С.Я. Долинная, В.А. Шелутко

ВОПРОСЫ ПРИМЕНЕНИЯ МЕТОДОВ ЛИНЕАРИЗАЦИИ СВЯЗЕЙ И НОРМАЛИЗАЦИИ ИСХОДНЫХ РЯДОВ ПРИ РАСЧЕТАХ ПО УРАВНЕНИЯМ РЕГРЕССИИ

S.Ya. Dolinnaya, V.A. Shelutko

QUESTIONS ON THE USE OF LINEARIZATION AND NORMALIZATION METHODS OF ORIGINAL SERIES IN REGRESSION EQUATIONS' CALCULATING

Рассматриваются вопросы применения методов линеаризации и нормализации исходной информации при расчетах по уравнениям регрессии. Исследуются особенности этих методов в связи с граничными условиями регрессионной модели. Даются рекомендации по их учету этих особенностей для повышения эффективности расчетов.

Ключевые слова: уравнение регрессии, коэффициент корреляции, линеаризация, нормализация, граничные условия, парная корреляция, множественная корреляция.

Questions on the use of linearization and normalization methods of original series in regression equations' calculating are regarded in this article. The features of these methods are investigated in relation to the boundary conditions of the regression model. Recommendations are given to register these features to improve the efficiency of calculations.

Key words: regression equations, correlation coefficient, linearization, normalization, boundary conditions, pair correlation, multiple correlation.

При исследовании различных природных явлений, значения которых во времени или пространстве представляются в виде последовательности значений случайных величин, часто встречаются ситуации, когда одна случайная величина, допустим Y, каким-либо образом определена по отношению к одной или множеству случайных величин: $X_1, X_2, ..., X_m$. Определенность или взаимосвязь различных природных явлений используется для построения статистического математического описания всех явлений в виде регрессионной математической модели с независимыми нормально распределенными ошибками и равноточными измерениями. Эта модель во многих случаях позволяет провести достаточно полный статистический анализ исследуемого явления в его связи с внешней средой. Такой анализ является основой для физической интерпретации и практического использования математической модели, например, в целях восстановления пропусков в рядах наблюдений или прогнозирования.

К настоящему времени опубликовано значительное число работ, в которых парный или множественный регрессионный анализ применялся для идентификации математических моделей процессов стока. Наряду с работами, содержащими

положительный опыт применения регрессионного анализа, отмечались случаи, когда достичь желаемой точности математического описания не удавалось. Последнее обстоятельство послужило поводом для скептического отношения некоторых авторов к методам парной и множественной регрессии.

Между тем часто основная причина неудач заключается в механическом приложении схемы регрессионного анализа без учета специфических особенностей используемого математического аппарата. Эти особенности, сформулированные в виде граничных условий, были опубликованы еще в 1983 г. [1]. Они весьма существенны и их нельзя игнорировать без риска получить отрицательный результат.

В частности большое влияние на результативность оценки взаимосвязей природных процессов на основе регрессионной модели оказывает несоблюдение двух основных граничных условий:

– связь между сопоставляемыми рядами должна быть линейной;
– сопоставляемые ряды должны подчиняться нормальному закону распределения.

В практике исходные ряды часто не соответствуют этим условиям. В этих случаях для учета названных условий используются методы линеаризации и нормализации связей. Среди них наибольшее распространение получили метод логарифмирования исходной информации и метод нормализации и линеаризации связи Г.А. Алексеева.

Между тем результаты применения этих методов при расчетах по уравнениям парной или множественной корреляции нередко оказываются неэффективными, несмотря на то, что они в достаточной степени учитывают названные граничные условия.

Целью настоящей работы является выяснение причин недостаточной эффективности названных методов и разработка рекомендаций по ее повышению

Как известно [1, 3] логарифмирование заключается в преобразовании исходных данных в десятичные или натуральные логарифмы. Полученные в результате логарифмирования сопоставляемых рядов связи, в случае если эти связи являются монотонно убывающими или монотонно возрастающими, становятся линейными. Таким образом, учитывается первое граничное условие. При этом расчеты параметров уравнения регрессии производятся обычным образом, но по логарифмам исходных рядов. Естественно, что результаты расчетов тоже получаются в логарифмах и для того, чтобы перейти к значениям исходного ряда, необходимо выполнить потенцирование.

