С.В. Шаночкин

СИНОПТИКО-СТАТИСТИЧЕСКИЙ ПРОГНОЗ ВЕСЕННЕГО СТОКА НА Р. ТУРЕ

S.V. Shanochkin

SYNOPTIC-STATISTICAL FORECAST OF A SPRING DRAIN ON THE RIVER TOURA

В работе представлена прогностическая схема, основанная на использовании ряда методов многомерного статистического анализа. Учет основных аргументов воднобалансовых зависимостей периода весеннего половодья осуществлен через параметры атмосферной циркуляции. Инструментом параметризации послужил метод главных компонентов. В качестве атмосферных носителей памяти использовались поля среднемесячных аномалий давления и температуры воздуха. Один из вариантов прогностической схемы является двухэтапным, первый из которых – прогноз класса водности. Прогностические решения получены в виде уравнений регрессии.

The forecasting scheme based on the use of a number of methods of the multivariate statistical analysis is discussed in the work. The account of the basic arguments of water balance dependences of the period of spring high water is carried out through parameters of atmospheric circulation is used. As the tool of parametrization the method of the main components. The fields of monthly average anomalies of pressure and air temperatures are used as the atmospheric carriers of memory. One of the variants of forecasting scheme is a two-stage one, the first of which is the forecast of a class of water content. Forecasting decisions are received as the equations of regress.

Прогнозы стока за весенний сезон базируются в основном на физико-статистических воднобалансовых зависимостях. Очевидно, основное достоинство такого подхода – достаточно ясный физический смысл получаемых связей. Однако имеется и ряд недостатков: малая заблаговременность прогноза, техническая сложность, связанная с многократным проведением снегомерных съемок, и тот факт, что объем гидрологической информации с каждым годом становится все меньше.

В данной работе реализован синоптико-статистический метод [Смирнов, 1986], где все основные аргументы воднобалансовых зависимостей учитываются косвенно, через значения интегральных характеристик атмосферной циркуляции. В качестве таких интегральных характеристик использовались доступные и легко измеряемые величины – приземные месячные аномалии давления воздуха.

Рассматриваемый район характеризуется большим разнообразием физико-географических условий формирования стока в смысле климата и общей увлажненности бассейнов, заболоченности, глубины залегания грунтовых вод и состава почво-грунтов. Реки этой территории отличаются значительной водностью, что нередко приводит к катастрофическим последствиям, в связи с чем прогнозы весеннего стока являются весьма актуальными.

В данной работе разрабатывалось два варианта модели:

– непрерывная (линейная) модель;

– дискретно-непрерывная (кусочно-линейная) модель.

– При использовании непрерывной (линейной) модели можно выделить следующие этапы разработки:

– сбор данных о весеннем стоке и метеорологической информации;

– выделение районов, атмосферные процессы над которыми обусловливают условия стокообразования в период весеннего половодья;

– получение объективных количественных характеристик атмосферных процессов (предсказателей) в показательных районах;

– выбор из общей совокупности предсказателей наиболее информативной группы предикторов;

– построение и оценка прогностических зависимостей.

При конструировании кусочно-линейной модели прогностические рекомендации вырабатывались в два этапа – этап определения класса (фазы характеристики) и этап предсказания собственно самой характеристики, выраженной в количественной шкале.

Модели первого варианта – дискретные, модели второго – непрерывные и являются внутриклассовыми уравнениями регрессии. Таким образом, был получен список уравнений, каждое из которых действительно в определенной области признакового пространства.

Для выделения двух классов (выше и ниже нормы) использован один из алгоритмов дискриминантного анализа.

Дискриминантный анализ относится к методам многомерной классификации [Берэтисс, 1974]. В ходе дискриминантного анализа новые кластеры (классы) не образуются, а формулируется правило, по которому новые члены совокупности относятся к одному из уже существующих классов. Основанием для отнесения каждого члена совокупности к определенному множеству служит величина дискриминантной функции, рассчитанная по существующим значениям дискриминантных переменных. На практике очень часто дискриминантный анализ используется в сочетании с другими методами многомерной статистики, например с методом множественной регрессии.

