ГЕОГРАФИЯ И ПРИРОДНЫЕ РЕСУРСЫ 2020 № 1 С. 166–174

Научный журнал http://www.izdatgeo.ru

МЕТОДИКА НАУЧНЫХ ИССЛЕДОВАНИЙ

УДК 556.06:519.237 DOI: 10.21782/GIPR0206-1619-2020-1(166-174)

А.В. ИГНАТОВ

Институт географии им. В.Б. Сочавы СО РАН,
664033, Иркутск, ул. Улан-Баторская, 1, Россия, ignatov@irigs.irk.ru

ПОДБОР ПРЕДИКТОРОВ ДЛЯ ПРОГНОСТИЧЕСКИХ МОДЕЛЕЙ СРЕДНЕГО РАСХОДА РЕКИ ОБИ В СТВОРЕ ГОРОДА БАРНАУЛА В ПЕРИОД ПОЛОВОДЬЯ

Рассмотрена задача построения статистической прогностической модели и связанная с этим проблема выбора предикторов для прогнозируемой переменной. В задаче используются данные о межгодовой изменчивости среднего расхода р. Оби в створе Барнаула в период половодья и о потенциально влияющих на него гидрометеорологических характеристиках. Утверждается, что результат выбора предикторов для прогностической модели зависит не только от используемых данных, но и от применяемого способа решения задачи. Этот способ определяется используемым для аппроксимации моделируемой зависимости математическим оператором, критерием оптимальности модели и алгоритмом подбора предикторов. Для исследования влияния способа моделирования на его результат проведен ряд компьютерных экспериментов, в каждом их которых применялись различающиеся между собой методы поиска оптимальной комбинации предикторов. Показано, что далеко не всегда наилучшее на обучающей выборке решение подтверждается на независимых данных. Для повышения устойчивости результата моделирования рекомендуется применять критерии выбора оптимальной модели, включающие в себя оценки ее надежности. Применение и сравнение разных способов построения моделей позволило выделить главные предикторы, на которые нагружается основная часть объясняемой дисперсии прогнозируемого расхода. Они определяются в первую очередь данными об объекте, имеют физическую интерпретацию, и их выбор менее зависим от используемого метода построения модели. Наибольшую эффективность показал способ построения прогностической модели как ансамбля частных моделей, в каждой из которых используется ограниченное число непересекающихся между собой предикторов. Такие модели, сохраняя устойчивость результата, позволяют наряду с главными предикторами учесть также влияние второстепенных факторов и за счет этого еще несколько улучшить качество разрабатываемой прогностической методики.

Ключевые слова: *формирование стока, стохастическая модель, операторы регрессии, условие выбора модели, компьютерный эксперимент, ансамблевый прогноз.*

A.V. IGNATOV

V.B. Sochava Institute of Geography, Siberian Branch, Russian Academy of Sciences,
664033, Irkutsk, ul. Ulan-Batorskaya, 1, Russia, ignatov@irigs.irk.ru

SELECTION OF PREDICTORS FOR PREDICTIVE MODELS OF THE AVERAGE DISCHARGE IN THE HYDROMETRIC SECTION OF THE OB RIVER NEAR BARNAUL DURING THE FLOOD PERIOD

The problem of constructing a statistical predictive model and the related problem of selecting predictors for the forecasted variable are considered. This study is based on using data on the interannual variability in average discharge of the Ob in the hydrometric section of Barnaul during the flood period and on hydrometeorological characteristics having a potential influence on it. It is argued that the result from selecting predictors for the predictive model depends not only on the data used but also on the method employed in solving the problem. Such a method is determined by the mathematical operator used to approximate the simulated dependence, the optimality criterion of the model and the algorithm for selecting predictors. To study the influence

of the modeling method on its result, a number of computer experiments were carried out, and each of them used different methods to find the best combination of predictors. It is shown that the best solution on the training sample is not always confirmed on independent data. To improve the sustainability of the simulation results, it is recommended that the criteria for selecting the optimal model should be used, which include assessments of its reliability. Use and comparison of different methods of constructing models made it possible to identify the main predictors which explain most of the variance of the forecasted discharge. They are determined primarily by data on the object and have a physical interpretation. Their selection is less dependent on the method used to construct the model. The highest effectiveness was shown by the method of constructing the predictive model as an ensemble of partial models, each of which uses a limited number of non-intersecting predictors. Retaining the sustainability of the simulation result, make it possible to take into account, along with the main predictors, also the influence of secondary factors and, hence, to improve somewhat the quality of the predictive technique being developed.