Метод линеаризации и нормализации Алексеева был предложен еще в 60-е гг. [2], но из-за трудоемкости расчетов, не получил особого распространения.

Полученные в результате применения этого метода связи, в случае если по исходным рядам они являются монотонно убывающими или монотонно возрастающими, также как и при логарифмировании становятся линейными. Кроме того, полученные в результате нормализации, ряды подчиняются нормальному закону распределения. Таким образом, в результате применения нормализации методом Алексеева учитываются оба рассматриваемых граничных условий

Основные этапы нормализации и линеаризации связи методов Г.А. Алексеева заключаются в следующем:

1. По каждому из рядов Y и X_j ($j = 1, 2, ..., m$) производится замена исходных значений их эмпирическими обеспеченностями, рассчитанными, например, по формуле[1]:

$$P_i = \frac{(i - 0,3)}{(n + 0,4)} \cdot 100, \quad (1)$$

где i – ранговые (порядковые) номера наблюденных значений исходного ряда Y или рядов X_j после их ранжирования в убывающем порядке. Эти эмпирические вероятности названы автором метода ранговыми переменными.

2. Ранговые переменные заменяются нормализованными переменными. Для этого по таблице значений ординат нормальной кривой обеспеченности по известным значениям P_i исходного ряда определяются нормированные ординаты t_i и рассчитываются соответствующие каждому значению исходного ряда Y и X нормализованные значения:

$$U_i = m_x + \sigma_x \cdot t_i, \quad (2)$$

где m_x и σ_x — математическое ожидание и среднее квадратическое отклонение преобразовываемого ряда.

3. По полученным нормализованным рядам, соответственно U и V_j ($j = 1, 2, ..., m$), определяются параметры уравнения регрессии нормализованных величин: коэффициенты веса a_j и сводный коэффициент корреляции R_0.
4. По нормализованным значениям рядов V_j рассчитываются по уравнению регрессии значения нормализованного ряда U:

$$\delta u_i = \Sigma a_j \cdot \delta v_{ji}, \quad (3)$$

где δu_i и δv_{ji} — отклонения от среднего значения соответственно по нормализованному ряду U и нормализованным рядам V_j.

5. Совершается переход от рассчитанных по уравнению регрессии нормализованных значений ряда Y к исходным значениям, Согласно предложения Г.А. Алексеева этот переход может быть совершен или путем обратного пересчета или по уравнению регрессии $Y = f(U)$.

   Обратный пересчет заключается в поэтапном определении по расчитанным нормализованным значениям ряда U ранговых переменных, определении по ранговым переменным и соответствующему закону распределения нормированных значений t_p и расчете по (2) значений исходного ряда Y.
6. Рассчитывается сводный коэффициент корреляции исходного ряда Y и ряда рассчитанного по уравнению регрессии Y_p и расчет сводного коэффициента

[1] Г.А. Алексеев использует другую формулу эмпирической обеспеченности [2]. В данном случае используется более распространенная формула эмпирической обеспеченности.

корреляции и критерий эффективности, как отношение дисперстии рассчитанного ряда к дисперсии исходного ряда.

При анализе эффективности рассмотренных методов было необходимо произвести довольно трудоемкие расчеты. Поэтому была разработана в среде Дельфи специальная программа, способная производить расчеты характеристик уравнений регрессии по любому заданному числу аргументов и их преобразованию.

