Civil and Environmental Science Journal
Vol. I, No. 01, pp. 027-033, 2018

Data generation in order to replace lost flow data using Bootstrap method and regression analysis

Gatot Eko Susilo[1]

[1]Civil Engineering Dept., Universitas Lampung, Bandar Lampung, 35145, Indonesia

gatot89@yahoo.ca

Received 28-02-2018; revised 23-03-2018; accepted 06-04-2018

Abstract. This paper aims to find method to generate data in order to replace lost flow data in the series of discharge data in Sungai Seputih River, Lampung Province. Bootstrap simulation is used to estimate the discharge data and complete the existing discharge data. Regression analysis is also used to find the pattern of data distribution. Results of the research show that both methods are able to generate new series of flow data that the distribution is similar to available field data. Results also show that the use of statistical methods is one way to tackle the problem of data limitations due to missing or unrecorded data. The weakness of data generation using a combination of Bootstrap methods and regression analysis is the disappearance of extreme values in the data series. Existing extreme values have been modified to ideal values that satisfy certain distributions. However, careful analysis is required in using statistical method, so that the results of analysis do not deviate from the field conditions.

Keywords: data generation, flow data, Bootstrap method, regression analysis.

1. Introduction

Hydrology is the study of the earth's water sundry which includes the process of its occurrence, its movement, its distribution, and its relation to the environment and the living creatures. Understanding of the science of hydrology is very useful in understanding the concept of water balance on a global scale on the surface of the earth. Hydrological events such as rain and flow are recorded in the information referred to as hydrological data. Almost all water resources development activities require hydrological information for basic planning and design. If the hydrological information used is not suitable and does not meet the requirements, it may result in incorrect and inaccurate planning and design. Interpretation of the hydrological phenomenon will be carried out properly if supported by sufficient data availability. Sufficient data collection tools and consistent data collection activities are essential for generating good hydrological data.

The most important hydrological data is the flow data of a river. Basically, all water resource planning requires flow data in the calculation. But since the flow recorder stations in rivers in Indonesia are not always available then the amount of discharge can be calculated by varying the rain to discharge. Rainfall data is more available in watersheds in Indonesia. The rainfall recorder station is easier to find than the flow recorder station. This is because rain stations are not only installed by the department of public work but are also installed by department of agriculture and department of transportation with various objectives.

Various methods have been created by people to diversify rain into debit. But the accuracy of each method is still being debated. The calibration and verification process of a hydrological model is an absolute process to determine the validity of a model or method. The problem is data for calibration and verification is rarely available in Indonesia. Automation of flow recorders and rain gauges in Indonesia has not been equally distributed in Indonesia. Consequently, most of the debit or rainfall data in Indonesia are not valid enough data to be used in water resource planning. As an effort to validate the data, the planners perform validation analysis with various statistical methods.

The problem of scarcity of hydrological data has been overwhelmingly faced by water resource planners. Some of them attempt to generate data to add or supplement lost data. One of the known data generation methods is Monte Carlo Simulation. Monte Carlo simulation is a simulation to determine a random number of sample data with a particular distribution. The goal of Monte Carlo simulation is to find a value close to the real value, or the value that will occur based on the distribution of the sampling data [1]. The Monte Carlo simulation involves the use of random numbers to model the system, where time does not play a substantive. Monte Carlo simulation is undertaken by artificial data generation using pseudo random numbers generator. Basically, a Monte Carlo simulation is performed based on a particular sampling distribution. The key is to identify the distribution of existing sample data. Randomly, simulations of numbers are performed so that a combination of near-fit distribution is most fit. Monte Carlo simulations have been used to generate rainfall data in stochastic hydrological modelling studies in Czech Republic [2]. This simulation has also been used as a method to quantify drainage discharge components in a stochastic drainage discharge model in South Africa [3].