Keywords: *runoff formation, stochastic model, regression operators, model selection condition, computer experiment, ensemble forecast.*

ПОСТАНОВКА ПРОБЛЕМЫ

Средний расход рек в период половодья существенно изменяется год от года, что обусловливает потребность в его прогнозировании [1]. На формирование межгодовой изменчивости расхода реки влияют различные физические характеристики окружающей среды, которые сами являются функциями времени и пространственных координат. Наблюдения за многими из этих характеристик обычно ведутся в отдельных точках или областях, тяготеющих к соответствующему водосбору. Исторически накапливаемые сведения о переменных систематизируются в виде числовых оценок, описывающих измеренные значения или некоторые их статистические обобщения.

Рассмотрим задачу построения прогностической модели для расхода реки в выбранном створе. Такая модель однозначно определяется списком предикторов и оператором описания зависимости от них предсказываемой переменной. Можно выделить два крайних подхода, применение одного из которых приводит к физико-математическим, а другого — к статистическим моделям. Первые чаще строятся как детерминистические соотношения с распределенными параметрами [2, 3], для формулировки которых максимально используются известные законы. Вторые — это обычно модели с сосредоточенными параметрами. Физическое содержание статистических моделей определяется выбором предикторов прогнозируемой переменной [4]. Оператор, описывающий ее зависимость от таких предикторов, в статистической модели представляет собой некоторую формальную математическую структуру, предназначенную для приближенного вычисления значений предсказываемой характеристики. Информация об этой зависимости извлекается из данных прошлых наблюдений за параметрами моделируемого объекта в процессе построения модели. Также существует множество моделей, представляющих собой компромисс между физико-математическими и статистическими моделями, и компьютерные средства их создания [4–8].

В рамках данной статьи исследованы различные алгоритмы построения прогностических статистических (или стохастических) моделей. Числовые оценки, описывающие изменчивость измеряемых характеристик окружающей среды, рассматриваются в таких задачах как значения возможных предикторов предсказываемой переменной. В прогностических моделях часто используются предикторы, запаздывающие по времени относительно прогнозируемой характеристики. В этом случае одна и та же физическая переменная, но взятая с разным смещением во времени, выступает в роли нескольких разных переменных. В итоге при постановке задачи построения подобной модели первоначальное число возможных предикторов может достигать большой величины (несколько сотен или тысяч) и значительно превышать количество их совместных реализаций в исходных данных.

Каждый из возможных предикторов может в большей или меньшей степени влиять на изменчивость прогнозируемой характеристики, но мера такой зависимости и ее математическая структура априори неизвестны. Часто истинная и ложная эмпирические зависимости от возможных предикторов на ограниченной выборке данных статистически неразличимы. Кроме того, сами предикторы могут быть связаны между собой, их значения в исходных данных могут быть известны с разной точностью, а некоторые оценки значений части переменных в тех или иных совместных реализациях вообще отсутствовать. Дополнительные проблемы поиска прогностической зависимости создают также аномальные значения переменных и тренды в их временных рядах, уменьшающие статистическую значимость результатов анализа данных.

Попытка напрямую построить аппроксимирующий оператор, описывающий зависимость сразу от всех возможных предикторов, далеко не всегда оказывается удачной. Проблема может выражаться в нехватке условий для нахождения решения задачи, в неустойчивости получаемого результата модели-

рования, в понижении точности модельной аппроксимации из-за включения в модель ложных аргументов прогнозируемой переменной. В связи с этим в реальных исследованиях возникает необходимость подбора ограниченного числа предикторов, которые совместно наиболее существенно определяют наблюдаемую изменчивость зависимой характеристики. Это самая трудоемкая и неопределенная часть процедуры построения статистической модели, особенно в условиях малой выборки данных [9] и слабости искомых связей. С учетом такой ситуации в качестве задачи настоящей работы было выбрано сравнение ряда алгоритмов подбора предикторов для прогностических моделей, разрабатываемых на основе материалов наблюдений за средним расходом половодья р. Оби и его возможными аргументами.

ИСХОДНЫЕ ДАННЫЕ И МЕТОДЫ РЕШЕНИЯ ЗАДАЧИ

Первичные исходные данные, использованные в работе, были переданы автору Н.Н. Завалишиным в рамках договора о научном сотрудничестве между СибНИГМИ Росгидромета РФ и Институтом географии СО РАН. Эти материалы включали в себя месячные значения ряда гидрометеорологических характеристик за период 1900–2016 гг. и были предназначены для решения задач, связанных с разработкой моделей прогноза расхода в р. Оби и притока в Новосибирское водохранилище. Гидрологические числовые данные сформированы на основе результатов регулярных измерений расхода Оби в створе Барнаула и суммарного притока воды к створу Новосибирской ГЭС, а метеорологические представлены температурой приземного воздуха и суммой осадков на 15 метеостанциях, расположенных на рассматриваемой территории, а также оценками максимального содержания воды в снеге в горных и равнинных частях водосбора Оби выше створа ГЭС. Каждая месячная или экстремальная характеристика, отнесенная к отдельной метеостанции или к створу реки, рассматривалась как самостоятельная переменная, значения которой заданы с годовым шагом по времени. Всего таких переменных было 408, и каждой из них был присвоен свой порядковый номер.

