УДК 556.165(282.247.41+282.247.13+282.247.21) «45»

О ПРЕДВЫЧИСЛЕНИИ ГОДОВОГО СТОКА КРУПНЫХ РЕК ЕВРОПЕЙСКОЙ ЧАСТИ РОССИИ НА ОСНОВЕ МЕТОДА ДЕРЕВЬЕВ РЕШЕНИЙ (DECISION TREES)

С.М. Гордеева, В.Н. Малинин

Российский государственный гидрометеорологический университет, gordeeva@rshu.ru

Рассматривается физико-статистический метод расчета годового стока крупных рек европейской части России (Волги, Северной Двины и Невы), в котором отбор предикторов осуществляется с помощью методов множественной линейной регрессии и метода деревьев решений (decision trees). Показано, что этот метод обладает целым рядом преимуществ, к которым относятся визуализация получаемых результатов, физически понятная их интерпретация и более высокая точность прогностических оценок годового стока рек.

Ключевые слова: метод деревьев решений, речной сток, осадки, Волга, Северная Двина, Нева, статистические модели.

ON PREDICTING ANNUAL RUNOFF OF LARGE RIVERS OF EVROPEAN RUSSIA BASED ON DECISION TREES METHOD

S.M. Gordeeva, V.N. Malinin

Russian State Hydrometeorological University

A physical-statistical calculation method of annual runoff of large rivers in the European Russia (Volga, Northern Dvina and Neva rivers) is considered. The selection of predictors is carried out using the methods of multiple linear regression and decision trees. It is shown that the method of decision trees has many advantages, which include the visualization of the results obtained, their physically comprehensible interpretation and higher accuracy of prognostic estimates of annual river runoff.

Keywords: decision trees method, runoff, precipitation, Volga river, Northern Dvina river, Neva river, statistical models.

Введение

Метод деревьев решений (decision trees) относится к числу наиболее популярных методов Data Mining (дословно «раскопка данных») — анализа экспериментальных данных с целью поиска неочевидных, объективных и полезных на практике закономерностей. На русском языке для этого вида анализа пока еще нет устоявшегося названия. Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, нейронных сетей, теории баз данных и др. Отсюда обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining [7—9, 11]. Метод деревьев решений (ДР) может быть использован для решения многих гидрометеорологических

задач, в частности задач классификации и прогнозирования. Однако единственной отечественной публикацией в данной области является работа [3], в которой построена прогностическая модель годового стока Северной Двины за период 1983—2013 гг., причем эта модель оказалась значительно точнее модели стока, построенной с помощью множественной линейной регрессии (МЛР). Описание метода ДР на русском языке можно найти, например, в работах [3, 6].

Целью настоящей работы является изучение возможности использования указанного метода для расчета годового стока крупных рек европейской части России (ЕЧР), находящихся в различных географических районах и в разных условиях увлажнения. Отметим, что актуальность оценок годового речного стока в настоящее время возросла, что в первую очередь связано с проблемой изменений глобального климата. Как показано во Втором оценочном докладе [2], интенсивность потепления на ЕЧР в последние десятилетия быстро возрастает, линейный тренд температуры воздуха за период 1976—2012 гг. составляет 0,52 °С/10 лет, что более чем в два раза превышает темпы повышения температуры воздуха в Северном полушарии (0,23 °С/10 лет). Естественно, это не может не приводить к значительным изменениям в характере увлажнения поверхности суши. Прежде всего, резко возросла повторяемость аномальных условий увлажнения, в том числе речных катастрофических наводнений и длительных маловодных периодов. Они вносят большой вклад в быстрое увеличение повторяемости опасных гидрометеорологических явлений (ОЯ), которые причиняют значительный экономический ущерб. Так, для территории России только за период 1996—2012 гг. (17 лет) повторяемость ОЯ возрастала со скоростью 188 случаев за 10 лет, в результате чего их число за этот период увеличилось более чем в два раза [2].

В настоящей работе для оценки годового стока выбраны три крупные реки ЕЧР: Волга (Волгоград), Северная Двина (Усть-Пинега) и Нева (Новосаратовка). В качестве основы для расчетов стока принят физико-статистический метод. Суть его состоит в том, что вначале устанавливаются физические связи функции отклика (речного стока) с определяющими факторами, а затем уже на статистической основе строится модель, заблаговременность которой зависит от инерционности воздействующих на изучаемый параметр факторов. Физической основой данного метода служит интегральное уравнение водного баланса речного бассейна. Для построения прогностической модели использованы метод деревьев решений и модель МЛР, сопоставление которых позволит более детально выявить достоинства и недостатки метода ДР.

Предвычисление годового стока Волги

С точки зрения прогноза межгодовых колебаний стока крупных рек определяющими являются климатические факторы [4]: запасы влаги в снежном покрове перед началом снеготаяния, предшествующее осеннее увлажнение почвы и предшествующее летнее увлажнение в апреле — сентябре.

