ОЦЕНКА ВЛИЯНИЯ ОТСУТСТВИЯ ДАННЫХ НА РАСЧЕТ КРИВЫХ ОБЕСПЕЧЕННОСТИ

А.Т. Горшкова, *канд. геогр. наук, заведующая лабораторией*
Д.А. Семанов, *канд. хим. наук, науч. сотр.*
О.Н. Урбанова, *ст. науч. сотр.*
Академия наук Республики Татарстан
(Россия, г. Казань)

Аннотация. *В нормативных документах, используемых для расчета обеспеченных значений гидрологических характеристик, приводятся только основные методы и схемы их определения без учета ошибок вычисления, особенно в части максимальных и минимальных значений. Предлагается упрощённый подход, достаточный не только для оценки влияния доли исключаемых данных и количества подвыборок на разброс параметров распределения, но и обоснования предположения о вариации в параметрах при появлении новых данных. В работе сделана оценка параметров статистического распределения гидрологических характеристик и определена величина погрешности этой оценки.*

Ключевые слова: *гидрологические характеристик, выборка, оценка погрешностей, параметры распределения*

Введение

Гидрологическим обоснованием всех проектных решений, связанных с водохозяйственным строительством, являются статистически обеспеченные значения лимитирующих гидрологических характеристик, для получения которых по фактическим рядам наблюдений строятся эмпирические кривые обеспеченности, аппроксимируемые теоретическим распределением кривых Пирсона III типа или трехпараметрическим распределением Крицкого-Менкеля. Подобный подход закреплен в периодически обновляемых и используемых в настоящее время нормативных документах - СН435-72, СП33-101-2003. Однако полного совпадения теоретических и эмпирических кривых достичь невозможно, так как ряд наблюденных гидрологических величин всегда будет несколько отличаться от теоретической вероятности.

В указанных сводах правил приводятся только основные методы и схемы определения расчетных гидрологических характеристик, но не указаны способы оценки погрешности вычислений, а использование подходов с тремя разными типами распределений не позволяют адекватно экстраполировать сами значения обеспеченностей до 1% и свыше 99%, тем более оценивать ошибку в их определении.

Подходы с проверкой адекватности распределений (соответствие фактическим статистическим данным), предложенные Христофоровым А.В. только осложняют ситуацию, существенно расширяя диапазоны параметров распределений, проходящих не слишком чувствительные статистические методы проверки их качества [1].

Для расчета кривых обеспеченности и определения с их помощью обеспеченных значений гидрологической величины (чаще всего это уровни или расходы воды) необходимо, чтобы длина ряда (продолжительность периода) наблюдений была достаточной для расчета и средняя квадратичная ошибка расчетного значения не превышает 10% для годового стока и 20% для максимального и минимального.

В настоящее время ряды гидрологических наблюдений имеют часто менее 20 значений, реже 30-50 и очень редко 80-100. Выбирая ряд значений за весь имеющийся период наблюдений, невозможно, однако, считать его в достаточной мере полным для того, чтобы непосредственно по нему устанавливать величины стока редкой повторяемости. В охваченном периоде наблюдений возможны пропуски в течение нескольких лет, когда измерения не проводились по ряду различных при-

чин. В тех случаях, когда длина имеющегося ряда данных недостаточна для определения параметров распределения гидрологической характеристики, производят его удлинение с использованием достаточно тесной корреляционной связи данной характеристики расчетного бассейна и бассейна-аналога с более длительным периодом наблюдений [2].

Несмотря на то, что наблюдаемые гидрологические характеристики (годовой сток и другие его однородные характеристики) отдельных лет не связаны между собой, что в известной мере подтверждается данными наблюдений и специальными исследованиями, все их значения расположенные в хронологическом порядке, представляют собой статистический ряд. Для построения кривой обеспеченности по такому ряду и ее экстраполяции за пределы наблюдений члены ряда располагаются в убывающем порядке. Такой ранжированный (вариационный) ряд значений за ограниченный период наблюдений рассматривается как выборка (часть) некоторой случайной величины, функция распределения вероятностей которой подлежит статистическому распределению. При этом устанавливается связь между возможными значениями гидрологической характеристики и ее повторяемостью.