Одной из моделей такого типа является кусочно-линейная модель, формальная запись которой имеет следующий вид:

$$y = \begin{cases} f_1(z), \text{ если } z \in D_1; \\ f_i(z), \text{ если } z \in D_i; \\ \dots\dots\dots\dots\dots\dots \\ f_k(z), \text{ если } z \in D_1, \end{cases}$$

где $f_i(z), i = 1, 2, ..., k$ – функция линейной регрессии предиктанта y на составляющие вектора-предиктора $z = (z_1, z_2, ..., z_n)$, полученные для каждого класса D_i.

С помощью дискриминантного анализа осуществляется альтернативный прогноз явления, который на следующем этапе конкретизируется на основе множественной регрессии.

Проблема репрезентативности выборки в задаче дискриминантного анализа стоит особенно остро, так как короткие ряды гидрометеорологических наблюдений необходимо разбивать на несколько частей, каждая из которых тоже должна быть представительной. В связи с этим в большинстве случаев дискриминантному анализу подвергается только два класса объектов (например, совокупность ситуаций выше и ниже нормы).

Пусть имеется множество единиц наблюдения, каждая из которых характеризуется несколькими признаками (переменными): z_{ij} – значение j-й переменной у i-го объекта; $i = 1, ..., n$; $j = 1, ..., p$.

Коэффициенты дискриминантной функции c определяются таким образом, чтобы для двух множеств (классов) максимальным было выражение:

$$\bar{f}_1(z) - \bar{f}_2(z) = \sum_{i=1}^{n_1} c_i z_{1i} - \sum_{i=1}^{n_2} c_i z_{2i}. \quad (1)$$

Тогда можно записать:

$$f_{kt}(z) - \bar{f}_k(z) = c_1(z_{1kt} - \bar{z}_{1k}) + c_2(z_{2kt} - \bar{z}_{2k}) + ... + c_p(z_{pkt} - \bar{z}_{pk}), \quad (2)$$

где k – номер группы; p – число переменных, характеризующих каждое наблюдение.

Обозначив дискриминантную функцию $f_k(z)$ как U_{kt} (k – номер группы, t – номер наблюдения в группе), внутригрупповую вариацию можно вычислить по выражению:

$$\sum_{i=1}^{n_k} (U_{kt} - \overline{U}_k)^2.$$

По обеим группам это будет выглядеть следующим образом:

$$\sum_{k=1}^{2} \sum_{t=1}^{n_k} (U_{kt} - \overline{U}_k)^2 = \sum_{k=1}^{2} \sum_{t=1}^{n_k} \left[c_1(z_{1kt} - \bar{z}_{1k}) + c_2(z_{2kt} - \bar{z}_{2k}) + ... + c_p(z_{pkt} - \bar{z}_{pk})\right]^2$$

или в матричной форме:

$$\sum_{i=1}^{2} \sum_{t=1}^{n_k} (U_{kt} - \overline{U}_k)^2 = C'(Z_1'Z_1 + Z_2'Z_2)C, \quad (3)$$

где C – вектор коэффициентов дискриминантной функции; Z_1' – транспонированная матрица отклонений наблюдаемых значений исходных переменных от их средних величин в первой группе; Z_2' – аналогичная матрица для второй группы.

$$Z_1' = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,n1} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,n2} \\ . & . & . & . \\ z_{p,1} & z_{p,2} & \cdots & z_{p,n2} \end{pmatrix};$$

$$Z_2' = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,n1} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,n2} \\ . & . & . & . \\ z_{p,1} & z_{p,2} & \cdots & z_{p,n2} \end{pmatrix}.$$