In addition to the Monte Carlo Simulation method, people often use the Bootstrap method. The bootstrap method is a method used to estimate the parameters of a population suspected of the statistical value obtained from the population sample. This method is often used because it does not base on certain distribution assumptions. Bootstrap is a method that can work without the need for distribution assumptions because the original sample is used as a population [4]. Bootstrap was first introduced by Efron in 1979. Bootstrap is a method based on data simulation for statistical inference purposes [5]. The Bootstrap method is performed by random sampling with a re-sampling with replacement. Some sources state that the bootstrap sample size (d) used is less than or greater than the sample data (n). However, the most optimum and effective way to guess parameters is the size of the bootstrap instance equal to the size of the sample data. The bootstrap method is a method based on re-sampling the sample data with the condition of return on the data in completing the statistics of the size of a sample in the hope that the sample represents the actual population data. Usually re-sampling size is taken thousands of times to represent the population data. This method is great for relatively small sample data sizes [6]. Bootstrap method has been used for quantifying uncertainty on sediment loads in Germany [7]. Previously, the method was also used in estimating the uncertainties related to the sample size in research of estimation of future discharge of the Rhine River [8]. The newest one, in China Bootstrap method was used to analyze the Influence of Rainfall spatial uncertainty on hydrological simulations [9].

This paper aims to generate data in order to replace lost flow data in the series of discharge data in Sungai Seputih River, Lampung Province. The corresponding flow data will be used to calculate the availability of water in the Way Seputih River for irrigation purposes in the Seputih Irrigation Area. Due to the lack of data, the Bootstrap simulation will be used to estimate the discharge data and complete the existing discharge data. Regression analysis is also used to find the pattern of data distribution.

2. Material and Methods

In this research, the Bootstrap method will be used to supplement lost discharge data in order to calculate the dependable flow to be used in calculating the allocation of irrigation water in the Pengubuan River. The river is located in the Central Lampung regions, Indonesia. The River is currently supplied water for irrigation areas which is Way Pengubuan Irrigation Area. The irrigation

area captured 5,000 ha and 3,500 ha as potential and functional paddy field, respectively. The calculation of irrigation water allocation is an activity to calculate the balance between water availability and water requirement in the irrigation area. In order to calculate irrigation water availability, a dependable flow is calculated with 80% reliability. For the calculation, the daily average discharge data in monthly period for 10 years have to be available. In fact, the daily average discharge data at Way Pengubuan Dam is only available for year 2011 until 2017. The available data is also not a complete data because there are some missing or undocumented data. The available discharge data view can be seen in Table 1.

Table 1. Existing flow data (in m^3/s) of Way Pengubuan River at Way Pengubuan Dam.

Year/Month	2011	2012	2013	2014	2015	2016	2017
Jan	-	4.85	4.03	5.23	3.34	3.21	4.58
Feb	-	5.04	5.06	4.36	5.11	4.35	3.77
Mar	-	3.86	5.05	3.16	4.27	5.69	5.40
Apr	-	4.39	4.99	4.41	3.67	5.87	4.76
May	-	3.17	5.41	4.96	3.92	5.37	-
Jun	-	-	-	-	-	-	-
Jul	-	1.82	3.00	2.45	0.83	2.11	-
Aug	-	-	-	-	-	-	-
Sep	-	-	-	-	-	-	-
Oct	-	-	-	-	-	-	-
Nov	-	-	-	-	-	-	-
Dec	3.60	3.46	3.78	4.15	1.13	4.69	-

Source: BBWS Mesuji Sekampung (2018)

To calculate the dependable flow required data flow with a data length of at least 10 years. To complete the missing data then the Bootstrap method is taken to generate the data. The data generation procedure with Bootstrap method is implemented as follows:

- Calculating the maximum and minimum values of monthly data for every year of data. This procedure will result in the maximum and minimum value of January, February, March, April, May, July, and December data for year 2011 to 2017.
- Generating data using Bootstrap method and complete flow data for the period of January, February, March, April, May, July, and December data for year 2011, 2017, 2018, 2019, and 2020.
- Generating data using Bootstrap method and complete flow data for the period of June for year 2011 to 2020. Value of June is random value between May and July values.
- To complete flow data of September, October, and November for year 2011 to 2020 then regression analysis is undertaken. Regression equation is formed for all data year and after that the shape of regression curves are modified to find ideal equation for the data of each year. Final equation of the regression then is used to generate new serial data.

