual streamflow, in this case). However the representation of such diagrams in a single figure is not feasible due to the huge amplitude of the values of the means and of the standard deviations when referred to flow units.

Notwithstanding this limitation, the appreciation of the preservation of the aforementioned statistical characteristics, however estimated on the basis of the annual flows themselves, showed that the inversion of the logarithmic transformation did not compromise the preservation of the means, of the standard deviations or even, in the majority of the cases, of the skewness coefficient of the annual streamflows. In fact, there was only one case, out of 54, where the skewness coefficient was not preserved.



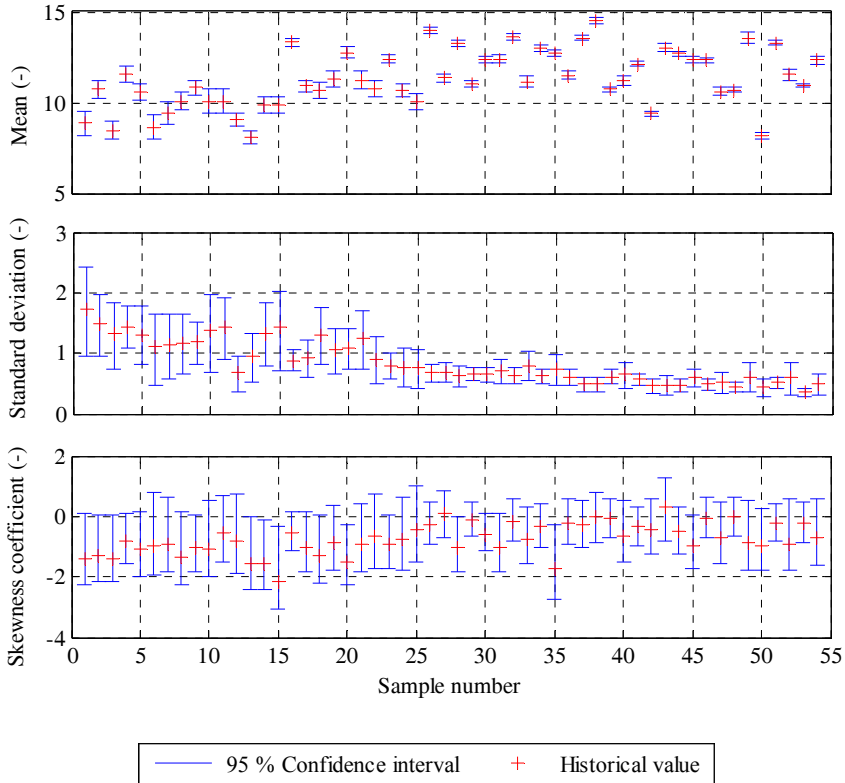


Figure 2: Confidence intervals at 95% of the means (top), the standard deviations (center), and the skewness coefficients (bottom) of the logarithms of annual streamflows. Comparison with the historical values of the same statistics.

Such results, along with those pertained to the logarithms of the streamflows are summarized in Table 1 which specifies the number of samples whose statistics were not preserved by the generation model.

Table 1: Results of the analysis of the preservation of the annual streamflow samples' statistics: number of samples – out of 54 – whose means, standard deviations and skewness coefficients were not preserved by the generation model.

Statistic	Logarithm of the flow	Flow
Mean	-	-
St. dev.	-	-
Skew. coeff.	-	1



4.3 Results at the monthly level

At the monthly level, the analyzed statistics were the means, standard deviations and skewness coefficients of the monthly flows in each month of the hydrologic year.

It should be mentioned that the calculation of the skewness coefficient was not always possible due to the fact that, in some samples, the flow in certain months is null in every year of the records, which is common in Portugal, especially in the Southern regions. Furthermore, the disaggregation process generates a number of synthetic monthly flow series of N years that present months with null flows in every year. In these cases it is also impossible to estimate the skewness coefficient. In order to standardize procedures, if null flows in one or more months always occur in one or more of the 1200 synthetic series generated from a given sample, the confidence intervals of the skewness coefficients of those months were not established, and therefore the analysis of the preservation of the skewness coefficient of monthly flows was not made in those months. There was a total of 16 months in such situation. It should be stressed, however, that the analysis of the preservation of the means and standard deviations focused on $54 \times 12 = 648$ samples of monthly flows, a number that for the skewness coefficient, was thus reduced to 632 samples.

