СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Доклад о состоянии санитарно-эпидемиологического благополучия населения в Воронежской области в 2016 году. – Воронеж : Упр. Федер. службы по надзору в сфере защиты прав потребителей и благополучия человека по Воронеж. обл., 2017. – 233 с.

2. Корчагина, В. А. Геоэкологическая экспресс-оценка качества поверхностных водных ресурсов Ближнего Подворонежья / В. А. Корчагина, Т. И. Прожорина, С. А. Куролап // Вестн. Воронеж. гос. ун-та. Сер.: География. Геоэкология. – 2008. – № 2. – С. 64–70.

3. Прожорина, Т. И. Оценка качества централизованного питьевого водоснабжения г. Воронежа / Т. И. Прожорина, И. П. Хруслова // Вестн. Воронеж. гос. ун-та. Сер.: География. Геоэкология. – 2013. – № 1. – С. 142–144.

4. Санитарно-эпидимеологические правила и нормативы «Питьевая вода. Гигиенические требования к качеству воды. Контроль качества» СанПиН 2.1.4.1074-01. – М. : Федер. центр Госсанэпиднадзора Минздрава России, 2002. – 103 с.

УДК 556.06(519.6)

С. В. СИДАК, А. А. ВОЛЧЕК
Беларусь, Брест, БрГТУ
E-mail: harchik-sveta@mail.ru

К ВОПРОСУ ОЦЕНКИ ГОДОВОГО СТОКА РЕКИ ПРИПЯТЬ НА ОСНОВЕ МЕТОДА ДЕРЕВЬЕВ РЕШЕНИЙ

Проблема оценки годового речного стока приобретает в последние годы все большее значение, так как она непосредственно связана с решением таких важнейших задач, как планирование и реализация дорогостоящих водохозяйственных мероприятий, надежное водообеспечение населения и экономики. Актуальность данной проблемы сегодня приобретает особое значение в связи с обострением воздействия глобального потепления на гидрологический режим рек.

В опубликованном в октябре 2018 г. Специальном докладе МГЭИК о глобальном потеплении на 1,5 °С освещается ряд последствий изменения климата, которых можно было бы избежать, ограничив глобальное потепление 1,5 °С по сравнению с 2 °С и более. Например, к 2100 г. при глобальном потеплении на 1,5 °С глобальное повышение уровня моря будет на 10 см ниже по сравнению с потеплением на 2 °С. В связи с этим исследование современных особенностей гидрологического режима рек, выявление долгосрочных тенденций изменения водного стока является важной гидрологической задачей.

Целью данной работы является анализ возможности применения метода деревьев решений для оценки годового речного стока реки Припять.

Основой для расчетов стока принят физико-статистический метод. Суть этого метода заключается в том, что изначально устанавливается физическая связь речного стока с определяющими факторами, а затем с помощью статистических инструментов происходит построение прогностической модели. В каче-

стве статистического аппарата использован метод множественной линейной регрессии (МЛР) [1, с. 235].

Оценка годового стока реки Припять. Рассмотрим задачу оценки годового стока реки Припять в створе Мозырь. Основными факторами, определяющими межгодовые колебания речного стока, являются влагозапасы в снежном покрове перед началом процесса снеготаяния, летнее и осеннее увлажнения почвы. То есть сток в i-й год зависит не только от увлажнения в (i – 1)-й год, но и в (i – 2)-й год. Учитывая то, что межгодовая изменчивость по такому фактору, как осадки, в значительной степени превышает изменчивость по фактору испарение, следует, что испарением с поверхности бассейна можно пренебречь. В связи с этим модель годового стока запишем в виде:

$$Q_i = f\left\{\sum_{j=1}^{c} P_{i-1,j}\ ,\sum_{j=1}^{w} P_{i-1,j}\ ,\sum_{j=1}^{c} P_{i-2,j}\ ,\sum_{j=1}^{w} P_{i-2,j}\right\}, \quad (1)$$

где c – количество месяцев в холодный период года (октябрь – март); w – количество месяцев в теплый период года (апрель – сентябрь); P – количество осадков; i – номер года.

Для построения модели (1) использовали алгоритм множественной линейной регрессии [1]. Годовой сток реки Припять брался за период с 1950 по 2006 г. Предиктором послужило количество осадков в теплый и холодный сезоны за период 1950–2006 гг. для семи станций. Таким образом, итоговое число предикторов, согласно формуле (1), составило n = 28. Вся выборка разделена на две части: зависимая (1950–2000 гг.), по которой строилась прогностическая модель МЛР, и независимая (2001–2006 гг.), используемая для получения оценки прогнозных свойств модели.

Согласно модели, построенной с помощью алгоритма МЛР, получили следующий результат: коэффициент детерминации – 0,79, но стандартная ошибка прогноза по независимой выборке превышает стандартное отклонение величины годового стока. Это означает, что необходим поиск альтернативных способов прогноза. Поэтому в дальнейшем исследовании был использован метод деревьев решений.

Метод деревьев решений. Метод деревьев решений может быть применен при решении многих гидрометеорологических задач, в том числе задач классификации и прогнозирования. Дерево решений можно представить в виде «ветвей», хранящих в себе значения атрибутов, от которых зависит целевая функция, и «листьев», на которых хранится значение целевой функции. В промежуточных узлах записаны атрибуты, по которым возможно различить один случай от другого.