Для задачи, описываемой в настоящей статье, из этих материалов была составлена таблица совместных реализаций исследуемого расхода Оби и его возможных предикторов. Значения этих предикторов характеризовали состояние водосбора Новосибирского водохранилища в каждый из шести месяцев, предшествующих началу второго квартала. Всего таких предикторов получилось 207. Использованные материалы отражали период с 1934 по 2016 г. Общий размер подготовленной таблицы числовых данных составил 83 строки оценок совместных значений 208 названных выше переменных. Часть оценок (около 2 % от общего числа) значений предикторов, относящихся преимущественно к началу рассматриваемого периода, отсутствовала. Имеющиеся оценки считались точными.

Данные с 1934 по 2007 г. использовались в качестве обучающей выборки для построения моделей. На данных с 2008 по 2016 г. (контрольная выборка) проверялась прогностическая способность построенных моделей. Изменчивость предсказываемой переменной (средний расход Оби в створе Барнаула во втором квартале года) на этих выборках характеризуется следующими числовыми параметрами. Стандартное отклонение на обучающей выборке составляет 592 $м^3/с$, на контрольной — 692 $м^3/с$. На разных фрагментах обучающей выборки, совпадающих по длине с контрольной, значение стандартного отклонения колеблется в диапазоне от 350 до 950 $м^3/с$. Среднее многолетнее значение ряда равняется 3058 $м^3/с$. Значимый тренд и автокорреляция по его временной структуре не наблюдаются.

Алгоритмы выбора предикторов основываются на оценке вклада аргумента (или комбинации аргументов) в формирование модельной оценки, объясняющей наблюдаемую изменчивость зависимой переменной [10–12]. Предварительный этап такой дифференциации осуществляет непосредственно разработчик модели, задавая из его собственных экспертных соображений список возможных предикторов прогнозируемой характеристики. На этом этапе предполагается, что все аргументы, включаемые в данный список, потенциально могут иметь больший или меньший вклад в формирование зависимой от них изменчивости модельной оценки. Для прочих характеристик этот вклад априори полагается ничтожным.

При дальнейшем решении задачи строится оператор модели [13], в рамках которого по заданному критерию оценивается вклад пробных комбинаций предикторов в изменчивость зависимой переменной. Комбинация предикторов, при которой достигается экстремум выбранного критерия, представляет собой искомый результат процедуры подбора аргументов. Более детальная оценка количественного вклада каждого включенного в оптимальную модель предиктора в изменчивость расчетной характеристики зависит от математической структуры оператора модели. Однако по причине часто имеющей место взаимосвязи предикторов между собой и характера зависимости от них прогнозируемой переменной их совместный вклад в формирование изменчивости функции не всегда может быть однозначно разделен на простые аддитивные или коммутативные составляющие.

Понятно, что результат подбора комбинации предикторов для стохастической модели помимо свойств исходных данных может зависеть от метода решения задачи. Вариант используемого метода определяется тремя основными факторами: математической структурой аппроксимирующего зависимость оператора, критерием оптимальности выбора предикторов и способом генерации их пробных комбинаций. Конкретизацию этих факторов осуществляет разработчик модели, что вносит субъективность и соответствующую неопределенность в результат решения задачи моделирования. Для оценки данной неопределенности и повышения объективности результата необходимо сравнивать между собой различные способы построения моделей и подбора предикторов для зависимой переменной. В нашем случае для выполнения этой работы была использована программа «Стохастическое моделирование», ряд инструментов которой специально предназначен для решения подобного рода задач. Свободно распространяемая версия [14] программы доступна любому заинтересованному лицу. В «Руководстве пользователя» [15] к ней можно найти пояснения к используемой далее специфической терминологии и к основным элементам технологии решения рассматриваемой задачи.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В рамках данной работы каждая постановка задачи моделирования и процедура ее решения рассматриваются как компьютерный эксперимент, результат которого — нахождение по тому или иному алгоритму оптимальной по заданному критерию модели с определенным списком предикторов. Комментарии к таким экспериментам изложены далее в тексте статьи, а их сводные результаты представлены в таблице.