Тогда объединенная ковариационная матрица $B*$ ищется в виде:

$$B* = \frac{1}{n_1 + n_2 - 2}\left(Z_1' Z_1 + Z_2' Z_2\right).$$

Выражение (3) – оценка внутригрупповой вариации и его можно представить в виде:

$$\sum_{i=1}^{2} \sum_{t=1}^{n_k} \left(U_{kt} - \overline{U}_k\right)^2 = C'\left[\left(n_1 + n_2 - 2\right)B\right]C. \tag{4}$$

Межгрупповая вариация вычисляется следующим образом:

$$\left(U_1 - \overline{U}_2\right)^2 = C'\left(\overline{Z}_1 - \overline{Z}_2\right)\cdot\left(\overline{Z}_1 - \overline{Z}_2\right)' C.$$

Для рассматриваемых объектов внутригрупповая вариация должна быть минимальной, а межгрупповая вариация – максимальной, т.е. необходимо, чтобы величина F была максимальной:

$$F = \frac{C'\left(\overline{Z}_1 - \overline{Z}_2\right)\left(\overline{Z}_1 - \overline{Z}_2\right)' C}{C'\left[\left(n_1 + n_2 - 2\right)B*\right]C}. \tag{5}$$

В точке, где функция F достигает максимума, частные производные по c_j будут равны нулю. Если вычислить частные производные $\frac{dF}{dc_1}; \frac{dF}{dc_2}; ...; \frac{dF}{dc_p}$ и приравнять их нулю, то после преобразований получим выражение:

$$C = B*^{-1}\left(\overline{Z}_1 - \overline{Z}_2\right). \tag{6}$$

Из этой формулы и определяется вектор коэффициентов дискриминантной функции (C).

Полученные значения коэффициентов подставляются в формулу $f(z) = c_1 z_1 + c_2 z_2$, и для каждого объекта в обеих группах (множествах) вычисляют дискриминантные функции, затем находятся средние значения для каждой группы. Таким образом, каждое *i*-е наблюдение, которое первоначально описывалось *m* переменными, будет как бы помещено в одномерное пространство, т.е. ему будет соответствовать одно значение дискриминантной функции, следовательно, размерность признакового пространства снижается. В качестве границы, разделяющей 2 класса, равноудаленные от $\bar{f}_1$ и $\bar{f}_2$, служит константа дискриминации $C = \frac{1}{2}\left(\bar{f}_1 - \bar{f}_2\right)$.

Выбор и обработка данных о слое стока талых вод производились по единой методике, основное содержание которой заключалось в следующем.

Для гидропоста г. Туринска на р. Туре определялись даты начала, окончания половодья и слой весеннего стока в миллиметрах [Гидрологический ежегодник]. Даты начала и окончания половодья устанавливались на гидрографах по резким переломным точкам, при этом за его начало принимался первый день подъема, а за конец – дата перехода к летней межени. В качестве исходной метеорологической информации использовались месячные аномалии давления и у поверхности земли, которые определялись по данным соответственно 70 и 46 станций.

Важный этап разработки методики заключался в выделении районов, синоптические процессы над которыми определяли бы условия стокообразования в период весеннего половодья на р. Туре. Получение прогностических рекомендаций на основе учета региональных особенностей синоптических процессов с физической точки зрения наиболее оправдано. Прогностические признаки синоптических процессов могут быть выявлены статистическими методами.

В работе рассматривалась годовая предыстория развития атмосферных процессов. Информация за столь продолжительный период позволяет учитывать динамику, являющуюся следствием сложных синоптических процессов, происходящих в атмосфере, и может, в некотором смысле, служить признаком их дальнейшего развития. Более того, представляется, что, рассматривая годовую предысторию развития атмосферных процессов с целью обнаружения прогностически полезной информации, мы в меньшей степени подвержены риску совершить ошибку за счет непродолжительного, но резко аномального развития атмосферных процессов.

Предсказывающую информацию можно представить в виде 840-мерного вектора-предиктора, размерность которого определяется как произведение чисел

$$a_1 \cdot a_2 \cdot a_3,$$

где a_1 – число физических переменных (давление воздуха); $a_2 = 70$ – число метеостанций; $a_3 = 12$ – число месяцев.

