Open Water Journal

Volume 4
Issue 1 *Open Water Science and Software*

Article 7

2017

R Code for Data Simulation with Moment Matching

Varvara Vetrova
University of Canterbury, New Zealand, vetrova.varvara.v@gmail.com

William E. Bardsley
University of Waikato, web@waikato.ac.nz

OPEN WATER

R Code for Data Simulation with Moment Matching

Software Introduction

Varvara Vetrova[1*], Earl Bardsley[2]

[1]*School of Mathematics & Statistics, University of Canterbury, New Zealand*
[2]*Faculty of Sciences & Engineering, University of Waikato, New Zealand*
[*]*Corresponding Author: vetrova.varvara.v@gmail.com*

ABSTRACT

Simulation from data may be carried out for various purposes in water resources and is most simply achieved when the recorded data set can be regarded as a sequence of independent random variables. The aim is then to simulate a data set from a proabability distribution with similar statistical properties to the recorded data. In a recent paper it was shown that a finite mixture of a large number of generalised beta distributions can be utilised to simulate univariate data which has the same mean, variance, skewness, and kurtosis as any recorded data (Bardsley, 2017). Attention is drawn here to an R implementation of the simulation procedure of that paper. The code is reasonably efficent and generating a million data values requires about one minute on a standard PC.

Keywords
data simulation, finite mixture, moment matching, R code

1. Introduction

This brief communication draws attention to an R implementation of a procedure for creating a simulated data set with similar statistical moment properties to an existing data set (Bardsley, 2017). As was noted, water resource simulations from data may require complex algorithms to allow for varying forms of temporal correlation in time series data. However, it can sometimes happen that an observed time series may be regarded as a sequence of independent random variables from some unknown probability distribution. Simulating from such data can be carried out easily if the data histogram is well matched by some standard parametric distribution. This standard method is restricted, however, because if the data has some degree of irregularity then sumulations from a fitted distribution will be reflective of that distribution and not the data.

As one approach to this problem, Bardsley (2017) proposed that simulations be carried out using a finite mixture distribution comprised of a large number of beta distributions, with parameters specified such that the simulated values would always have the same true mean, variance, skewness, and kurtosis as the original data. At the same time as matching moments, the flexible nature of the mixture distribution allows irregular data to be better approximated than by using a single parametric distribution, or a mixture of a small number of parametric distributions.

An efficient Matlab implementation of the simulation algorithm is available from Bardsley (2017), with a million simulations noted as taking typically less than 10 seconds. For those without Matlab access this paper makes available a corresponding R implementation.

2.0 Algorithm

The algorithm is a translation to R of the Matlab code of Bardsley (2017). The R code here is not as fast as the Matlab equivalent but is still reasonably fast, with a million simulations taking around 1 minute to complete. Some further information is given as comments on the code. An aspect of the data simulation is that the simulations are random variables from beta distributions with differing parameters. There is potential for numerical instability for certain combinations of the two beta shape parameters. The same safeguards were used in the R code as for the Matlab code, with checks indicating that the implemented R beta random number generator is at least as numerically stable as the Matlab equivalent. See Bardsley (2017) for the applicability of beta distribution simulation over the beta shape parameter range.

3.0 Example

By way of example application of the R code, Fig. 1a shows the results of a million simulations from the same New Zealand autumn lake inflow data that was used for the Matlab simulations of Fig. 6 of Bardsley (2017). As expected, with such a large number of simulations there is little difference between the two simulation histograms (Fig. 1b). It may be possible to modify the R code with a view to increasing computational speed. However, the current capability of generating a million simulations in about a minute on a standard PC is probably acceptable for most purposes.

Software Availability

Name: SimData, Version number 0.1, Developer: V.Vetrova, Date first available: 23/01/206, How to retrieve it:

https://github.com/vetrovav/Data-simulation

Reference

Bardsley, W.E. (2017), A finite mixture approach to univariate data simulation with moment matching, Environ. Model. Softw., 90, 27–33.