Характеристики построенных в разных экспериментах моделей, предназначенных для прогноза среднего расхода реки Оби в период половодья

Номер эксперимента	Оператор регрессии*	Число предикторов в модели	Все или наиболее существенные предикторы (обозначены их номерами в исходной таблице описаний переменных), последовательно включаемые в модель в процедуре пошагового регрессионного анализа	Среднеквадратичная ошибка аппроксимации функции, $м^3/с$	
				Обучающая выборка	Контрольная выборка
1	1	207	Все переменные, включенные в список возможных аргументов	555	522
2	2	207		542	501
3	1	4	26, 53, 14, 124	286	207
4	2	5	26, 235, 236, 237	310	351
5	3	24	26, 53, 126, 61, 92, 361, 12, 89, 139 ...	136	363
6	4	24	26, 53, 178, 11, 191, 126, 235, 88, 61 ...	139	349
7	3	4	26, 53, 126, 36	274	232
8	4	4	26, 53, 126, 36	270	257
9	1	2	26, 53	312	284
10	2	3	26, 235, 236	311	388
11	3	3	26, 53, 124	292	236
12	3	3	26, 53, 126	281	200
13	С3	17	Список предикторов в данных моделях второго уровня формировался путем объединения предикторов, входящих в модели первого уровня. Процедура пошагового регрессионного анализа при построении итоговых моделей в этих экспериментах не применялась	198	340
14	С4	16		222	276
15	А3	33		296	168
16	А3	30		297	163
17	А4	18		272	195
18	А4	27		297	185
19	3А	6	26, 53, 126, 11, 104, 127	274	156
20	3А	7	26, 53, 126, 11, 235,92, 36	240	181
21	4А	9	26, 53, 178, 11, 191, 126, 154, 64, 263	247	144

* Операторы регрессии: 1 — локальная линейная экстраполяция, 2 — локальная квадратичная экстраполяция, 3 — линейный аддитивный, 4 — линейный коммутативный. С3 и С4 — сумма разложения функции по операторам 3 и 4. А3 и А4 — взвешенное осреднение операторов 3 и 4. 3А и 4А — соответственно операторы 3 и 4, используемые в моделях, построенных на основе новой постановки задачи со списком возможных аргументов, сформированным по результатам экспериментов 15–17.

Эксперименты 1, 2. Первой была сделана попытка решить задачу путем построения моделей зависимости расхода Оби от всех его заданных 207 возможных предикторов. Для ее реализации были использованы операторы «локальной линейной» и «локальной квадратичной» [15] экстраполяции, оптимизируемые по условию минимизации остаточной дисперсии на обучающей выборке. Они относятся к непараметрическим инструментам статистического моделирования [16, 17] и аппроксимируют искомую регрессию в многомерном пространстве предикторов с использованием локальных усредненных численных оценок функции, а также ее первой и второй производных, вычисляемых вдоль определенного направления. При заданной комбинации предикторов они оптимизируются только по одному коэффициенту — статистике осреднения данных, но требуют обязательного использования в этой операции метода «выбрасываемой точки» [18]. Такой алгоритм позволяет формально решать задачу построения модели при любом соотношении заданного числа предикторов и числа их совместных реализаций с зависимой переменной.

Эксперименты 3, 4. В этих экспериментах непараметрические операторы использовались для подбора наиболее существенных предикторов расхода половодья в р. Оби. Рекомендуемая комбинация выбиралась с использованием пошагового регрессионного анализа [19], а именно путем формирования пробных комбинаций последовательно увеличивающегося числа аргументов зависимой переменной и сравнения между собой результатов такого моделирования. Выбор наилучшей модели осуществлялся по условию минимума дисперсии аппроксимации функции на обучающей выборке. При таком алгоритме решения задачи искомый минимум реализуется при небольшом числе наиболее существенных предикторов в модели. С ростом их числа дисперсия аппроксимации функции начинает постепенно нарастать, совершая небольшие флуктуации вокруг тенденции этого нарастания.

Эксперименты 5, 6. В этой постановке задачи для построения моделей межгодовой изменчивости среднего расхода половодья использованы параметрические операторы регрессии — «линейный аддитивный» и «линейный коммутативный» [15]. У них число подбираемых на обучающей выборке постоянных коэффициентов (параметров) на единицу больше, чем число предикторов. С учетом объема обучающей выборки максимальное число предикторов в таких моделях было ограничено значением 24, чтобы на один подбираемый параметр для самой сложной модели приходилось не менее трех совместных реализаций функции и аргументов.