Заметим, что чем больше площадь водосборного бассейна, тем больший вклад этих факторов в колебания стока и тем более длительную их предысторию

следует учитывать. Это означает, что сток в i-й год может зависеть не только от увлажнения в ($i-1$)-й год, но и частично в ($i-2$)-й год. Поэтому основная рабочая гипотеза может быть сформулирована следующим образом: накопление влаги (общее увлажнение) в бассейне за два предшествующих началу половодья года практически полностью определяет речной сток в его замыкающем створе до начала следующего половодья. Естественно, основное влияние на сток оказывает первый предшествующий год. Влияние второго года сказывается главным образом в аномальные по характеру увлажнения годы. Учтем также, что межгодовая изменчивость осадков значительно превышает аналогичную изменчивость суммарного испарения. Это означает, что испарением с поверхности бассейна можно пренебречь без существенной потери точности в расчетах. В результате модель для годового стока может быть записана в следующем виде:

$$Q_i = f\left(P^{\text{х}}{}_{(i-1)j}, P^{\text{т}}{}_{(i-1)j}, P^{\text{х}}{}_{(i-2)j}, P^{\text{т}}{}_{(i-2)j}\right), \qquad (1)$$

где Q_i — годовой сток реки; $P^{\text{х}}{}_j$ и $P^{\text{т}}{}_j$ — суммы осадков за холодный (октябрь — март) и теплый (апрель — сентябрь) период года на j–й станции; i — номер текущего года; $i-1$, $i-2$ — номера двух предыдущих лет соответственно.

В соответствии со сформулированной выше гипотезой в качестве годового стока должен приниматься период с апреля по март следующего года. Однако подобный годовой период усреднения значений речного стока не получил распространения на практике. Поэтому целесообразно в формуле (1) перейти к календарному годовому периоду, т.е. стоку с января по декабрь. В результате осадки за холодный период (октябрь — март) (i–1)-го года частично перекрывают рассчитываемый речной сток текущего i-го года (январь — март). В этом случае минимальная теоретическая заблаговременность расчета годового стока составляет девять месяцев, а реальная будет зависеть от заблаговременности получения данных об осадках. С одной стороны, формулу (1) можно рассматривать как прогностическую модель, поскольку между входящими в модель переменными и рассчитываемым параметром (годовым стоком) существует определенная заблаговременность. С другой стороны, формула (1) представляет собой физико-статистическую параметризацию годового стока крупной реки на основе легко определяемых параметров, а именно осадков, измеряемых на сети станций, имеющихся в открытом доступе.

Достаточно эффективным методом построения такой модели является пошаговый алгоритм МЛР [5]. Однако отбор предикторов может осуществляться и другими методами, например методом деревьев решений. Отметим, что стокоформирующая зона Волги находится выше Самары, ниже которой боковая приточность к реке практически отсутствует. Это означает, что южнее Самары увлажнение почти не сказывается на стоке Волги в замыкающем створе, поэтому его межгодовой изменчивостью можно пренебречь.

Работоспособность параметрической модели (1) была доказана в работе [4], в которой приводятся результаты оценки годового стока Волги методом МЛР за период 1891—1990 гг. В данной работе предиктором послужило количество осадков в теплое и холодное полугодие за период 1966—2012 гг. для станций,

расположенных в стокоформирующей зоне бассейна, которые выбирались с сайта ВНИИГМИ — МЦД из архива aisori.meteo.ru [1]. Первоначально в списке было около 40 станций, однако большая часть данных имела значительные пропуски. Поэтому были взяты ряды данных, имеющие минимальное число пропусков, а также представляющиеся наиболее достоверными, т.е. относящиеся к крупным метеостанциям. В результате осталось 26 станций. Таким образом, общее число предикторов в соответствии с формулой (1) составило $m = 104$. Конечно, данных указанного числа станций мало для репрезентативного описания поля осадков в бассейне. Однако использование в расчетах современных архивов информации, в которых данные об осадках заданы в узлах сетки (например, NCEP/DOE AMIP-II Reanalysis, GPCP) и получены методами интерполяции, может приводить к ошибкам, не поддающимся идентификации.

Зависимая выборка, по которой строилась модель МЛР, включала 35 лет, независимая — 10 лет (2003—2012 гг.). Среднеквадратическое отклонение (СКО) годового стока Волги у г. Волгограда составило 1435 м^3/с. Очевидно, оценка СКО может быть принята в первом приближении в качестве допустимой ошибки долгосрочного прогноза годового стока.

С помощью пошаговой процедуры МЛР методом включения переменных удалось построить оптимальную модель годового стока Волги, содержащую пять переменных. Оптимальность модели понимается в смысле ее адекватности по критерию Фишера и значимости всех предикторов по критерию Стьюдента при $\alpha = 0{,}05$. Первым предиктором в этой модели является количество зимних осадков в ($i-1$)-й год в пункте Кумены (водосбор реки Вятки). Корреляция их с годовым стоком Волги равна $r = 0{,}63$.