Для статистического оценивания параметров в выборке в настоящее время разработано большое количество способов, среди которых наиболее употребительными в отечественной гидрологии являются метод моментов (для оценки не более четырех параметров), метод максимального (наибольшего) правдоподобия (особенно полезен при малых выборках) и метод квантилей (схожий с методом моментов). Универсального ответа на вопрос, какой из рассмотренных методов лучше и какой из них следует применять для решения гидрологических задач, нет. Вычисленные этими методами параметры гидрологических характеристик несколько различаются между собой. Оценить отличие параметров трудно, так как в каждом конкретном случае, вычисленный параметр отличается от истинного значения на неизвестную величину. Иначе говоря, существует некоторая доля неопределенности в знании действительного значения параметра [3].

Оптимальными методами могли бы оказаться оценки с использованием генерации большого количества подвыборок из имеющихся данных с определением параметров распределения таких подвыборок и обеспеченностей гидрологических характеристик для каждой подвыборки с последующим усреднением и определением разброса значений получившихся величин. Для проверки возможности такого подхода допустимо использовать любую из рекомендуемых функций распределения. Правомерность такого подхода основывается на ограниченности исходных данных, несмотря на то, что имеющийся ряд данных может быть либо продлен, либо дополнен [4, 5].

По этой причине, любая подвыборка из имеющихся данных столь же «правомерна» для получения параметров распределения. Исключение некоторого случайного набора данных позволяет проверить насколько отсутствие определенной части значений повлияет на изменение параметров распределения. Это поможет сделать обоснованные предположения о вариации в параметрах при появлении новых данных.

Авторы на основании большого количества подвыборок оценили не только параметры статистического распределения гидрологических характеристик, но и определили величину погрешности этой оценки, что и явилось целью данного исследования.

Материалы и методы исследования

Исходными данными для получения параметров статистического распределения явились годовые и меженные (август) расходы воды р.Свияга в створах гидрологических наблюдений, расположенных у сел Ивашевка и Вырыпаевка, а также р. Казанка у г. Арска. В данном исследовании более подробно изложены результаты обработки и анализа исходных данных по посту Свияга-Ивашевка, выполненные с помощью программы, написанной на языке python (версия 2.72).

В качестве функции распределения гидрологических характеристик было принято трёхпараметрическое гамма-распределение

$$\varphi(x) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} a^{\alpha} b x^{b\alpha-1} e^{-\alpha a x^{b}}$$

[6,7], а расчёт эмпирической вероятности проводился по формуле Алексеева

P = (m-0.25) / (n+0.5) * 100%, рекомендуемые СП 33-101-2003 [8].

Подбор функции распределения осуществлялся оценкой параметра «a» с помощью коэффициента вариации «Cv» с последующим подбором параметров «α» и «b» вычислением линейной регрессии в логарифмических координатах. Расчёты проводились для выборок, генерируемых из исходного набора данных, путём исключения случайным образом двух и более значений. Количество подвыборок составило 50, 200, 500 и 2000. Вычислялись параметры распределения и обеспеченности 1%, 5%, 50%, 75%, 95%, 99%.

Конечные параметры обеспеченности определялись как взвешенные средние. В качестве весового коэффициента использовалось обратное значение квадрата отклонения линейной регрессии при расчёте параметров «α» и «b». Ошибки оценок параметров определялись упрощённо, как если бы статистики соответствовали нормальному распределению с 95% вероятностью, за исключением 99% и 1% обеспеченности, для которых диапазон ошибки в оценке среднего захватывает 99% данных. Упрощённый подход без учёта формы распределения получившихся статистик был посчитан достаточным для оценки влияния доли исключаемых данных и количества подвыборок на разброс параметров распределения.

Результаты исследования и обсуждение

На основе значений годовых расходов воды, наблюдаемых у с. Ивашевка, расположенного на р. Свияга оценивалась степень влияния количества исключённых данных и количества подвыборок на результаты оценки параметров распределения и их погрешностей. Исходная выборка состояла из 54 значений годовых расходов воды. Проверялось исключение 2 (менее 4% данных) и 5 (более 9%) значений. Результаты расчетов представлены в таблице 1.