Одним из самых известных способов построения деревьев решений является алгоритм CART. В отличие от нейронных сетей, применяемых при решении подобных задач, алгоритм CART обладает высочайшей скоростью и возможностью визуализации получаемых результатов. К важным достоинствам алгоритма CART следует также отнести возможность специальной обработки пропущенных значений [2, с. 53].

При использовании алгоритма CART каждый узел в дереве решений име ет двух потомков. Для того чтобы принять решение, к какому классу отнести си туацию, достаточно ответить на вопросы-правила типа *if – then*, формируемые узлах этого дерева. В ходе ответов на вопросы все множество примеров (обуча ющая выборка) делится на две части: часть, в которой правило выполняетс (ветвь right), и часть, в которой правило не выполняется (ветвь left). При выбор оптимального правила используется функция оценки качества разбиения (ил критерий расщепления). В алгоритме CART в качестве такой функции использу ется индекс *Gini*, определяемый по формуле

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2,$$

где p_i – вероятность класса i в узле T; n – число классов.

Наиболее сложным этапом в процессе построения дерева является момен остановки ветвления дерева или же, говоря иными словами, определение опти мального размера дерева. С одной стороны, чем больше размерность дерева, те точнее прогноз, а с другой – труднее интерпретация результатов.

В программе Statistica для алгоритма CART реализованы два способ остановки: *по отклонению,* когда «подходящее» дерево классификации выбира ется из «усеченных» с помощью специального правила стандартной ошибки и *прямая остановка* (FACT),при котрой пользователь сам устанавливае размеры дерева классификации, до которых оно может расти [3, с. 48].

В данной работе моделирование стока реки Припять реализовано помощью алгоритма CART с априорными вероятностями, пропорциональным численности классов. Исходными данными послужила выборка из 28 временны рядов предикторов. На рисунке 1 представлено распределение значений цен проверки на обучающей выборке и цены ошибки кросс-проверки в зависимост от номера дерева. Из рисунка 1 следует, что с увеличением числа вершин в дере ве ошибки обучения значительно уменьшаются.

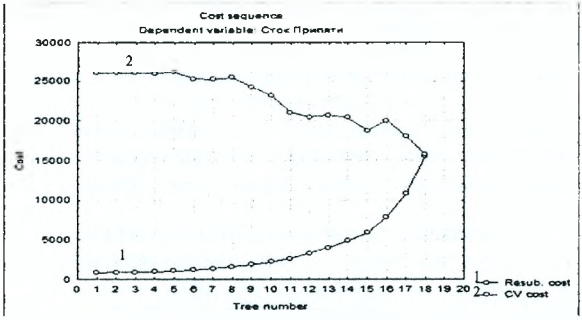

Рисунок 1 – Распределение цены проверки на обучающей выборке и цены ошибки кросс-проверки

Для определения дерева оптимального размера последовательно рассмотрены результаты прогноза стока реки Припять после каждого ветвления, вплоть до последнего дерева. На рисунке 2 представлено дерево решений 16, на первом ветвлении которого разделителем выступают зимние осадки за предшествующий i – 1 год в пункте Житковичи. Если количество осадков меньше 311 мм, то прогнозируемый сток реки равен 365,94 $м^3/с$, а если осадков выпало больше 311 мм, то ожидается сток больше нормы.

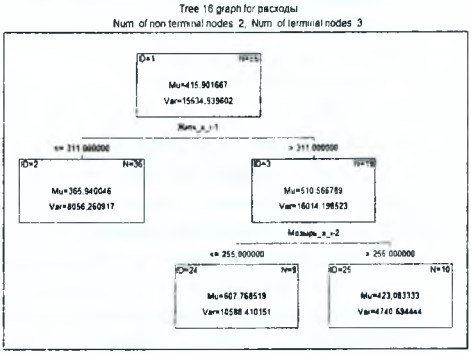

Рисунок 2 – Дерево решений 16, описывающее формирование годового стока Припяти

После расчета коэффициентов детерминации и стандартной ошибки стока между наблюдаемыми и вычисленными значениями стока для каждого полученного дерева пришли к выводу, что наилучшие оценки прогноза стока реки Припять получили из дерева 10, имеющего 9 терминальных вершин и 8 нетерминальных. Стандартная ошибка стока Припяти по независимой выборке составила 0,58 от величины стандартного отклонения стока, коэффициент детерминации составил $R^2 = 0,91$. Эти результаты в значительной степени отличаются от прогноза стока на основе модели МЛР.

Заключение. Сформулируем основные результаты и выводы, полученные в работе: 1) использование метода деревьев решений является перспективным направлением при оценке годового стока рек Беларуси; 2) используемый в работе метод построения дерева решений CART прост в понимании и интерпретации; 3) даже на самых первых этапах ветвления, при небольшом числе предикторов, имеется возможность получить оценки годового стока с приемлемой точностью.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Малинин, В. Н. Статистические методы анализа гидрометеорологической информации / В. Н. Малинин. – СПб. : Изд-во РГГМУ, 2008. – 407 с.

2. Гордеева, С. М. О предвычислении годового стока крупных рек европейской части России на основе метода деревьев решений (decisiontrees) / С. М. Гордеева, В. Н. Малинин // Учен. зап. РГГМУ. – 2016. – № 50. – С. 53–65.

3. Андреев, И. М. Описание алгоритма CART / И. М. Андреев // Exponenta Pro. Математика в приложениях. – 2004. – № 3–4. – С. 48–53.