Статистические параметры модели приводятся в табл. 1. Нетрудно видеть, что коэффициент детерминации модели $R^2 = 0{,}82$, стандартная ошибка стока по зависимой выборке 646,5 м^3/с, что составляет 45 % СКО. Стандартная ошибка модели по независимым данным за 2003—2012 гг. составила 794 м^3/с. Отсюда следует, что, несмотря на резкое сокращение плотности осадкомерной сети в 90-е годы XX века, сохраняется довольно высокая точность оценок годового стока Волги.

Таблица 1

Статистические характеристики оптимальных моделей множественной линейной регрессии зависимости годового стока рек от зимних и летних осадков на метеорологических станциях водосборов Волги и Северной Двины в ($i-1$)-й и ($i-2$)-й годы, а также стока Невы от уровня Ладожского озера и суммы годовых осадков в Санкт-Петербурге в ($i-1$)-й и ($i-2$)-й годы

Река	Характеристика зависимой выборки				Стандартная ошибка по независимой выборке в долях СКО
	Период, годы	Число переменных	Коэффициент детерминации	Стандартная ошибка в долях СКО	
Волга	1968—2002	5	0,82	0,45	0,55
Северная Двина	1968—2002	5	0,47	0,79	1,16
Нева	1898—1997	4	0,51	0,71	1,00

Рассмотрим теперь способ построения модели годового стока Волги на основе метода ДР. Исходными данными послужила та же самая выборка данных об осадках на 26 станциях, используемых при построении модели стока Волги методом МЛР. Моделирование годового стока Волги выполнялось в пакете Statistica алгоритмом CART с априорными вероятностями, пропорциональными численности классов, и ценой ошибки классификации, одинаковой для всех классов [10]. В этом случае минимизация потерь эквивалентна минимизации доли неправильно классифицированных наблюдений. Ограничение на число ветвлений может приниматься по ошибке классификации с кроссвалидацией по независимой выборке, что обеспечивает отсечение дерева по минимальным цене — сложности.

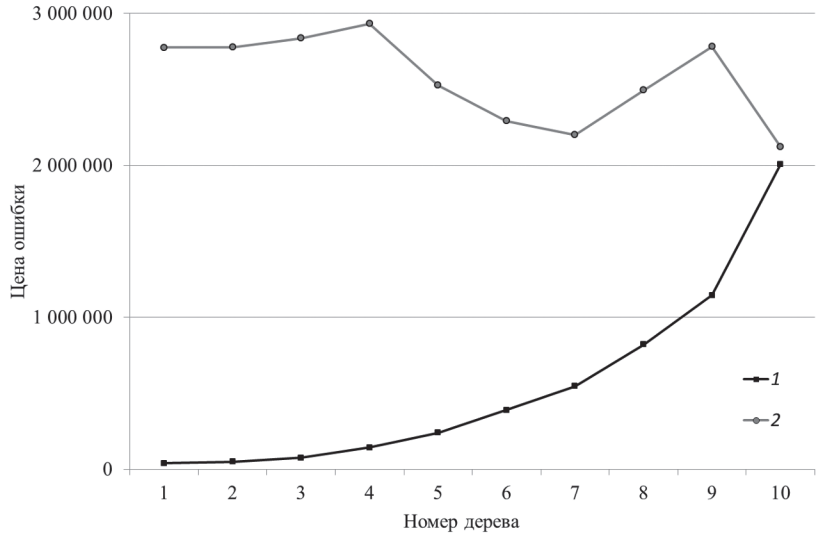
На рис. 1 *а* представлено распределение значений цены проверки на обучающей выборке *(Resubstitution cost)* и цены ошибки кроссвалидации по независимой выборке *(CV-cost*) в зависимости от числа узлов дерева. Дерево номер 1 имеет десять терминальных вершин и девять нетерминальных, а последнее дерево номер 10 — только одну вершину, что соответствует исходной выборке. Из рисунка следует, что с увеличением числа вершин цена ошибок обучения *(Resubstitution cost)* быстро уменьшается. Очевидно, чем «толще» становится дерево, тем точнее оно будет описывать зависимую переменную. В принципе, величина *Resubstitution cost* обратно пропорциональна коэффициенту детерминации, полученному по обучающей выборке. Распределение ошибок кроссвалидации *(CV-cost)* показывает, что дерево номер 7 имеет наименьшую ошибку независимых оценок, поэтому оно может быть принято как оптимальное. Дополнительно для оценки точности воспользуемся стандартными статистическими характеристиками: коэффициентом детерминации и стандартной ошибкой модели для зависимой выборки, которые были рассчитаны для периода 1968—2002 гг. для всех деревьев (табл. 2).