Таблица 1. Степень влияния количества исключенных данных на оценку параметров распределения в подвыборках

Кол-во исключенных значений	Количество подвыборок	Параметры распределения
2	50	a=0.0617 ± 0.00052 b=0.896 ± 0.00276 alpha=9.74 ± 0.10 обеспеченность 99%=7.98 ± 0.071, P=0.99 обеспеченность 95%=11.1 ± 0.071 обеспеченность 75%=16.7 ± 0.064 обеспеченность 50%=21.6 ± 0.059 обеспеченность 5%=37.3 ± 0.12 обеспеченность 1%=45.7 ± 0.25, P=0.99
2	200	a=0.0617 ± 0.00029 b=0.897 ± 0.0014 alpha=9.71 ± 0.051 обеспеченность 99%=7.96 ± 0.032, P=0.99 обеспеченность 95%=11.1 ± 0.031 обеспеченность 75%=16.7 ± 0.027 обеспеченность 50%=21.5 ± 0.024 обеспеченность 5%=37.3 ± 0.064 обеспеченность 1%=45.6 ± 0.13, P=0.99
2	500	a=0.0614 ± 0.00021 b=0.898 ± 0.0011 alpha=9.81 ± 0.038 обеспеченность 50%=21.5 ± 0.018 обеспеченность 75%=16.7 ± 0.022 обеспеченность 5%=37.2 ± 0.047 обеспеченность 95%=11.2 ± 0.027 обеспеченность 99%=8.01 ± 0.028, P=0.99 обеспеченность 1%=45.4 ± 0.098, P=0.99
2	2000	a=0.0615 ± 9.9e-05 b=0.898 ± 0.00050 alpha=9.76 ± 0.017 обеспеченность 50%=21.5 ± 0.0088 обеспеченность 75%=16.7 ± 0.010 обеспеченность 5%=37.2 ± 0.021 обеспеченность 95%=11.1 ± 0.012 обеспеченность 99%=7.99 ± 0.012, P=0.99 обеспеченность 1%=45.5 ± 0.044, P=0.99
5	50	a=0.0618 ± 0.00079 b=0.896 ± 0.0040 alpha=9.73 ± 0.179 обеспеченность 50%=21.5 ± 0.073 обеспеченность 75%=16.7 ± 0.086 обеспеченность 5%=37.2 ± 0.20 обеспеченность 95%=11.1 ± 0.10 обеспеченность 99%=7.96 ± 0.105, P=0.99 обеспеченность 1%=45.6 ± 0.43, P=0.99
5	200	a=0.0621 ± 0.00053 b=0.896 ± 0.0027 alpha=9.86 ± 0.104 обеспеченность 99%=7.94 ± 0.057, P=0.99 обеспеченность 95%=11.1 ± 0.062 обеспеченность 75%=16.7 ± 0.054 обеспеченность 50%=21.5 ± 0.048 обеспеченность 5%=37.1 ± 0.13 обеспеченность 1%=45.4 ± 0.27, P=0.99

Кол-во исключенных значений	Количество подвыборок	Параметры распределения
5	500	a=0.0618 ± 0.00033 b=0.896 ± 0.0017 alpha=9.82 ± 0.056 обеспеченность 99%=7.98 ± 0.038, P=0.99 обеспеченность 95%=11.1 ± 0.037 обеспеченность 75%=16.7 ± 0.032 обеспеченность 50%=21.5 ± 0.028 обеспеченность 5%=37.2 ± 0.070 обеспеченность 1%=45.5 ± 0.15, P=0.99
5	2000	a=0.0617 ± 0.00017 b=0.897 ± 0.00084 alpha=9.84 ± 0.029 обеспеченность 99%=8.00 ± 0.020, P=0.99 обеспеченность 95%=11.2 ± 0.020 обеспеченность 75%=16.7 ± 0.017 обеспеченность 50%=21.5 ± 0.015 обеспеченность 5%=37.2 ± 0.035 обеспеченность 1%=45.5 ± 0.074, P=0.99