Содержательная интерпретация такого массива информации представляется практически невозможной. Понижение размерности вектора-предиктора осуществлялось с помощью карт изокоррелят, которые строились для каждой из двух характеристик (стока и максимального расхода воды) за разные месяцы. С этой целью вычислялись коэффициенты корреляции между стоком за половодье и максимальными расходами воды весеннего половодья на р. Туре со значениями аномалий давления воздуха. Карты изокоррелят, построенные для различных месяцев, имеют весьма существенные отличия: области корреляций одного знака различаются местоположением, занимаемой площадью, величиной коэффициентов корреляции и т. д. Информация каждого месяца характеризовалась своими показательными областями. В качестве таковых выбирались крупномасштабные области локализации значимых коэффициентов корреляции, отдельно стоящие станции отбору не подлежали. Информативная ценность конкретных станций оценивалась по величинам полученных коэффициентов корреляции. В последующих вычислительных процедурах информация выделенных областей описывалась данными тех станций, которые имеют наибольшие коэффициенты корреляции со стоком и максимальным расходом воды за половодье в соответствующих зонах. Далее отобранная в результате релятивной компрессии информация обрабатывалась методом разложения метеорологических полей на естественные ортогональные составляющие. Анализ показал, что при формировании совокупностей достаточно ограничиться данными 7 – 11 станций. Как показали расчеты, основное содержание исходной совокупности метеорологических полей сосредоточено в нескольких первых членах разложения. В среднем первые 3–4 члена разложения охватывают 80 – 90 % суммарной дисперсии.

Прогнозирование стока выполнялось по двум регрессионным моделям, первая из которых была линейной, вторая – кусочно-линейной. Период с 1940 по 1983 г. был взят в качестве зависимого, а период с 1984 по 1989 г. – в качестве независимого. Рабочая совокупность предикторов формировалась путем отбора из коэффициентов разложения (главных компонент) наиболее информативных – в смысле вклада в коэффициент множественной корреляции. Для этой цели был применен регрессионный анализ. Он позволяет довольно просто отобрать подмножество информативных предикторов из заданного множества. В рамках линейной прогностической модели строились единые для всего зависимого 44-летнего периода уравнения связи, которые впоследствии использовались для прогноза максимального расхода и слоя стока весеннего половодья в течение последующих пяти лет.

В табл. 1 приведены уравнения, связывающие слой весеннего стока и максимальный расход воды с наиболее информативными главными компонентами среднемесячных полей давления воздуха.

По расчетным уравнениям можно дать прогноз с различной заблаговременностью. Учитывая, что начало весеннего половодья на р. Туре приходится в среднем на середину апреля, по первому и четвертому уравнениям можно дать прогноз со средней заблаговременностью 15 сут., в остальных случаях заблаговременность составляет 45 сут.

Таблица 1

Прогностические уравнения и их оценка (линейная модель)

№ п/п	Уравнение регрессии	S/σ
1	$y = 4{,}77T_1^{\text{III}} - 5{,}03T_1^{\text{VI}} + 3{,}89T_1^{\text{IX}} + 6{,}33T_2^{\text{IX}} - 4{,}10T_1^{\text{X}} + 1{,}35T_1^{\text{XI}} + 79{,}55$	0,50
2	$y = -5{,}47T_1^{\text{VI}} + 5{,}37T_1^{\text{IX}} + 6{,}03T_2^{\text{IX}} - 4{,}19T_1^{\text{X}} + 1{,}13T_1^{\text{XI}} + 79{,}55$	0,55
3	$y = -5{,}93T_1^{\text{IV}} + 5{,}26T_1^{\text{IX}} - 5{,}35T_1^{\text{X}} + 79{,}55$	0,59
4	$Q_{\max} = -66{,}30T_1^{\text{II}} + 18{,}69T_1^{\text{III}} - 79{,}54T_1^{\text{V}} + 36{,}60T_1^{\text{VII}} + 74{,}89T_3^{\text{VII}} - 29{,}18T_1^{\text{X}} + 743{,}91$	0,46
5	$Q_{\max} = -69{,}36T_1^{\text{V}} + 53{,}01T_1^{\text{VII}} + 66{,}38T_3^{\text{VII}} - 46{,}20T_1^{\text{X}} + 743{,}91$	0,56
6	$Q_{\max} = -58{,}34T_1^{\text{II}} + 49{,}40T_1^{\text{VII}} - 43{,}33T_1^{\text{X}} + 743{,}91$	0,60