Таблица 2

Статистические оценки прогноза годового стока Волги по зависимой (1968—2002 гг.) и независимой (2003—2012 гг.) выборке для всех моделей деревьев решений

Номер дерева	Число вершин терминальных / нетерминальных	Коэффициент детерминации по зависимой выборке	Стандартная ошибка годового стока в долях СКО	
			по зависимой выборке	по независимой выборке
9	2 / 1	0,43	0,74	0,78
8	3 / 2	0,59	0,63	0,77
7	4 / 3	0,73	0,51	0,63
6	5 / 4	0,81	0,43	0,67
5	6 / 5	0,88	0,34	0,67
4	7 / 6	0,93	0,26	0,56
3	8 / 7	0,96	0,19	0,52
2	9 / 8	0,98	0,15	0,52
1	10 / 9	0,98	0,14	0,55

а)



б)

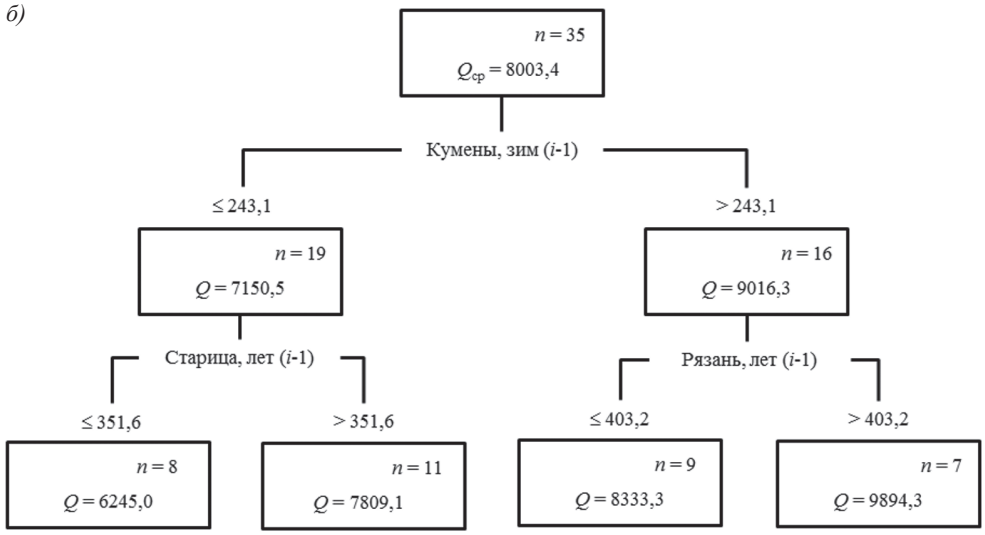


Рис. 1. Характеристики дерева решений, описывающего формирование годового стока Волги ($м^3/с$) в *i*-й год в зависимости от сумм зимних и летних осадков в ($i-1$)-й и ($i-2$)-й годы (мм) на метеорологических станциях, расположенных на территории бассейна, за период 1968—2002 гг.

а — значения цены проверки на обучающей выборке *(Resubstitution cost)* (*1*) и цены ошибки кросс-проверки *(CV-cost)* (*2*) в зависимости от числа узлов дерева; *б* — дерево решений номер 7.

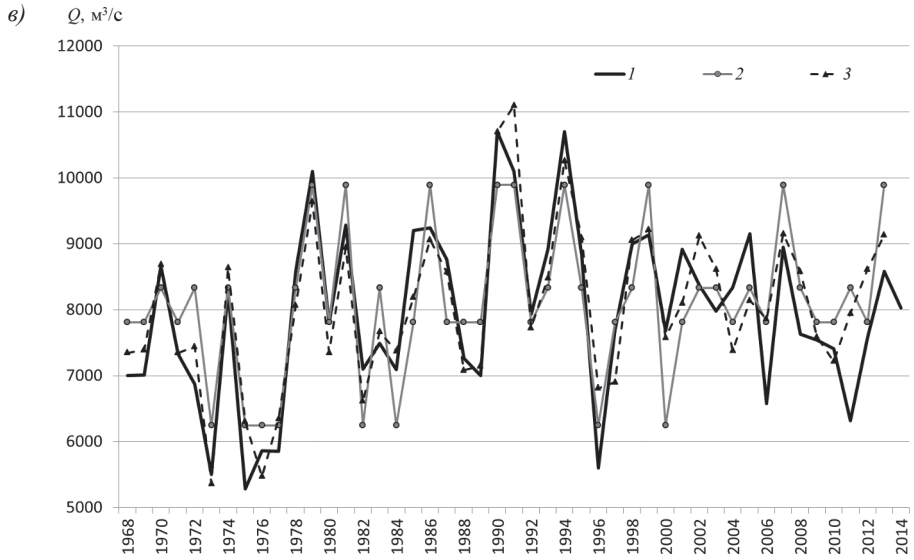


Рис. 1 *(окончание)*. Характеристики дерева решений, описывающего формирование годового стока Волги (м³/с) в *i*-й год в зависимости от сумм зимних и летних осадков в (*i*−1)-й и (*i*−2)-й годы (мм) на метеорологических станциях, расположенных на территории бассейна, за период 1968—2002 гг.

в — сопоставление фактических (*1*) и вычисленных по модели МЛР (*2*) и методом ДР (*3*) значений годового стока Волги у г. Волгограда (независимая выборка начинается с 2003 г.).

Из табл. 2 следует, что с увеличением числа ветвлений коэффициент детерминации для зависимой выборки довольно быстро увеличивается, а стандартная ошибка оценки годового стока уменьшается до дерева 3. Далее их изменения минимальны. Для данного дерева коэффициент детерминации между исходными и вычисленными значениями стока для зависимой выборки составляет $R^2 = 0{,}96$, стандартная ошибка стока равна 272 м³/с, или 19 % СКО. Одновременно для этого же дерева отмечается минимальная ошибка стока Волги по независимой выборке (752 м³/с, или 52 % СКО).