Рекомендуемая комбинация предикторов в этом случае выбиралась также с использованием пошагового регрессионного анализа. Выбор наилучшей модели с одинаковым числом предикторов осуществлялся по условию минимума дисперсии аппроксимации функции на обучающей выборке. При такой постановке задачи критерий оптимальности модели монотонно уменьшается с увеличением числа используемых в ней предикторов. Вследствие этого максимальная точность аппроксимации значений функции достигалась при максимальном разрешенном числе ее аргументов, т. е. при 24.

Эксперименты 7–10. Постановка задачи отличалась от экспериментов 3–6 тем, что вместо оценки дисперсии аппроксимации функции на обучающей выборке в качестве критерия выбора модели применялась характеристика, дополнительно учитывающая оценку вероятности сохранения аппроксимируемой зависимости на контрольных данных [15]. В таких условиях оптимум модели на обучающей выборке достигался при использовании не более четырех предикторов, как для параметрических, так и для непараметрических операторов регрессии.

Эксперименты 11–12. В этих экспериментах для поиска оптимальных моделей применялись более сложные алгоритмы, потенциально повышающие вероятность выбора «правильных» предикторов. В результате данной работы в двух разных вариантах решения задачи были выбраны модели с рекомендуемыми операторами, использующими по три предиктора.

Эксперименты 13–18. Для решения задачи был применен подход, использующий предварительное агрегирование первичных аргументов. Он был реализован с помощью технологии «двухуровневого моделирования», формирующего оператор регрессии для модели второго уровня путем объединения группы операторов частных моделей первого уровня с ограниченным числом предикторов. Частные модельные оценки в данном случае рассматриваются как новые переменные, формируемые из первичных возможных аргументов расхода половодья по алгоритмам, определяемым описаниями моделей первого уровня. В каждой такой частной модели используется от двух до пяти исходных предикторов, но их общее количество в объединенной модели второго уровня может достигать значительной величины.

Целесообразность применения такого подхода основывается на предположении о том, что с использованием оператора одного типа, как правило, нельзя описать все особенности фактической зависимости, которые могут отражаться в исходных данных. Но если разложить ее на некоторые компоненты, подобрать из стандартного набора для каждой компоненты свой оптимальный оператор, а

затем собрать по соответствующей схеме более сложную модель, то можно надеяться на извлечение большей информации из исходных данных и соответствующее улучшение результатов моделирования.

В этих экспериментах модели второго уровня строились двумя способами. Первый (применен в экспериментах 13, 14) — это разложение функции на аддитивные компоненты, каждая из которых формируется своей оптимальной частной моделью, и последующее суммирование операторов регрессии, использованных для построения таких моделей. Второй (эксперименты 15–18) — формирование ансамбля различных частных моделей (с разными комбинациями разноименных первичных предикторов), предназначенных для оценки одной и той же зависимой переменной. Модель второго уровня по этому способу строится как взвешенное среднее соответствующих операторов регрессии, используемых в моделях первого уровня. Весовые коэффициенты при таком осреднении ограничены значениями 0 и 1, а их сумма нормирована на единицу.

Эксперименты 19–21. В этих экспериментах для поиска оптимальной прогностической модели использовался повторный подбор предикторов для среднего расхода половодья р. Оби по правилам экспериментов 7 и 8, но уже не из первичного, а из более сжатого множества его возможных аргументов. Такие усеченные множества формировались на основе перечня наиболее существенных первичных предикторов, выделяемых специальной процедурой из предикторов, входящих в ансамблевые модели второго уровня, построенные в экспериментах 15–17.

Использование всех 207 возможных аргументов в качестве фактических предикторов в непараметрических моделях локальной экстраполяции функции приводит к необходимости осреднения данных по большому количеству реализаций для подавления влияния случайных факторов. В нашем случае локальное среднее значение функции вычисляется по ее 30–36 измеренным значениям, расположенным наиболее близко в пространстве аргументов к точке, в которой вычисляется модельное значение функции. Такое осреднение приводит к подавлению влияния не только ложных, но и истинных факторов. В результате доля объясненной дисперсии моделями зависимости от всех возможных предикторов оказывается малой как на обучающей, так и на контрольной выборке.

Если же при построении модели дополнительно использовать алгоритмы подбора наиболее подходящей комбинации предикторов, то результат аппроксимации функции непараметрическими операторами заметно улучшается (эксперименты 3, 4). Локальное осреднение в оптимальных моделях с четырьмя лучшими предикторами, при которых наблюдается минимальная остаточная дисперсия, осуществляется в этом случае не более чем по 20 значениям функции. Однако оптимальные на обучающей выборке комбинации предикторов при применении различных операторов регрессии могут оказаться разными, что может привести и к существенно различающимся результатам предсказания значений функции на контрольной выборке. Этот эффект также можно наблюдать при сравнении результатов моделирования в экспериментах 9 и 10.