Для иллюстрации результатов исследований в данном случае было отобрано 4 примера, разнообразных по характеру связи между сопоставляемыми рядами и эффективности рассматриваемых преобразований. В табл. 1 представлены результаты расчетов сводных коэффициентов корреляции R_0 по исходным, логарифмированным и нормализованным рядам рассматриваемых примеров. В первой колонке таблицы указан вид исходной информации; во второй колонке указано число использованных аргументов при построении уравнения регрессии; в третьей и четвертой колонке указаны соответственно сводный коэффициент корреляции и средняя квадратическая погрешность его расчетов по преобразованным значениям исходных рядов; в пятой и шестой колонке указаны соответственно сводный коэффициент корреляции и средняя квадратическая погрешность его расчетов по рядам, полученным после перехода от преобразованных значений к исходным.

Как следует из приведенных в таблице данных, по большинству представленных примеров нормализация и логарифмирование исходной информации повышает сводные коэффициенты корреляции уравнений регрессии сопоставляемых рядов и, следовательно, повышает точность расчетов по уравнениям регрессии.

Так, при расчете значений $БПК_5$ р. Великой в г. Опочка (нижний створ) — Y по уравнению регрессии по двум аргументам (растворимый кислород нижний створ — X_1; $БПК_5$ верхний створ — X_2) получен сводный коэффициент корреляции $R_0 = 0{,}67$, после нормализации исходных рядов методом Алексеева $R_0 = 0{,}77$, по логарифмированным рядам $R_0 = 0{,}74$.

При расчете разбавления сточных вод в период весеннего половодья на реке Юг у села Подосиновца (Y) по запасу воды в снеге плюс осадки (X_1) и показателю осеннего увлажнения почвы (X_2) получен сводный коэффициент корреляции $R_0 = 0{,}65$, по нормализованным рядам $R_0 = 0{,}81$, по логарифмированным рядам $R_0 = 0{,}33$.

При расчете объема стока за декабрь месяц по бассейну р. Уайт-Холлоу $R_0 = 0{,}79$, а по нормализованным и логарифмированным рядам $R_0 = 0{,}86$.

Наибольшее повышение сводного коэффициента корреляции, судя по приведенным примерам, получается после нормализации исходных рядов методом Алексеева. Логарифмирование дает несколько меньшую эффективность.

В двух случаях в выбранных примерах указанные преобразования оказываются неэффективными, в частности они не улучшили результатов расчетов уравнения регрессии максимальных расходов воды по р. Амур в пункте Богородское по связи с максимальными расходами в пунктах Хабаровск и Комсомольск-на-Амуре, а при расчетах уравнения регрессии по разбавлению сточных вод в период весеннего половодья (река Юг у села Подосиновца) по логарифмированным рядам результаты оказались значительно хуже, чем по исходным рядам.

Таблица 1

Результаты расчетов уравнений парной и множественной корреляции по исходным и преобразованным рядам

Исходные данные	Число аргументов	По преобразованным значениям исходных рядов		При переходе к исходным значениям	
		R_0	σ	R_0	σ
Максимальные расходы воды по р. Амур в пунктах Богородское (Y), Комсомольск-на-Амуре (X_1) и Хабаровск (X_2) за период 1893–2013 гг.					
По исходным рядам	2			0,98	0,004
По нормализованным рядам	2	0,97	0,01	0,96	0,007
По логаримированным рядам	2	0,97	0,01	0,97	0,006
Р. Великая в пункте Опочка н.ст. БПК (Y), раств. О2 (X_1), расход н.ст. (X_2), БПК в.ст. (X_3)					
По исходным рядам	3			0,67	0,05
По нормализованным рядам	3	0,77	0,03	0,73	0,04
По логаримированным рядам	3	0,74	0,04	0,67	0,05
Расчет разбавления сточных вод в период весеннего половодья. Река Юг у села Подосиновца (Y) коэффицент разбавления, (X_1) запас воды в снеге + осадки, (X_2) показатель осеннего увлажнения почвы					
По исходным рядам	2			0,65	0,13
По нормализованным рядам	2	0,81	0,08	0,85	0,06
По логаримированным рядам	2	0,33	0,2	0,44	0,19
Бассейн р. Уайт-Холлоу, сток за декабрь (Y), осадки за октябрь (X_1), ноябрь (X_2), декабрь (X_3)					
По исходным рядам	3			0,79	0,8
По нормализованным рядам	3	0,86	0,6	0,78	0,09
По логаримированным рядам	3	0,86	0,6	0,90	0,04