Итак, при использовании независимой выборки наилучшей является модель дерева 3. Отметим эффективность метода ДР, которая состоит в том, что на всех шагах ветвления, начиная в первого, стандартная ошибка вычисленных значений стока для независимой выборки меньше стандартного отклонения стока. При этом уже на третьем шаге ветвления, соответствующем оптимальному дереву 7, можно рассчитать годовой сток Волги с приемлемой для практических целей точностью (63 % СКО).

Дерево 7, представленное на рис. 1 *б*, имеет довольно простой вид. На первом ветвлении разделителем выступают зимние осадки за предшествующий, (*i*−1)-й год в пункте Кумены. Если осадков выпадает меньше 243,1 мм, то в 19 случаях из 35 отмечается низкий сток Волги со средним значением 7150,5 м³/с. Если осадков

выпадает больше 243,1 мм, то, наоборот, в 16 случаях сток становится высоким (в среднем 9016,3 м^3/с). Очевидно, ст. Кумены можно рассматривать как важнейший индикатор оценки межгодовых колебаний стока Волги, поскольку она одновременно является первым предиктором в модели МЛР.

На втором ветвлении происходит уточнение формирования 19 значений низкого стока Волги за счет летних осадков в пункте Старица за (i−1)-й год. Если осадков выпадало мало (< 351,5 мм), то отмечалось восемь значений аномально малого стока Волги (среднее 6245 м^3/с); если осадков выпадало больше 351,5 мм, то значения стока были близки к норме (среднее 7809 м^3/с). Следующим разделителем служат летние осадки на ст. Рязань за (i−1)-й год. Если осадков выпало больше 403 мм, то отмечалось семь значений аномально высокого стока (среднее 9894 м^3/с), если меньше 403 мм, то сток был чуть больше нормы (среднее 8333 м^3/с).

Сравнение фактических и вычисленных значений стока Волги у г. Волгограда по модели дерева 7 и модели МЛР представлено на рис. 1 *в*. По своим характеристикам модель дерева 7 несколько превосходит оптимальную модель МЛР. Весьма важно, что метод деревьев решений позволяет выявить особенности формирования стока различной водности. Так, индикатором аномально высокого стока Волги являются летние осадки в (i−1)-й год на ст. Рязань, а аномально низкого стока — летние осадки в (i−1)-й год на ст. Старица.

Если в модели МЛР описание дисперсии речного стока происходит за счет корреляции с исходными предикторами на всей длине временных рядов, то метод деревьев решений, очевидно, минимизирует расстояния между значениями стока и предикторами на отдельных временных отрезках ряда стока, причем с увеличением толщины дерева длина отрезков уменьшается.

Предвычисление годового стока Северной Двины

Северная Двина является важнейшей рекой ЕЧР, несущей свои воды на север. Ближайший к устью створ с. Усть-Пинега, замыкает площадь водосбора, равную 348 000 км². Годовой сток Северной Двины брался за период 1968—2012 гг. К сожалению, осадки за указанный период времени доступны только для восьми станций водосбора реки [3]. Прогноз стока данной реки очень сложен из-за слабого покрытия речного водосбора гидрометеорологическими данными.

В соответствии с формулой (1) был сформирован архив из 32 временных рядов предикторов. При этом архив был разделен на две части: зависимую выборку (1968—2002 гг.), которая использовалась для построения модели, и независимую выборку (2003—2012 гг.), по которой выполнялось сравнение фактических и вычисленных по модели значений годового стока.

Вначале для прогноза стока Северной Двины применялась модель МЛР. С помощью пошагового алгоритма МЛР получена оптимальная модель (см. табл. 1), в которой ее стандартная ошибка составляет 332 м^3/с при стандартном отклонении стока 419 м^3/с, или 79 % СКО. В принципе, для зависимой выборки результаты неплохие. Однако по независимой выборке стандартная ошибка прогноза составила 486 м^3/с, что существенно превышает СКО годового стока. Это означает, что при

таком слабом покрытии территории бассейна осадкомерными станциями данная регрессионная модель не позволяет рассчитывать годовой сток Северной Двины с необходимой точностью.

Совершенно иные результаты прогноза стока Северной Двины получены при использовании алгоритма CART. Исходное дерево номер 1 имеет девять терминальных вершин и восемь нетерминальных (табл. 3). Для этого дерева коэффициент детерминации и стандартная ошибка годового стока по зависимой выборке равны 0,85 и 0,38 % СКО соответственно. Стандартная ошибка годового стока по независимой выборке для всех деревьев меньше СКО. Из табл. 3 видно, что ее быстрое уменьшение наблюдается до дерева 4 (59 % СКО) с последующей стабилизацией на других шагах ветвления. Поэтому модель дерева 4 с шестью терминальными вершинами можно считать оптимальной. Итак, метод ДР по сравнению с моделью МЛР даже при слабом покрытии территории бассейна осадкомерными станциями обеспечивает решение задачи долгосрочного прогноза годового стока Северной Двины. При этом уже на первых шагах ветвления можно получить оценку стока с достаточной точностью.