Анализ полученного результата показывает, что с увеличением количества подвыборок (200 и выше), оценка погрешностей существенно не изменяется. Для крайних обеспеченностей (1% и 95%) величина оценки погрешностей растет с увеличением количества отбрасываемых значений. Результаты оценок годовых и меженных (август) расходов воды для пунктов наблюдения на р. Свияга по постам Ивашевка и Вырыпаевка при 200 подвыборках и количестве исключенных значений 5 представлены в таблице 2.

Таблица 2. Результаты оценок годовых и меженных (август) расходов воды крайних обеспеченностей по постам Свияга-Ивашевка и Свияга-Вырыпаевка

Пункт наблюдения и период	Количество подвыборок (Кол-во исключенных значений)	Параметры распределения
Ивашевка, год	200 (5)	a=0.0621 ± 0.00053 b=0.896 ± 0.0027 alpha=9.86 ± 0.104 обеспеченность 95%=11.1 ± 0.062 обеспеченность 1%=45.4 ± 0.27, P=0.99
Ивашевка, август	200 (5)	a=0.112 ± 0.00073 b=0.946 ± 0.0028 alpha=3.021 ± 0.015 обеспеченность 95%=2.56 ± 0.017 обеспеченность 1%=30.2 ± 0.185, P=0.99
Вырыпаевка, год	200 (5)	a=0.128 ± 0.00066 b=0.934 ± 0.0024 alpha=7.36 ± 0.048 обеспеченность 95%=4.12 ± 0.018 обеспеченность 1%=19.6 ± 0.078, P=0.99
Вырыпаевка, август	200 (5)	a=0.205 ± 0.00083 b=0.972 ± 0.0019 alpha=7.54 ± 0.068 обеспеченность 95%=2.42 ± 0.0089 обеспеченность 1%=10.6 ± 0.046, P=0.99

Количество повторов, начиная с 200 подвыборок, практически не влияет на величину ошибки. Полученные данные позволяют предположить, что исключение 10 - 15% значений и количества подвыборок свыше 200 может быть достаточно для получения оценок параметров распределения.

На приведенных ниже графиках зависимости среднеквадратичного отклонения от количества отброшенных значений показаны примерно одинаковые оценки погрешностей при 200 и 500 подвыборках, что подтверждает предыдущий вывод (рисунки 1 и 2).

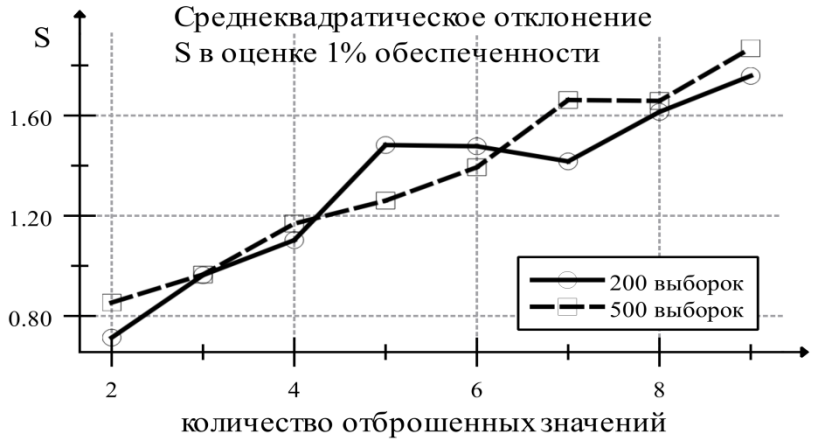

Рис. 1. Погрешность оценки в определении стока 1% обеспеченности в зависимости от исключенных значений