Вообще говоря, метод нормализации и линеаризации связей хорошо теоретически обоснован [2]. Он компенсирует несоответствие подавляющего большинства исходных рядов наблюдений указанным выше граничным условиям парной и множественной корреляции. В отличие от методов основанных на логарифмировании исходной информации он не только выпрямляет монотонно убывающие или монотонно убывающие связи но и нормализирует исходную информацию. Этим и объясняется более высокая эффективность этого метода. Дело в том, что законы распределения многих рядов наблюдений в науках о Земле являются асимметричными и для их описания, например, в гидрологии обычно используется или закон распределения Пирсона 3 типа или трехпараметрическое гамма распределение (закон распределения Крицкого-Менкеля).

Поэтому метод Алексеева, полностью линеаризующий монотонно убывающие и монотонно возрастающие связи и, что особенно важно, приводящий путем преобразований исходные ряды к нормальному закону распределения, казалось бы, должен быть эффективным во всех случаях. Теоретически доказано [2] что во всех случаях сводный коэффициент корреляции, рассчитанный по нормализованным рядам, должен быть

всегда больше или равен сводному коэффициенту корреляции по исходным рядам. Однако при практических расчетах в ряде случаев это положение не соблюдается.

Поэтому, для обоснованного использования метода преобразования Алексеева необходимо было выяснить причины расхождения ожидаемых и действительных результатов расчетов уравнений множественной корреляции на основе линеаризации и нормализации.

В табл. 1 в качестве примера представлены характеристики уравнения регрессии максимальных расходов воды по р. Амур в пунктах Богородское (Y), Комсомольск-на-Амуре (X_1) и Хабаровск (X_2) за период 1893–2013 гг. В примере приводится предельный случай, когда сводный коэффициент корреляции уравнения регрессии по исходным рядам достаточно велик: $R_0 = 0{,}98$. В то же время сводный коэффициент корреляции уравнения множественной корреляции, полученный по нормализованным данным $R_0 = 0{,}97$, то есть меньше сводного коэффициента корреляции полученного по исходным рядам наблюдений.

Конечно, исходя из чисто практических соображений, в данном случае вряд ли стоило применять какие-то преобразования, так как сводный коэффициент корреляции и так достаточно велик. И все же в чем дело, почему даже в этом случае, случае почти функциональной связи, нормализация не только не улучшает результат расчета по уравнению регрессии, а в некоторой степени ухудшает их.

По нашему мнению это определяется тем, что переходы от исходных к нормализованным рядам и обратный переход совершаются по теоретической кривой обеспеченности. Между тем эмпирические кривые обеспеченности, как и в данном случае, не полностью совпадают с теоретическими кривыми обеспеченности (рис. 1). Вследствие этого при переходе от исходных значений ряда к нормализованным и обратном переходе происходит сглаживание исходной информации, в тем большей степени, чем больше рассеивание эмпирических точек относительно теоретической кривой обеспеченности. Например, для исходного ряда максимального стока р. Амур у п. Богородское дисперсия $D = 38450000$, а дисперсия нормализованного ряда, то есть ряда значений, снятых по соответствующим обеспеченностям с теоретической кривой обеспеченности, равна 36780205. Таким образом, за счет нормализации ряда дисперсия снижается на 4,3 % по отношению к дисперсии исходного ряда. Очевидно, что тоже самое происходит при нормализации и с другими рядами. Важно также, что и обратный переход от нормализованных, рассчитанных по уравнению регрессии значений к исходным, также совершается по теоретической кривой обеспеченности. Это еще больше снижает дисперсию рассчитанных значений и в конце концов при большом разбросе значений ряда относительно теоретической кривой обеспеченности может привести к значительному снижению эффективности метода.