В качестве примера на рис. 2 приводится модель дерева 7. Первым ветвлителем являются зимние осадки на ст. Великий Устюг в ($i-1$)-й год. Если их выпадало больше 278 мм, то в четырех случаях из 35 отмечался аномально высокий сток Северной Двины (среднее 3865 м^3/с), если меньше 278 мм, то в 31 случае наблюдался сток около нормы (среднее 3126 м^3/с). На втором ветвлении разделителем служат зимние осадки на ст. Каргополь в ($i-1$)-й год. Если их выпало больше 217 мм, то в 26 случаях из 31 сток должен быть около нормы (среднее 3197 м^3/с), если меньше 217 мм, то сток оказывался аномально низким (среднее 2758 м^3/с). На следующих шагах происходит уточнение формирования 26 значений «среднего» стока. Для дерева 4, помимо указанных выше станций, ветвлителями выступают летние осадки на ст. Мезень за ($i-2$)-й год, зимние осадки на ст. Шенкурск за ($i-2$)-й год, зимние осадки на ст. Онега за ($i-1$)-й год.

Таблица 3

Статистические оценки прогноза годового стока р. Северной Двины по зависимой (1968—2002 гг.) и независимой (2003—2012 гг.) выборке для всех деревьев решений

Номер дерева	Число вершин терминальных / нетерминальных	Коэффициент детерминации по зависимой выборке	Стандартная ошибка годового стока в долях СКО	
			по зависимой выборке	по независимой выборке
8	2 / 1	0,32	0,81	0,86
7	3 / 2	0,46	0,72	0,80
6	4 / 3	0,66	0,57	0,81
5	5 / 4	0,74	0,50	0,75
4	6 / 5	0,78	0,47	0,59
3	7 / 6	0,81	0,43	0,58
2	8 / 7	0,84	0,40	0,59
1	9 / 8	0,85	0,38	0,60

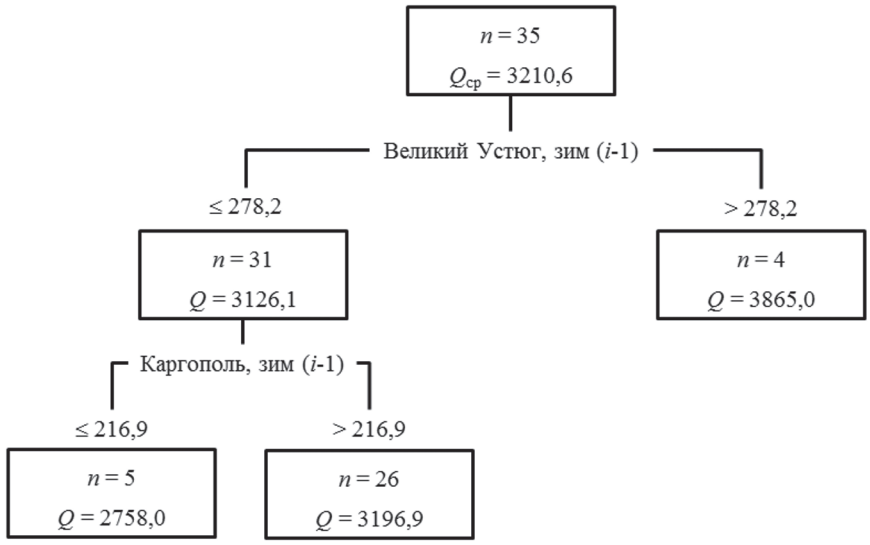


Рис. 2. Дерево решений номер 7, описывающее формирование годового стока Северной Двины ($м^3/с$) в i-й год, в зависимости от суммы зимних и летних осадков в ($i-1$)-й и ($i-2$)-й годы (мм) на метеорологических станциях, расположенных на территории бассейна за период 1968—2002 гг.

Предвычисление годового стока Невы

Как известно, Ладожское озеро и река Нева являются важнейшим звеном Волго-Балтийского и Беломоро-Балтийского водных путей. Долгосрочный прогноз стока Невы относится к числу актуальных региональных гидрологических проблем, и ее решение имеет большое практическое значение для многих отраслей экономики Санкт-Петербурга и Ленинградской области.

По мнению авторов, речной сток Невы и колебания уровня Ладожского озера целесообразно рассматривать как единую гидрологическую систему, функционирование которой почти полностью определяется естественными факторами, и прежде всего процессами влагообмена в системе океан — атмосфера — суша. Поэтому к расчету стока Невы полностью применим физико-статистический метод. При этом определяющим фактором служат изменения уровня Ладожского озера, который интегрирует запасы влаги водосборного бассейна озера в предшествующий период. Увлажнение вне водосборного бассейна Ладожского озера охарактеризуем количеством осадков в г. Санкт-Петербурге, поскольку внутри водосборного бассейна ни на одной станции не имеется временных рядов данных об осадках большой продолжительности.