Примечания: здесь y – слой стока за период весеннего половодья, мм; $Q_{\max}$ – максимальный расход воды, м^3/с; T_n^m – коэффициенты разложения (нижний индекс предиктора n обозначает номер главной компоненты, верхний m – месяц); S/σ – критерий применимости и качества методики (методика прогноза при длине ряда больше 25 членов считается удовлетворительной, если $S/\sigma \leq$ 0,80).

Рассмотренная выше прогностическая модель является обычной линейной моделью. Для реализации кусочно-линейной модели необходимо было получить уравнения связи для каждого из классов стока в отдельности. Получены они так же, как и выше описанные, т. е. с помощью регрессионного анализа. Различие состоит в том, что одни из них построены по данным, когда сток превышал норму, другие – когда он был ниже нормы.

Кусочно-линейная модель требует некоторого решающего правила, с использованием которого сначала определяется ожидаемая фаза (или класс) стока, а затем уже прогнозируется его численное значение по уравнению регрессии. Такое правило было получено с помощью дискриминантного анализа.

Расчеты выполнялись по следующей схеме. Величины весеннего стока р.. Туры были разбиты на два класса. В первый класс были отнесены годы, когда сток был выше нормы, во второй – годы со стоком ниже нормы. Далее, в соответствии с полученными классами величин стока были выделены обусловившие их классы предшествующих метеопроцессов, характер которых описывался значениями аномалий давления воздуха.

При выполнении дискриминантного анализа объем обучающей выборки состоял из данных об аномалиях давления, характеризующих группы многоводных и маловодных лет (по восемь лет в каждой группе соответственно). По этим данным были сформированы матрицы Z_1 и Z_2. Далее рассчитывались средние значения каждой переменной в отдельных группах для определения положения центров этих групп (осреднение по признаку).

Дискриминантная функция $f(z)$ данном случае имеет вид:

$$f(z) = c_1 z_1 + c_2 z_2 + ... + c_n z_n \,. \tag{7}$$

Коэффициенты $c_1,\ c_2,\ ...,\ c_n$ вычислялись по формуле:

$$C = B*^{-1}\ (\overline{Z}_1 - \overline{Z}_2)\,, \tag{8}$$

где $\overline{Z}_1,\ \overline{Z}_2$ – векторы средних в первой и во второй группах; C – вектор коэффициентов; $B*^{-1}$ – матрица, обратная совместной ковариационной матрице.

Для определения совместной ковариационной матрицы $B*$ рассчитывались матрицы B_1 и B_2. Каждый элемент этих матриц представляет собой разность между соответствующим значением исходной переменной и средним значением этой переменной в данной группе.

Совместная ковариационная матрица имеет следующий вид:

$$B* = \frac{1}{n_1 + n_2 - 2}(B_1 + B_2)\,,$$

где n_1 и n_2 – число объектов в 1-й и 2-й группе соответственно.

По совместной ковариационной матрице находилась обратная матрица. Далее определялся вектор коэффициентов дискриминантной функции по формуле (8).

В результате описанных вычислений были получены следующие значения коэффициентов:

c_1 = –0,38; c_2 = 0,42; c_3 = –0,83; c_4 = 0,19; c_5 = 0,58; c_6 = 0,97; c_7 = –1,02.