Если для построения непараметрической модели данных недостаточно или в статистических распределениях значений переменных присутствуют существенные неоднородности, то более удачным для аппроксимации моделируемой зависимости может оказаться применение параметрических операторов регрессии. Проверка этой гипотезы применительно к решаемой задаче была выполнена в экспериментах 5–8. Результаты экспериментов 5 и 6 показали, что оптимизация таких операторов по условию минимума остаточной дисперсии на обучающей выборке приводит к излишней подгонке модели к данным. Сама оптимальная модель оказывается неустойчивой, что проявляется в большой ошибке тестового прогноза на контрольной выборке. Из этого следует, что применение таких операторов для подбора предикторов требует использования при выборе оптимальной модели дополнительных условий, регуляризирующих решение задачи. В нашем случае это включение в критерий выбора модели оценки вероятности, с которой аппроксимируемая зависимость сохранится на независимых материалах. Учет указанного параметра ограничивает чрезмерную подгонку модели к данным на обучающей выборке (эксперименты 7–10). Однако следует заметить, что при неудачном выборе оператора даже использование такого дополнительного условия не гарантирует успешности построения модели (эксперимент 10).

Применение более сложного алгоритма для совместного подбора предикторов и операторов с повышенными требованиями к устойчивости результата (эксперименты 11 и 12) показало, что для построения достаточно надежной модели с небольшим числом предикторов для анализируемых данных лучше всего подходит «линейный аддитивный» оператор.

По результатам экспериментов 3–6 можно построить графики ошибки предсказания значений функции на контрольной выборке в зависимости от числа предикторов в модели (рис. 1). Поведение этих графиков показывает, что такие зависимости немонотонны и их абсолютные минимумы не со-

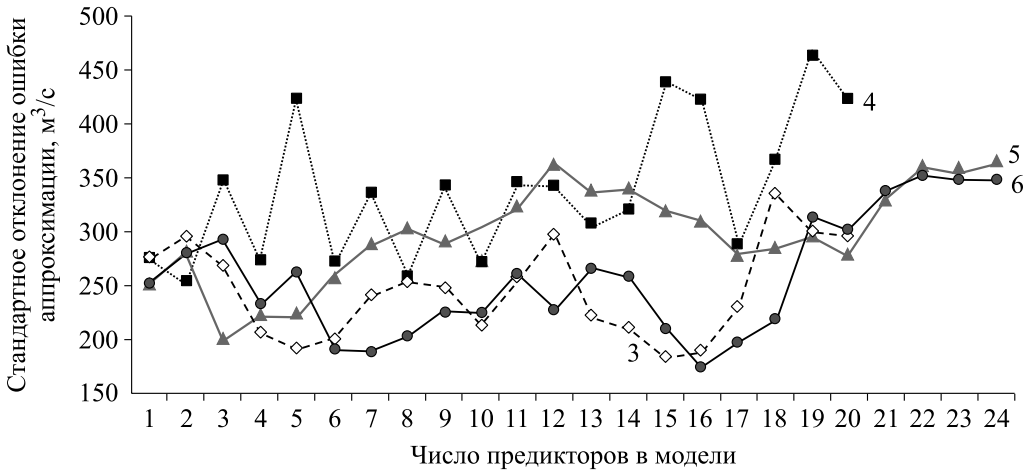

Рис. 1. Зависимость фактической ошибки предсказания расхода р. Оби (на контрольной выборке) от числа предикторов в модели.

Номера графиков соответствуют номерам экспериментов (см. текст и таблицу), для которых сравниваются результаты моделирования.

впадают по положению с соответствующими минимумами аппроксимации обучающей выборки (см. таблицу). Анализ графиков 3 и 6 (см. рис. 1) позволяет также предположить, что при использовании в прогностических моделях некоторых расширенных комбинаций предикторов могут быть получены результаты лучшие, чем в экспериментах 3–12. Попытка построения таких моделей сделана в экспериментах 13–21.