В результате рабочая прогностическая формула годового стока Невы в i-й год Q_i примет следующий вид:

$$Q_i = f\left(P^*_{i-1}, P^*_{i-2}, H_{i-1}, H_{i-2}\right), \qquad (2)$$

где P^* — годовая сумма осадков в Санкт-Петербурге, H — среднегодовой уровень Ладожского озера, i — номер текущего года; $i-1$ и $i-2$ — номера двух предшествующих лет соответственно. Исходные данные по уровню и количеству осадков брались начиная с 1896 г., по стоку — с 1898 г. Зависимая выборка принималась за период 1898—1997 гг. (100 лет), независимая — за период 1998—2007 гг. (10 лет). Среднеквадратическое отклонение годового стока Невы на зависимой выборке составило 430,9 м^3/с.

Распределение основных статистических характеристик оценки стока Невы для полного дерева, построенного алгоритмом CART, представлено в табл. 4. Нетрудно заметить, что оптимальным следует считать дерево 11, для которого отмечается минимум стандартной ошибки стока по независимой выборке (0,74 СКО).

На рис. 3 приводится дерево 11, имеющее три нетерминальные вершины. Первым разделителем служит уровень озера в $(i-1)$-й год. Если он выше 455,5 см, то в 68 случаях из 100 сток Невы является высоким (среднее 2654 м^3/с), если меньше 455,5 см, то в 32 случаях — низким (среднее 2140 м^3/с). На втором шаге ветвителем опять выступает уровень озера в $(i-1)$-й год (H = 563 см), который отделяет просто высокий сток (среднее 2604 м^3/с) от аномально высокого стока (среднее 3170 м^3/с), наблюдавшегося в шести случаях из 100. Осадки в Санкт-Петербурге являются хорошим индикатором формирования маловодного стока. Если осадков в $(i-1)$-й год выпадает меньше 463 мм, то отмечается аномально низкий годовой сток Невы (среднее 1602,5 м^3/с).

Что касается предвычисления годового стока Невы на основе модели МЛР (см. табл. 1), то она состоит из четырех предикторов (уровня озера и количества осадков в Санкт-Петербурге в $(i-1)$-й и $(i-2)$-й годы). При этом ее стандартная ошибка по независимой выборке равна СКО, т.е. она значительно уступает в точности моделям деревьев до седьмого шага ветвления включительно.

Таблица 4

Статистические оценки прогноза годового стока р. Невы по зависимой (1898—1997 гг.) и независимой (1998—2007 гг.) выборке для всех деревьев решений

Номер дерева	Число вершин терминальных / нетерминальных	Коэффициент детерминации по зависимой выборке	Стандартная ошибка годового стока в долях СКО	
			по зависимой выборке	по независимой выборке
13	2 / 1	0,31	0,83	0,80
12	3 / 2	0,41	0,76	0,77
11	4 / 3	0,48	0,72	0,74
10	5 / 4	0,54	0,67	0,82
9	8 / 7	0,65	0,59	0,93
8	9 / 8	0,68	0,56	0,96
7	10 / 9	0,71	0,53	0,96
6	11 / 10	0,73	0,52	1,03
—	—	—	—	—
1	16 / 15	0,77	0,48	1,00

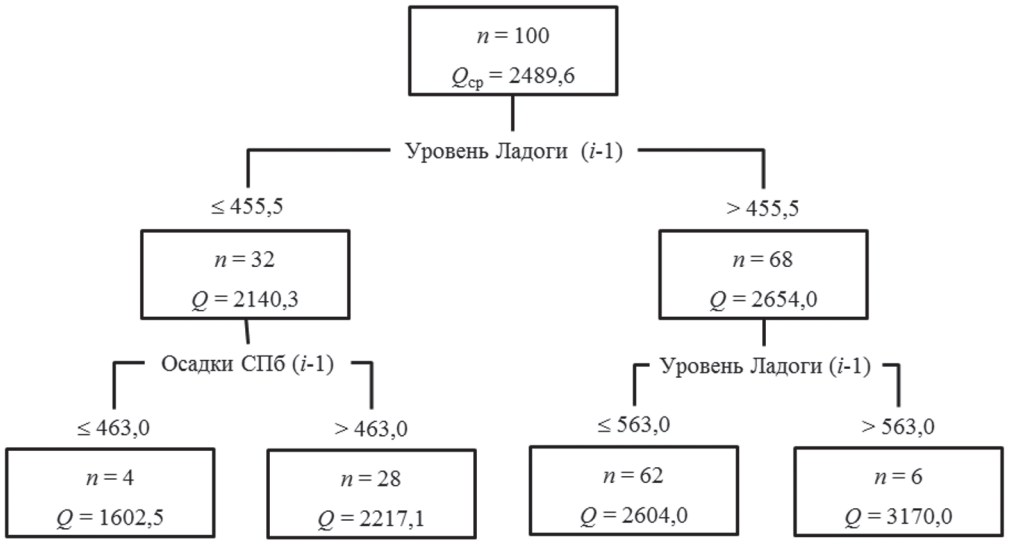


Рис. 3. Дерево решений номер 11, описывающее формирование годового стока Невы ($м^3/с$) в *i*-й год в зависимости от годовой суммы осадков в Санкт-Петербурге (мм) и среднегодового уровня Ладожского озера (см) в ($i-1$)-й и ($i-2$)-й годы за период 1898—1997 гг.