Подстановкой полученных значений коэффициентов в формулу (7) производился расчет дискриминантной функции для каждого объекта (года), вычислялись дискриминантные функции для каждого множества:

$$\bar{f}_1 = 2{,}159;\ \bar{f}_2 = -0{,}521.$$

Тогда константа дискриминации будет равна:

$$C = \frac{1}{2}(2{,}159 - 0{,}521) = 0{,}819\,.$$

После получения константы дискриминации проверялась правильность распределения объектов в уже существующих двух классах, а также проводилась классификация новых объектов (лет). Для того чтобы отнести объект к одному из двух множеств, для них рассчитывались значения дискриминантных функций. При $f > C$ объект (год) относился к первому множеству, т. е.

к многоводной группе лет, а при $f < C$ – ко второму множеству, т. е. к маловодной группе.

Методика дала положительные результаты, так как при классификации 49 лет ошибка распознавания имела место лишь в двух случаях, что составило 4,1 % от общего числа лет.

Регрессионные зависимости строились для каждой фазы (класса) в отдельности. Вклады первых трех коэффициентов разложения в общую дисперсию полей аномалий давления составили 75– 85 %. Уравнения регрессии строились с использованием наиболее информативных коэффициентов разложения. Построенные уравнения приведены в табл. 2.

Таблица 2

Прогностические уравнения и их оценка (кусочно-линейная модель)

№ п/п	Класс стока	Уравнение регрессии	S/σ
1	Выше нормы	$y = 4{,}68T_2^{\mathrm{V}} - 3{,}42T_1^{\mathrm{VI}} + 5{,}84T_3^{\mathrm{VI}} + 1{,}30T_1^{\mathrm{VIII}} + 2{,}73T_1^{\mathrm{X}} + 104{,}84$	0,57
2		$y = 8{,}75T_2^{\mathrm{III}} + 2{,}31T_2^{\mathrm{V}} - 2{,}54T_1^{\mathrm{VI}} + 104{,}84$	0,49
3		$y = 8{,}19T_2^{\mathrm{III}} + 1{,}69T_2^{\mathrm{V}} - 2{,}07T_1^{\mathrm{VI}} + 5{,}78T_3^{\mathrm{VI}} + 1{,}70T_1^{\mathrm{VIII}} + 0{,}78T_1^{\mathrm{X}} + 104{,}84$	0,42
4	Ниже нормы	$y = -2{,}39T_1^{\mathrm{VIII}} + 1{,}58T_1^{\mathrm{IX}} - 7{,}73T_3^{\mathrm{X}} + 53{,}21$	0,55
5		$y = -3{,}46T_2^{\mathrm{IV}} + 7{,}35T_4^{\mathrm{IV}} - 3{,}19T_1^{\mathrm{VIII}} + 53{,}21$	0,50
6		$y = -2{,}72T_2^{\mathrm{IV}} + 10{,}14T_4^{\mathrm{VI}} - 3{,}03T_1^{\mathrm{VIII}} + 0{,}81T_1^{\mathrm{IX}} - 10{,}44T_3^{\mathrm{X}} + 53{,}21$	0,44

Сравнительный анализ данных таблиц 1 и 2 дает основания считать, что кусочно-линейная модель эффективней обычной линейной модели. Полученные результаты позволяют также говорить о возможности успешного прогноза основных характеристик весеннего стока рек с использованием характеристик атмосферной циркуляции, являющихся, в свою очередь, косвенными оценками аргументов воднобалансовых зависимостей.

Литература

1. *Берэтисс А.Т.* Структуры данных. – М.: Статистика, 1974. – 408 с.
2. Гидрологический ежегодник. Т. 6. Бассейн Карского моря (западная часть). Вып. 4 – 9. – Свердловск, 1939 – 1998.
3. *Смирнов Н.П., Скляренко В.Л.* Методы многомерного статистического анализа в гидрологических исследованиях. – Л.: Изд-во ЛГУ, 1986. – 189 с.