Следует отметить, что Г.А. Алексеевым был предложен и другой метод перехода от значений, рассчитанных по уравнению регрессии нормализованных величин, к исходным значениям [2] Этот переход предлагается производить по связи исходных и нормализованных значений (рис. 2). Но и в этом случае дисперсия рассчитанного ряда будет зависеть от тесноты этой связи, которая в свою очередь определяется разбросом точек относительно теоретической кривой обеспеченности. Это и приводит в некоторых случаях к понижению сводного коэффициента корреляции.

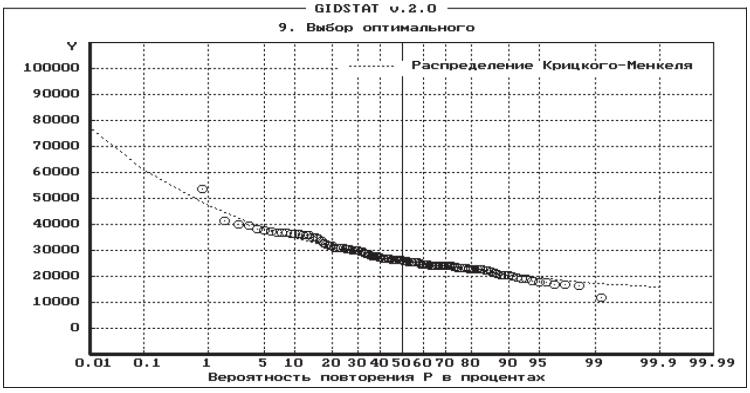

Рис. 1. Кривая обеспеченности максимальных расходов воды; р. Амур, п. Богородское

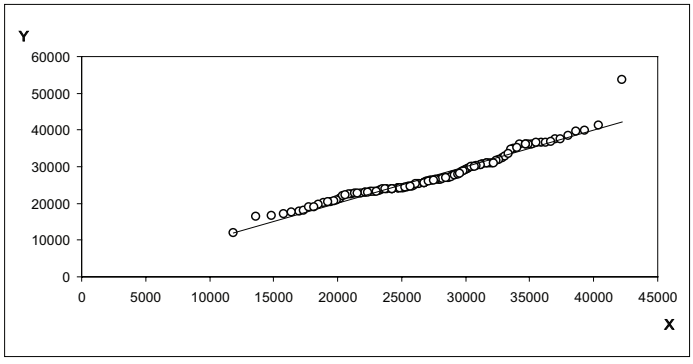

Рис. 2. График связи исходных (*Y*) и нормализованных (*X*) значений максимального стока; р. Амур, п. Богородское

Таким образом, эффективность применения метода нормализации и линеаризации связей Алексеева во многом зависит от согласия эмпирической и теоретической кривой обеспеченности для всех исходных рядов использованных при построении уравнения регрессии. Учитывая, что разброс значений эмпирической кривой относительно теоретической в некоторой степени зависти от длины исходного ряда, Как показывает опыт, для эффективного использования метода нормализации и линеаризации длинна совместного периода наблюдений по исходным рядам должна быть не меньше 15–20 лет.

При расчетах уравнения регрессии разбавления сточных вод в период весеннего половодья (р. Юг у села Подосиновца) (табл. 1) по исходным рядам сводный коэффициент корреляции $R_0 = 0{,}65$, а по логарифмированным рядам $R_0 = 0{,}33$. То есть, логарифмирование исходных рядов резко уменьшило тесноту связи. В тоже время по существующим представлениям логарифмирование исходных данных при построении монотонно убывающих или монотонно возрастающих связей должно или увеличивать тесноту связи или оставлять ее неизменной.

Для разрешения этого противоречия был проанализирован график связи (рис. 3) рассчитанных по уравнению регрессии логарифмов исходного ряда *X* и значений исходного ряда *Y*. Как и следовало ожидать в соответствии со сводным коэффициентом корреляции $R_0 = 0{,}33$, связь выражена очень слабо. Особенно сильно отклоняются от линии регрессии две точки в левой части графика при минимальных значениях ряда логарифмов. При исключении значения этих точек из исходных рядов, связь существенно изменилась и стала более выраженной (рис. 4). Увеличился и сводный коэффициент корреляции исходного ряда и ряда рассчитанного на основе уравнения множественной корреляции по логарифмам. Так по исходным рядам сводный коэффициент корреляции, при исключении этих двух значений стал равным $R_0 = 0{,}86$, а по логарифмам $R_0 = 0{,}88$.