Заключение

Полученные результаты свидетельствуют о перспективности использования метода деревьев решений для оценок годового стока крупнейших рек ЕЧР (Волга, Северная Двина, Нева). В отличие от модели МЛР, в которой описание дисперсии речного стока происходит за счет корреляции с исходными предикторами по всей их длине, метод ДР минимизирует расстояния между значениями стока и предикторами на отдельных временных отрезках ряда стока, причем с увеличением толщины дерева длина этих отрезков уменьшается. При этом метод ДР обладает существенными преимуществами по сравнению с классическим методом МЛР, которые особенно ярко проявились для прогностических моделей стока Северной Двины и Невы. Так, для Северной Двины вследствие слабого покрытия бассейна осадкомерными станциями модель МЛР не позволяет рассчитывать ее годовой сток с необходимой точностью. В то же время, при использовании метода ДР уже на первых шагах ветвления можно получить оценку годового стока с достаточной точностью, а для наилучшей модели (дерево 3) стандартная ошибка расчета годового стока по независимой выборке почти в два раза меньше оценки его стандартного отклонения!

Для расчета годового стока Невы четырех предикторов (уровень Ладожского озера и количество осадков в Санкт-Петербурге за два предшествующих года) явно недостаточно, чтобы построить эффективную модель МЛР. Однако их вполне достаточно для получения относительно надежных значений стока Невы, начиная

с первого шага ветвления до седьмого включительно. Оптимальная модель (дерево 11) имеет стандартную ошибку расчета годового стока по независимой выборке, равную 0,74 СКО.

Важным преимуществом метода ДР, особенно для понимания формирования аномального по водности стока, является визуализация получаемых результатов и более понятная их интерпретация. Так, аномально высокий сток Волги отмечается, когда на ст. Рязань количество летних осадков в ($i-1$)-й год превышает 403 мм, аномально низкий сток — когда количество летних осадков в ($i-1$)-й год на ст. Старица меньше 355 мм. Индикатором аномально высокого стока Северной Двины служит количество зимних осадков на ст. Великий Устюг в ($i-1$)-й год (> 278 мм), а аномально низкого стока — количество зимних осадков на ст. Каргополь в ($i-1$)-й год (< 217 мм). Если уровень Ладожского озера превышает 563 см, то на следующий год ожидается аномально высокий сток Невы; если годовая сумма осадков в Санкт-Петербурге меньше 463 см, сток Невы на следующий год будет аномально низким.

Наконец, еще одно достоинство метода деревьев решений состоит в том, что уже на первых шагах ветвления, т.е. при малом числе предикторов, удается получить оценки годового стока с приемлемой для практических целей точностью. При этом в отличие от МЛР при ветвлении в модели деревьев неоднократно может входить один и тот же предиктор.

Список литературы

1. *Булыгина О.Н., Разуваев В.Н., Коршунова Н.Н., Швец Н.В.* Описание массива данных месячных сумм осадков на станциях России. Свидетельство о государственной регистрации базы данных № 2015620394. Электронный ресурс. Доступ: http://meteo.ru/it/178-aisori (свободный, 20 января 2018 г.).
2. Второй оценочный доклад Росгидромета об изменениях климата и их последствиях на территории Российской Федерации. М.: изд-во Росгидромета, 2014. 1003 с.
3. *Гордеева С.М., Малинин В.Н.* Использование Data Mining в задаче гидрометеорологического прогнозирования // Ученые записки РГГМУ. 2016. № 44. С. 30—44.
4. *Малинин В.Н.* Проблема прогноза уровня Каспийского моря. СПб: изд-во РГГМИ, 1994. 160 с.
5. *Малинин В.Н.* Статистические методы анализа гидрометеорологической информации. СПб: изд-во РГГМУ, 2008. 407 с.
6. *Чубукова И.А.* Data Mining. М.: Интернет-университет информационных технологий, Бином — лаборатория знаний, 2008. 384 с.
7. *Bramer M.* Principles of Data Mining. Springer, 2007. 344 p.
8. *Breiman L., Friedman J., Olshen R., Stone C.* Classification and Regression Trees. Wadsworth, Belmont, CA, 1984. 358 p.
9. Classification and Regression Trees: textbook. Electronic resource. URL: http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf (free, 20.01.2018).
10. Interactive Trees (C&RT, CHAID): Statistica Help. Electronic resource. URL: http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Gxx/Indices/ InteractiveTreesCRTCHAID_HIndex (free, 20.01.2018).
11. Popular Decision Tree: Classification and Regression Trees (C&RT). Electronic resource. URL: http://documents.software.dell.com/Statistics/Textbook/ Classification-and-Regression-Trees (free, 20.01.2018).