Результаты экспериментов 13 и 14 позволяют заметить, что разложение функции на аддитивные составляющие с последующим их суммированием приводит к моделям, промежуточным по качеству между экспериментами 5, 6 и экспериментами 7, 8. Более устойчивые модели с бóльшим количеством аргументов зависимой переменной строятся как ансамбль множества частных моделей с небольшим числом (2–5) предикторов путем объединения их операторов (эксперименты 15–18). Эта операция, выполняемая при прогностических расчетах, эквивалентна взвешенному осреднению множества прогнозов, сформированных по разным моделям для одной и той же характеристики. Еще немного продвинуться вперед в процессе совершенствования прогностической методики можно, если результаты построения ансамблевых моделей использовать для новой постановки задачи моделирования с суще-

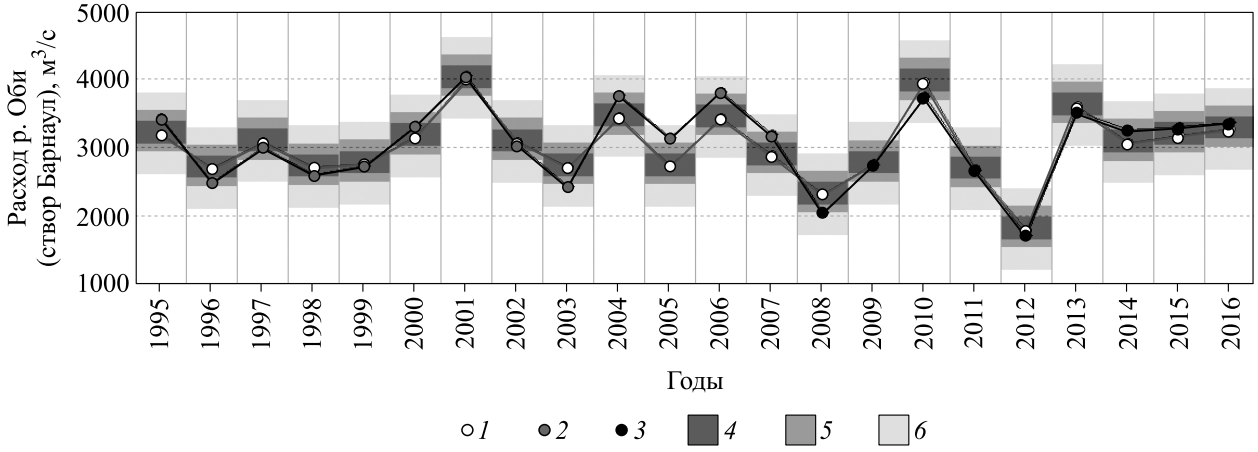

Рис. 2. Аппроксимация среднего расхода Оби в период половодья фрагмента (1995–2007 гг.) обучающей выборки и контрольных (2008–2016 гг.) данных с применением модели, построенной в эксперименте 21.

1 — медиана расчетной вероятностной оценки; значения фактической оценки: *2* — на обучающей выборке, *3* — на контрольной выборке; интервалы гистограммы приближенной модельной оценки с доверительной вероятностью, %: *4* — 50, *5* — 80, *6* — 96.

ственно меньшим (по сравнению с исходной постановкой) числом возможных предикторов (эксперименты 19–21). Фрагмент аппроксимации части обучающей выборки и контрольного прогноза по одной из таких «продвинутых» моделей в графической форме показан на рис. 2.

ЗАКЛЮЧЕНИЕ

Отбор лучшей модели по точности аппроксимации функции только на обучающей (см. таблицу) или только на контрольной выборке (см. рис. 1) не обеспечивает достаточную устойчивость получаемого решения. Сравнение между собой оптимальных моделей, построенных по разным алгоритмам с использованием обучающей выборки, а также независимая оценка их предсказательной способности на контрольной выборке совместно позволяют более объективно выбрать лучшую прогностическую методику.

Анализ последовательности подключения предикторов в различных процедурах пошаговой регрессии, использованных для построения моделей, подтвердивших свою устойчивость на независимых данных, позволяет выделить главные факторы формирования среднего расхода половодья Оби в створе г. Барнаула. Первый из них — это предиктор 26 (фиксируемый в конце марта накопленный запас воды в снеге в горной части водосбора). Второй фактор, скорее всего, отражает величину альбедо поверхности снега в конце зимы – начале весны, лучше всего индицируемую предикторами 53 и 178 (суммы февральских осадков, регистрируемые на метеостанциях Кулунда и Купино соответственно). Третий существенный фактор — накопленные к началу зимы запасы грунтовой влаги на водосборе. Основные индикаторы этого фактора — переменные под номерами 126, 124, 11, 36 (см. таблицу). Остальные переменные, входящие в построенные модели, отражают действие более слабых факторов формирования стока половодья.

Присутствие в модели с небольшим числом предикторов, отражающих действие трех основных факторов формирования стока половодья, обеспечивает приемлемую ошибку модельной аппроксимации фактических значений зависимой переменной как на обучающей, так и на контрольной выборке. Такие модели рекомендуется искать с учетом оценки их надежности.

Два главных предиктора (номера 26 и 53) определяются использованными исходными данными об объекте и входят во все построенные модели. Добавление других предикторов в модель в значительной степени зависит от способа их подбора и задаваемого пользователем или рекомендуемого применяемой компьютерной программой оператора регрессии.