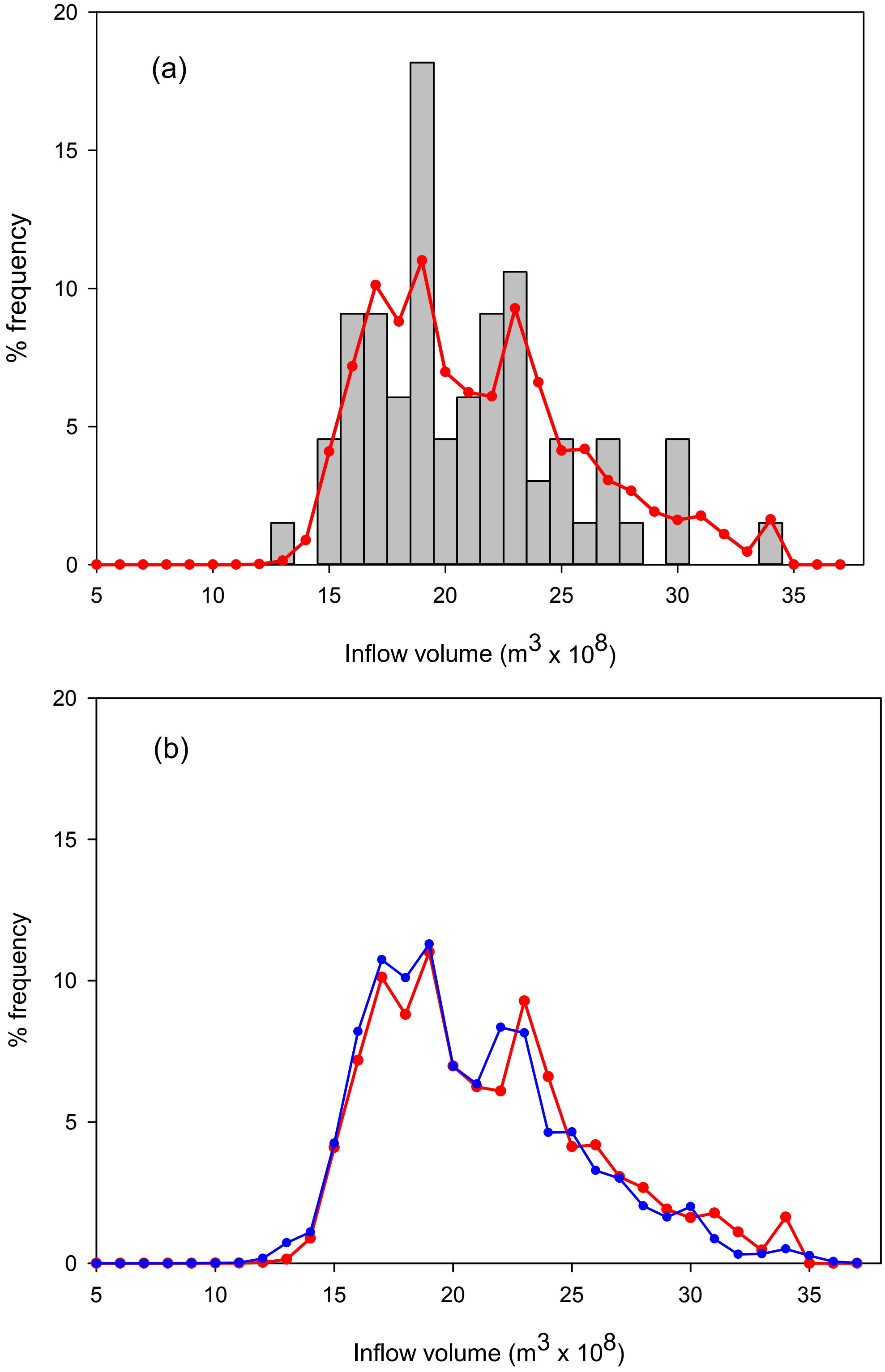

Figure 1a,b. (a) Data histogram and % frequency (red) from a million simulations carried out with the R code. (b) replot of the R code simulations and a million simulations from Matlab code. Data and Matlab simulations from Bardsley (2017).