Once the new serial of data is found, dependable flow with 80% reliability is calculated using Weibull probability equation [10]. The probability of each daily average monthly data is calculated using following formula:

$$P = \frac{m}{n+1} \quad (1)$$

where, P is the probability of data that explain the reliability percentage of data, m is the number of data after sorted from maximum to minimum value, and n is number of data.

3. Result and Discussion

Result of procedure 1 is given in Table 2. The result shows the maximum and minimum value of January, February, March, April, May, July, and December data for year 2011 to 2017.

Table 2. Maximum and minimum flow data (in m^3/s) of Way Pengubuan River

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Max	5.23	5.11	5.69	5.87	5.41	-	3.00	-	-	-	-	4.69
Min	3.21	3.77	3.16	3.67	3.17	-	0.83	-	-	-	-	1.13

Source: Calculation

Bootstrap method is used to generate data and complete flow data for the period of January, February, March, April, May, July, and December data for year 2011, 2017, 2018, 2019, and 2020. The value of particular month is actually random data between maximum and minimum value of corresponding month. Another generating data process using Bootstrap method is undertaken and complete flow data for the period of June for year 2011 to 2020. Value of June is random value between May and July values. The results are given as follows:

Table 3. Result of data generation using Bootstrap method

Year/Month	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Jan	4.16	4.85	4.03	5.23	3.34	3.21	4.58	4.70	3.65	4.00
Feb	4.75	5.04	5.06	4.36	5.11	4.35	3.77	5.29	4.26	3.66
Mar	5.42	3.86	5.05	3.16	4.27	5.69	5.40	5.17	4.33	3.61
Apr	5.34	4.39	4.99	4.41	3.67	5.87	4.76	4.51	3.99	3.82
May	4.31	3.17	5.41	4.96	3.92	5.37	5.05	3.50	3.39	4.10
Jun	3.50	2.97	3.95	4.66	1.08	4.34	3.27	2.34	2.66	4.26
Jul	5.00	3.64	6.00	4.89	1.66	4.22	2.50	1.20	1.94	4.18
Aug	-	-	-	-	0.89	2.81	-	0.27	1.37	3.90
Sep	-	-	-	-	-	-	-	0.00	1.08	3.69
Oct	-	-	-	-	-	-	-	0.00	1.22	3.80
Nov	-	-	-	-	-	2.76	-	0.62	1.92	4.17
Dec	3.60	3.46	3.78	4.15	1.13	4.69	3.60	2.40	3.32	4.13

Source: Calculation

Regression analysis is undertaken to form ideal shape of data distribution. The example of regression curve formed by regression analysis is given for year 2011 in Figure 1. The curve of actual data is modified into new curve formed by regression analysis. Using the equation of the new regression analysis, a serial of new data is generated. This serial data is finally used as serial data for dependable flow calculation.

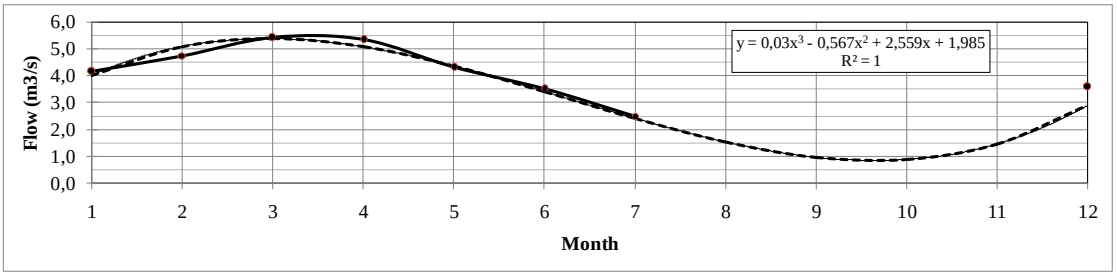

Figure 1. Curve of actual data (black line and black dot) and regression curve formed by regression analysis (dashed line) for year 2011 data

Using same procedures regression curve formed by regression for each year is given as follows:

Table 4. Regression equations formed by modified data

Year	Regression equation formed
2011	$y = 0.03x^3 - 0.567x^2 + 2.559x + 1.985$
2012	$y = 0.015x^3 - 0.239x^2 + 0.579x + 4.483$
2013	$y = 0.025x^3 - 0.484x^2 + 2.317x + 2.097$
2014	$y = 0.011x^3 - 0.181x^2 + 0.4x + 4.784$
2015	$y = 0.035x^3 - 0.646x^2 + 2.666x + 1.494$
2016	$y = 0.046x^3 - 0.893x^2 + 4.465x - 0.786$
2017	$y = 0.025x^3 - 0.464x^2 + 2.053x + 2.42$
2018	$y = 0.031x^3 - 0.545x^2 + 2.009x + 3.206$
2019	$y = 0.023x^3 - 0.409x^2 + 1.676x + 2.36$
2020	$y = 0.022x^3 - 0.411x^2 + 1.66x + 2.343$

Source: Calculation

Using equations above the final generated data is given in Table 5. Dependable flow is undertaken by sorting daily average monthly flow data. Weibull probability equation is used to calculate the value of reliability for each month. The results of reliability calculation Weibull probability equation are given in Table 5 below.

Table 5. Serial flow data based on regression curve

m	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	P
1	5.01	5.29	5.81	5.73	4.96	4.26	4.18	3.90	3.69	3.80	4.17	4.13	0.09
2	4.84	5.08	5.37	5.22	4.71	3.98	3.18	2.46	1.97	1.87	2.30	3.69	0.18
3	4.70	5.00	5.37	5.07	4.36	3.79	2.63	2.03	1.74	1.68	2.13	3.44	0.27
4	4.03	4.95	5.17	4.81	4.21	3.43	2.49	1.95	1.54	1.55	1.92	3.41	0.36
5	4.01	4.94	5.08	4.51	3.69	3.41	2.49	1.53	1.27	1.37	1.92	3.32	0.45
6	3.81	4.87	4.65	4.19	3.63	3.04	2.41	1.50	1.08	1.22	1.90	2.94	0.55
7	3.96	4.81	4.62	4.06	3.50	2.66	1.97	1.37	0.96	0.88	1.50	2.89	0.64
8	3.65	4.52	4.47	3.99	3.39	2.59	1.94	1.33	0.60	0.56	1.46	2.53	0.73
9	3.55	4.26	4.33	3.94	3.28	2.34	1.20	0.27	0.00	0.00	0.62	2.40	0.82
10	2.83	3.66	3.61	3.85	3.05	1.79	0.51	0.00	0.00	0.00	0.00	0.94	0.91

Source: Calculation

Using interpolation technique dependable flow with 80% reliability for each month are presented in Figure 2. The resulting dependable flow above illustrates the pattern of rainy and dry seasons occurring in the study area. Therefore, it can be concluded that the data can be statistically used as a material calculation of water allocation in the area concerned. For irrigation purposes, usually the dependable flow value is calculated based on a period of 15 days. To calculate the dependable flow with a period of 15 days, it can be done by taking two random numbers whose average is the value in the corresponding month. For example, to calculate the value of dependable flow in the period January I and January II, we have to take two random numbers where the average of the two random numbers is the January dependable flow value.

The weakness of data generation with a combination of Bootstrap methods and regression analysis is the diminished extreme value of a distribution. Existing extreme values have been modified to ideal values that satisfy certain distributions. The influence of climate anomalies such as El Nino must be closely watched because the minimum extreme values sometimes appear this year. Minimal extreme values will affect the dependable flow calculation. Sometimes a minus number appears in the

generated data. If the minus value appears in serial data then the value must be replaced with the value of zero because logically there is no minus flow value.