Таким образом, для построения надежных моделей, отражающих влияние на зависимую переменную не только самых существенных, но и более слабых факторов, рекомендуется применять методы моделирования, позволяющие извлекать из исходных данных большее количество информации и одновременно обеспечивать достаточную устойчивость получаемого решения. В нашем примере такие свойства продемонстрировали ансамблевые модели, построенные с применением технологии «двухуровневого моделирования».

Работа выполнена в рамках НИР Института географии им. В.Б. Сочавы СО РАН (0347–2016–003).

СПИСОК ЛИТЕРАТУРЫ

1. **Абасов Н.В., Бережных Т.В., Резников А.П.** Долгосрочное прогнозирование природообусловленных факторов в энергетике // Системные исследования проблем энергетики. — Новосибирск: Наука, 2000. — С. 415–429.
2. **Виноградов Ю.Б.** Математическое моделирование процессов формирования стока. — Л.: Гидрометеоиздат, 1988. — 312 с.
3. **Кучмент Л.С.** Математическое моделирование речного стока. — Л.: Гидрометеоиздат, 1972. — 192 с.
4. **Резников А.П.** Предсказание естественных процессов обучающейся системой. — Новосибирск: Наука, 1982. — 287 с.
5. **Бураков Д.А.** Математическая модель расчета весеннего половодья для равнинных заболоченных бассейнов // Метеорология и гидрология. — 1978. — № 1. — С. 49–59.
6. **Гельфан А.Н.** Динамико-стохастическое моделирование формирования талого стока. — М.: Наука, 2007. — 280 с.
7. **Абасов Н.В.** Система долгосрочного прогнозирования и анализа природообусловленных факторов энергетики ГеоГИПСАР // Материалы междунар. совещ. APN (MAIRS/NEESP/SIRS) «Экстремальные проявления глобального изменения климата на территории Северной Азии»: Enviromis–2012. — Томск: Изд-во Том. центра науч.-техн. информации, 2012. — С. 63–66.

8. **Solomatine D.P., Dual K.N.** Model trees as alternative to neural networks in rainfall — runoff modeling // Hydrol. Sci. Journ. — 2003. — N 3. — P. 399–411.
9. **Гаскаров Д.В., Шаповалов В.И.** Малая выборка. — М.: Статистика, 1978. — 248 с.
10. **Кузьмин В.А.** Отбор и параметризация прогностических моделей речного стока // Метеорология и гидрология. — 2001. — № 3. — С. 85–90.
11. **Lorrai M., Sechi G.M.** Neural nets for modelling rainfall–runoff transformations // Water Resources Management. — 1995. — N 9. — P. 299–313.
12. **Дубров А.М.** Обработка статистических данных методом главных компонент. — М.: Статистика, 1978. — 134 с.
13. **Бураков Д.А., Гордеев И.Н., Игнатов А.В., Петкун О.Э., Путинцев Л.А., Чекмарёв А.А.** Прогнозирование притока воды в Красноярское и Саяно-Шушенское водохранилища во втором квартале года // География и природ. ресурсы. — 2016. — № 2. — С. 175–183.
14. **Игнатов А.В., Чекмарёв А.А.** Стохастическое моделирование. Версия 01 — Формулировка и проверка гипотез. Свидетельство о государственной регистрации программы для ЭВМ № 2017662529 от 10.11.2017 [Электронный ресурс]. — http://irigs.irk.ru/science/im.html (дата обращения 31.01.2018).
15. **Игнатов А.В.** Руководство пользователя программы «Формулировка и проверка гипотез» [Электронный ресурс]. — http://irigs.irk.ru/science/im.html (дата обращения 31.01.2018).
16. **Березин О.П.** Определение законов распределения малых выборок методом прямоугольных вкладов // Доклады к НТК по надежности судового электрооборудования. — Л.: Изд-во Науч.-техн. объед. судостроит. промышленности им. А.Н. Крылова, 1965. — Вып. 65. — С. 190–198.
17. **Добровидов А.В.** Об одном алгоритме непараметрической оценки случайных многомерных сигналов // Автоматика и телемеханика. — 1971. — № 2. — С. 88–89.
18. **Христофоров А.В.** Особенности задачи прогноза гидрологических характеристик по уравнениям регрессии // Метеорология и гидрология. — 1975. — № 11. — С. 72–80.
19. **Пошаговый** регрессионный анализ [Электронный ресурс]. — https://megalektsii.ru/s9796t1.html (дата обращения 31.01.2018).

Поступила в редакцию 29.05.2018
После доработки 15.02.2019
Принята к публикации 19.09.2019