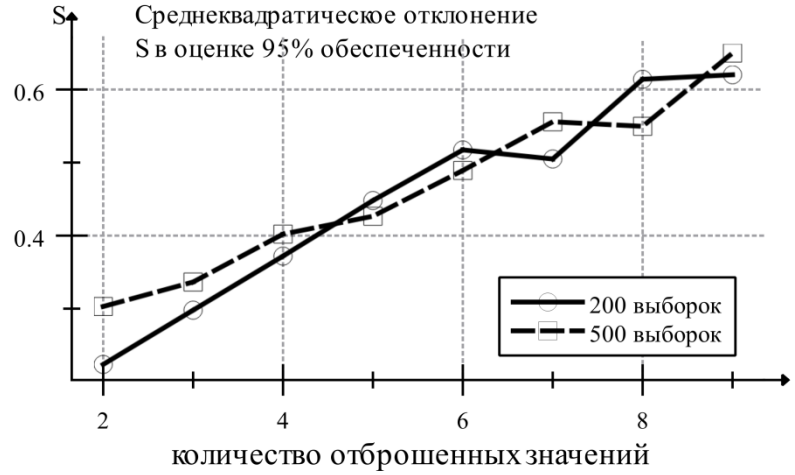

Рис. 2 Погрешность оценки в определении стока 95% обеспеченности в зависимости от исключенных значений

Кроме того, графики показывают наличие роста ошибки погрешности определения при уменьшении количества данных в подвыборках. Исключение более 9% значений (5 из 54 исходных значений) из ряда расчётной 95% обеспеченности для большого числа подвыборок (2000) показывает некоторую несимметричность встречаемости величин, и даже появление дополнительных пиков (рисунок 3).

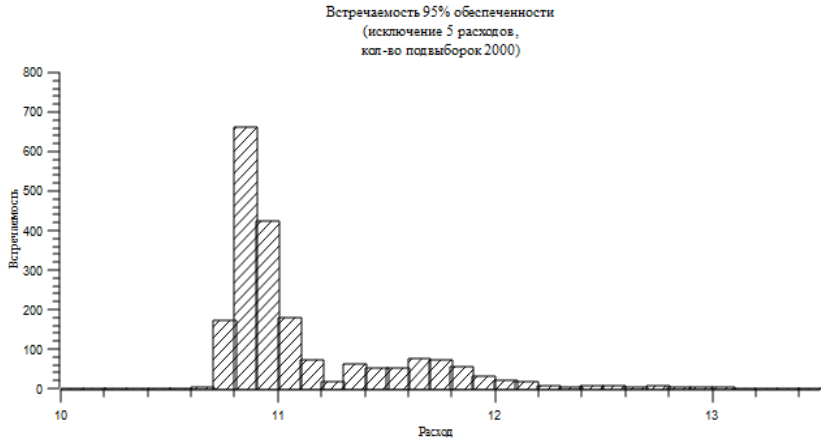

Рис. 3. Встречаемость 95% обеспеченности в количестве подвыборок 2000

Часто встречаемые значения гидрологических величин 95% обеспеченности (10.8-10.9) имеют меньшее значение, чем средняя величина (11.2). Наличие небольшого пика встречаемости при значениях (11.6-11.8) смещает среднее значение в большую сторону относительно медианы. Этот дополнительный пик встречаемости возник, по всей видимости, при исключении в части выборок данных с крайними значениями, редкими и потому существенно влияющими на результаты. Возможно, что наилучшей оценкой в данном случае будет не весовое среднее, а медиана или значение в районе максимальной вероятности.

Зависимость количества величин и их встречаемости для большого числа подвыборок (2000) при исключении более 5 значений (9% данных) из ряда расчётной 1% обеспеченности показана на рисунке 4.