The application of the models at the monthly level showed that, for all 54 analyzed samples, the means of the monthly flow were always preserved. The standard deviations and the skewness coefficients were not preserved in a total of 3 and 10 months respectively.

The results of the analysis at the monthly level are displayed in Table 2 which specifies the number of samples (out of 54), and months whose statistics were not preserved by the disaggregation models.

Table 2: Results of the analysis of the preservation of the streamflow samples' statistics: number of samples – out of 54 – whose means, standard deviations, and skewness coefficients were not preserved by the disaggregation model.

Statistic	Month											
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Mean	-	-	-	-	-	-	-	-	-	-	-	-
St. dev.	2	-	-	-	-	-	-	-	-	-	1	-
Skew. coeff.	1	-	1	2	-	-	1	2	-	-	-	3

The results represented in Figure 2 and in Tables 1 and 2, clearly show that, generally, the proposed methodology for generating annual and monthly synthetic streamflow series accurately preserves the main statistical characteristics of the historical streamflow records in Portuguese rivers.



5 Conclusions

A procedure for generating synthetic annual and monthly streamflow series was developed and tested, integrating a probabilistic generation model at the annual level, and a disaggregation model – the method of fragments – at the monthly level. The procedure was applied to 54 streamflow samples from gauging stations geographically spread over Mainland Portugal. A total of 22440 monthly flows were collected, analyzed and utilized to generate nearly 27 million monthly streamflows.

The probabilistic model used for generating annual streamflows, which is based on a random sampling of the log-Pearson III law of probability, proved to be robust. The application of the model resulted in the preservation of the statistics of the historical samples, although those samples exhibit quite different statistical characteristics among them. Since the statistical model generated logarithms of annual flows, the occurrence of synthetic negative streamflows is naturally avoided.

The method of fragments also proved to be appropriate to disaggregate annual streamflows into monthly streamflows. The procedure developed to automatically define the classes of fragments, which proved to lead to good results, completely eliminates any uncertainty regarding the establishment of those classes, thus increasing the robustness and generality of the disaggregation model.

The results at both temporal levels show that the proposed methodology for generating annual and monthly streamflow series accurately preserves the main statistical characteristics of the historical streamflow records in Portuguese rivers.

References

- [1] Quintela, A., *Recursos de águas superficiais em Portugal Continental*, Ph.D. thesis, Instituto Superior Técnico, Lisbon, Portugal, 1967.
- [2] Santos, E.G., *Disaggregation modeling of hydrologic time series*, Ph.D. thesis, Colorado State University, Colorado, USA, 1983.
- [3] Guimarães, R.C.P.d.C., *Simulação no dimensionamento e gestão de sistemas de recursos hídricos. Geração de séries sintéticas de escoamento*, Ph.D. thesis, Universidade de Évora, Évora, Portugal, 2005.
- [4] Chow, V.T., Maidment, D. & Mays, L., *Applied Hydrology*, McGraw-Hill International Publications, Singapore, 1988.
- [5] Arsénio, J.J.G., *Gestão de Albufeiras: avaliação do potencial dos sistemas com recurso a séries sintéticas de escoamentos*, Master's thesis, Instituto Superior Técnico, Lisbon, Portugal, 2003.
- [6] Portela, M.M., Quintela, A.C., Evaluation of the water resources in Portuguese watersheds without streamflow data. *Conferencia Internacional de Organismos de Cuenca (International Conference of Basin Organizations)*, Madrid, Spain, 2002.



- [7] Salas, J.D., Delleur, J.W., Yevjevich, V. & Lane, W.L., *Applied modeling of hydrologic time series*, Water Resources Publications, Colorado, USA, 1980.