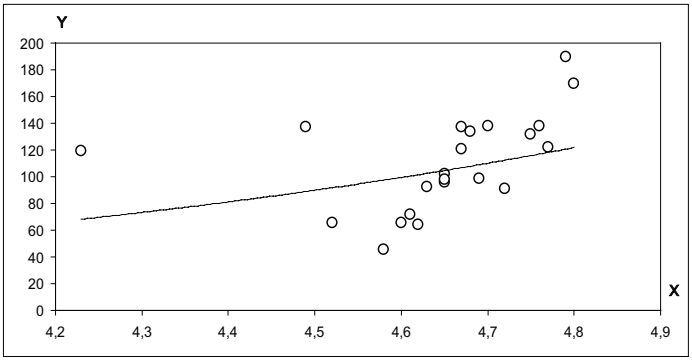

Рис. 3. График связи исходных (*Y*) и рассчитанных по уравнению регрессии логарифмов (*X*) значений разбавления сточных вод в период весеннего половодья на р. Юг у села Подосиновца

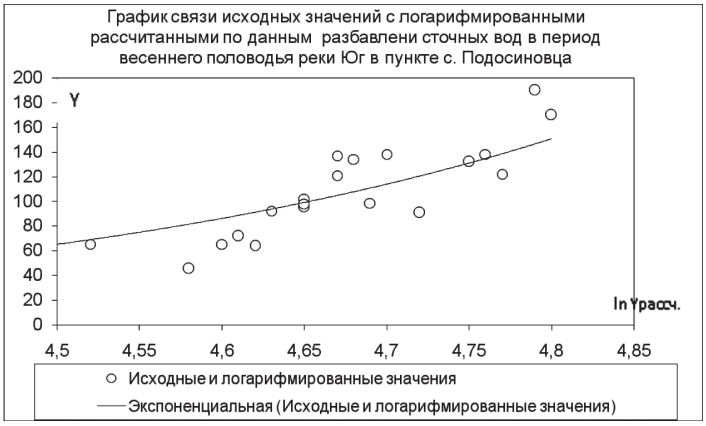

Рис. 4. График связи исходных (*Y*) и рассчитанных по уравнению регрессии логарифмов (*X*) значений разбавления сточных вод в период весеннего половодья на р. Юг у села Подосиновца при исключении двух минимальных значений

Для обоснования возможности исключения этих двух значений при логарифмировании исходной информации была проведена проверка однородности исходных рядов и рядов их логарифмов по критерию Диксона для минимальных значений. Оказалось, что гипотеза об однородности ряда по критерию Диксона не опровергается по всем исходным и логарифмированным рядам, за исключением ряда логарифмов X_2 (показатель осеннего увлажнения почвы). По этому ряду рассчитанное значение статистики Диксона $D = 0{,}32$ превосходит теоретическое значение при уровне значимости α равном одному проценту. На этом основании с большой вероятностью можно утверждать, что минимальные значения логарифмов ряда X_2 являются выбросами. Следовательно, эти значения и соответствующие им по дате значения рядов Y и X_1 необходимо исключать при построении уравнения регрессии [4].

Вместе с тем, тогда возникает другой вопрос, почему исходный ряд X_2 является однородным, а ряд логарифмов этого ряда имеет выбросы. Возможно дело в том, что в результате логарифмирования получается новый ряд с особыми свойствами, не присущими исходному ряду. Причем, эти отличия возрастают от меньших значений членов ряда к большим. Так если значения исходного ряда изменялись бы от 0 до 10 то логарифмы этого ряда изменяялись бы от 0 до 1, если от 10 до 100 то логарифмы изменялись бы от 1 до 2, если от 100 до 1000, то логарифмы от 2 до 3 и т.д. Особенно сильно эта непропорциональность изменений логарифмов по отношению к изменению значений исходного ряда сказывается на минимальных значениях, так как градиенты изменений на единицу логарифма, особенно при малых значениях исходных рядов, имеют гораздо большие приращения, чем на больших.