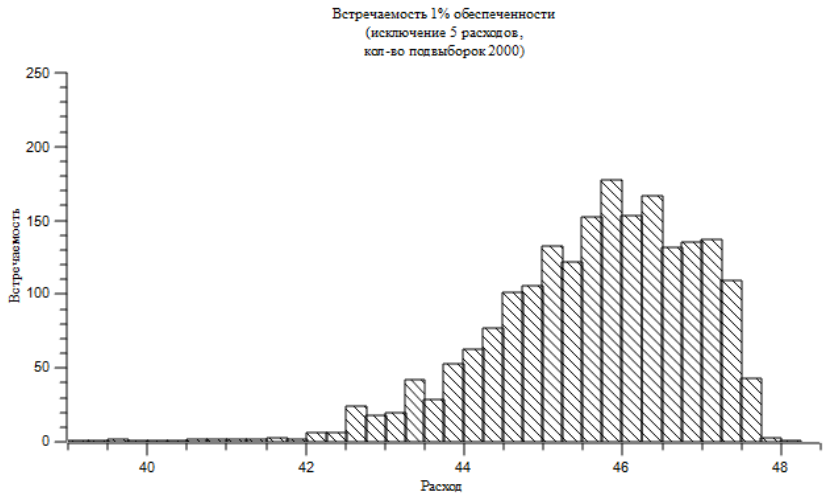

Рис. 4. Встречаемость 1% обеспеченности в количестве подвыборок 2000

Судя по графику для 1% обеспеченности медиана (45.8) немного больше, чем среднее (45.7), хотя разница в этом случае несущественна.

Заключение

Используемые в настоящее время методы и схемы определения обеспеченных гидрологических характеристик, рекомендуемые в нормативных документах, не позволяют адекватно экстраполировать значения обеспеченностей до 1% и свыше 99%, тем более оценивать ошибки в их определении.

Предлагаемый упрощённый подход (метод) определения параметров распределения гидрологических величин по принципу случайного исключения некоторой части данных и количества подвыбо-

рок, генерируемых из исходного набора данных, а также оценки их изменчивости показал свою эффективность. С одной стороны, метод не требует дополнения случайно генерируемыми данными, которые могут привнести свой, заданный параметрами генерации, вклад в оценку параметров. С другой стороны он не требователен к компьютерным ресурсам.

Результатом проведенного исследования явился расчет не только величин обеспеченного годового и меженного стока р. Свияга, но и оценки их разброса при изменении набора данных измерений. Разовые расчёты проводились с помощью программы, написанной на языке python (версия 2.72) на основе современных компьютеров, в том числе, на процессорах Intel Atom Z3740.

Библиографический список

1. *Христофоров А.В.* Оценка параметров распределения вероятностей величин речного стока/ Метеорология и Гидрология, 1981, №8. С.78-86.

2. Пособие по определению расчетных гидрологических характеристик. Л.: Гидрометеоиздат, 1984.- 447 с.

3. http://opds.sut.ru/old/electronic_manuals/oed/index.htm#r001.

4. Politis, D.N. and Romano, J.P. (1994). The stationary bootstrap. Journal of American Statistical Association, 89, 1303-1313.

5. *Efron, B.* (1982). The jackknife, the bootstrap, and other resampling plans. 38. Society of Industrial and Applied Mathematics CBMS-NSF Monographs.

6. *Крицкий С.Н., Менкель М.Ф.* Гидрологические основы речной гидротехники. Изд. АН СССР, 1950.

7. *Соколовский Д.Л.* Речной сток. Л.: Гидрометеоиздат, 1968. - 540 с.

8. *СП* 33-101-2003. Определение основных расчетных гидрологических характеристик. М.: Госстрой России, ФГУП ЦПП, 2004. - 73 с.

ESTIMATION OF INFLUENCE OF NULL DATA ON CALCULATION OF CURVES OF MATERIAL WELL-BEING

A.T. Gorshkova, *candidate of geographical sciences, head of laboratory*
D.A. Semanov, *candidate of chemistry, researcher*
O.N. Urbanova, *senior researcher*
Academy of sciences of the Republic of Tatarstan
(Russia, Kazan)

Abstract. *The norms for calculating the provided values of hydrological characteristics contain only the basic methods without taking into account inaccuracy, especially in the part of the maximum and minimum values. Today we propose a simple method for assessing the impact of data exclusion, the number of subsamples and the justification for the assumption of variation in parameters under the conditions of the appearance of new data. Based on the evaluation of the statistical distribution of hydrological characteristics, the inaccuracy is determined.*

Keywords: *characteristics of hydrology, design sample, inaccuracy, distribution parameters*