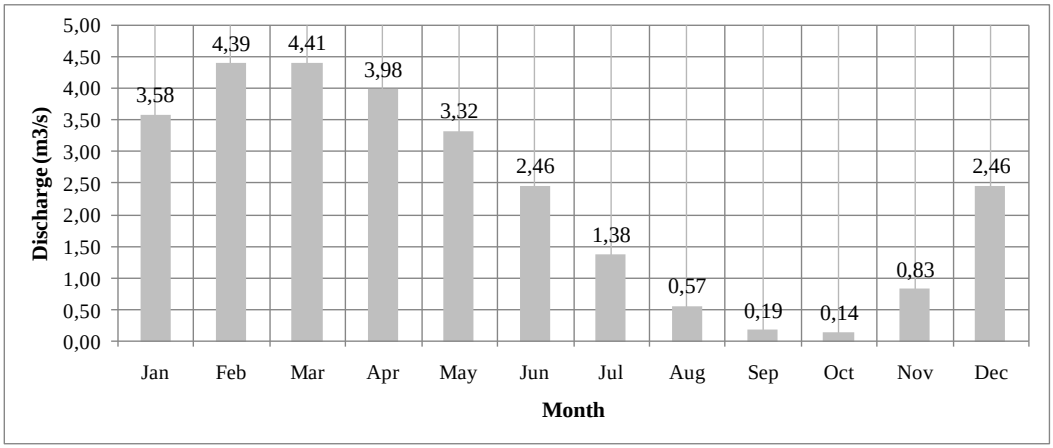

Figure 2. Dependable flow with 80% reliability of Pengubuan River

Basically, the best way to calculate dependable flow is to collect as much historical data as possible. But to get a lot of data flow and complete is not easy in Indonesia. Therefore, statistical analysis is the best way to do it. Statistical analysis is probably the best way to process a small amount of data. But statistical analysis should be accompanied by empirical analysis to test the accuracy of the preceding analysis.

4. Conclusions

Data generation in order to replace lost flow data in the series of discharge data in Sungai Seputih River, Lampung Province has been analyzed. The results show that the combination of Bootstrap method and regression analysis is able to generate new data whose distribution is similar to available field data. However, complete data is a key requirement in a water resource plan. Statistical methods can be taken to solve the problem of data availability. However, careful analysis is required in using statistical methods so that the results of analysis do not deviate from the field conditions.

Acknowledgements

The author would like to express his deep gratitude to Mrs. Eka Desmawati and Mr. Ankavisi Nalaralagi from BBWS Mesuji Sekampung for their support of this research especially in providing hydrological data.

References

[1] Huang, H. 2018. Monte Carlo simulation using Excell (In Bahasa Indonesia), Globalstats Academic Publication. http://www.globalstatistik.com/simulasi-monte-carlo-dengan-excel/. March 19th (19:05).
[2] Březková, L., Starý, S., and Doležal, P. The Real-time Stochastic Flow Forecast. *Soil & Water Res.*, 5(2): 49–57.
[3] Flores, G. 2015. A stochastic model for sewer base flows using Monte Carlo simulation. *Master Thesis*, Stellenbosch University, South Africa.
[4] Sungkono, J. 2015. Bootstrap re-sampling observation on estimation parameter regression using software R (In Bahasa Indonesia). *Magistra* No. 92 Year XXVII.
[5] Efron, B. and Tibshirani, R. J. 1993, *An introduction to the Bootstrap*, Chapman and Hall, New York.
[6] Monalisa, A. 2016. Use of Bootstrap resampling method for simulation data time series model using ARIMA, *Undergraduate thesis* (In Bahasa Indonesia), University of Jember.

[7] Slaets, J. I. F., Piepho, H., Schmitter, P., Hilger, T. and Cadisch, G. 2017. Quantifying uncertainty on sediment loads using Bootstrap confidence intervals. *Hydrol. Earth Syst. Sci*., 21: 571–588.
[8] Lenderink, G., Buishand, A. and Deursen, W. 2007. Estimates of future discharge of the river Rhine using two scenario methodologies: direct versus delta approach. *Hydrol. Earth Syst. Sci*., 11(3): 1145–1159.
[9] Zhang, A., Shi, H., Li, T. and Fu, X. 2018. Analysis of the influence of rainfall spatial uncertainty on hydrological simulations using the Bootstrap method. *Atmosphere* 9(71): 2–24.
[10] Weibull, W. 1951. A statistical distribution function of wide applicability, *J. Appl. Mech.-Trans. ASME*, 18(3): 293–297.