Таким образом на результаты расчетов по уравнению регрессии по логарифмированным рядам может в существенной степени сказаться наличие выбросов в исходных и логарифмированных рядах. Поэтому перед построением уравнений парной или множественной корреляции по логарифмированным рядам, необходимо проанализировать исходные ряды на наличие выбросов, то есть проверить не выходят ли экстремальные значения и их логарифмические преобразования в критическую область ординат кривой обеспеченности, например, Пирсона 3 типа.

В заключение следует отметить, что нередко при использовании различных преобразований в качестве окончательного результата принимается сводный коэффициент корреляции преобразованных рядов. На самом деле, теснота связи преобразованных рядов и исходных рядов может существенно отличаться [3]. Поэтому для объективной оценки эффективности разработанных методик в качестве окончательных результатов необходимо принимать результаты, полученные по исходным рядам данных. Для этого, рассчитав по уравнению регрессии значения преобразованных величин, необходимо, перевести их в действительные и затем уже по ним оценить эффективность метода. Кроме того, учитывая, что связи в результате ретрансформации могут быть нелинейными, в качестве показателя эффективности необходимо использовать широко известные соотношения:

– критерий случайности

$$\delta = D_{\Delta} / D \qquad (4)$$

– или критерий эффективности

$$F = D_p / D. \quad (5)$$

Выводы

1. На результаты расчетов по уравнениям парной и множественной корреляции в значительной степени влияет несоблюдение ряда граничных условий применения этих уравнений. Особенно важно соблюдать условия линейности связей между всеми исходными рядами и подчинения исходных рядов нормальному закону распределения.
2. С помощью методов линеаризации и нормализации можно во многих случаях добиться увеличения тесноты связи сопоставляемых рядов и, следовательно, повышения точности расчетов по уравнению регрессии
3. Для учета первого условия о линейности связей сопоставляемых рядов, в случае если связи являются монотонно убывающими или монотонно возрастающими, во многих случаях достаточно эффективным оказывается метод логарифмирования исходной информации. Вместе с тем при наличии выбросов в исходных или логарифмированных рядах эффективность этого метода уменьшается, поэтому желательно перед анализом связи проверить исходные ряды на наличие выбросов и, если они есть, убрать их.
4. Во многих случаях более эффективным методом, является метод линеаризации и нормализации связи Г.А. Алексеева. Однако, эффективность этого метода резко понижается при периоде совместных наблюдений меньше 10–15 лет и при большом разбросе значений эмпирической кривой обеспеченности относительно теоретической кривой.
5. В качестве показателя эффективности применения указанных преобразований при использовании уравнений регрессии необходимо использовать широко известные соотношения — критерий случайности или критерий эффективности.

Литература

1. *Шелутко В.А.* Численные методы в гидрологии. — Л.: Гидрометеоиздат, 1991. — 240 с.
2. *Алексеев Г.А.* Объективные методы выравнивания и нормализации корреляционных связей. — Л.: Гидрометеоиздат, 1971. — 362 с.
3. Статистические методы в гидрологии. Перевод с английского. — Л.: Гидрометеоиздат, 1967. — 271 с.
4. *Виссмен У.мл., Харбоф Т.И., Кнэпп Д.У.* Введение в гидрологию. — Л.: Гидрометиздат, 1979. — 470 с.
5. *Малинин В.Н.* Статистические методы анализа гидрометеорологической информации. — СПб.: РГГМУ, 2008. — 408 с.
6. *Шитиков В.К., Розенберг Г.С., Зинченко Т.Д.* Количественная гидроэкология: в 2 кн. — М.: Наука, 2005. 281 и 336 с.
7. *Голованова Е.Ю.* Статистические характеристики рядов многолетних изменений суммарных влагозапасов речных бассейнов (на примере России). // Ученые записки РГГМУ, 2014, № 33, с. 24–30.