УДК 551.509.314

МЕТОДИЧЕСКИЕ АСПЕКТЫ ОПРЕДЕЛЕНИЯ НЕОБХОДИМОГО ЧИСЛА ФАКТОРОВ НА ОСНОВЕ КОЛИЧЕСТВА ИНФОРМАЦИИ

И. И. Ульшин, В. С. Иванов

Военный учебно-научный центр Военно-воздушных сил «Военно-воздушная академия имени профессора Н.Е. Жуковского и Ю.А. Гагарина», г. Воронеж

При разработке новых прогностических метеорологических способов для отбора предикторов предлагается использовать факторный анализ. Показано, что для устранения существующей неопределенности в оценивании достаточного числа факторов необходимо рассчитывать количество информации, получаемой при переходе от пространства исходных признаков к факторному пространству.

КЛЮЧЕВЫЕ СЛОВА: ИНФОРМАЦИЯ, ФАКТОРНОЕ ПРОСТРАНСТВО, ПРОГНОЗ, МЕТЕОРОЛОГИЯ, ФАКТОРНЫЙ АНАЛИЗ

Одним из основных направлений совершенствования метеорологического обеспечения различных потребителей является разработка новых прогностических метеорологических способов. Его актуальность не снижается и в настоящее время, несмотря на продолжающееся развитие технических средств наблюдений за погодой и информационно-коммуникационных технологий.

В ходе разработки прогностических способов наиболее сложным этапом является отбор предикторов. Причина заключается в том, что, с одной стороны, в силу «табу статистического прогноза» перечень предикторов не может быть слишком большим, поскольку в противном случае трудно обеспечить надежность и статистическую значимость получаемых результатов. С другой стороны, его чрезмерное сокращение может приводить к потерям информации об исходном состоянии атмосферы. Для разрешения подобной проблемы обычно используются специальные процедуры последовательного присоединения или последовательного исключения признаков. Однако достаточно высокие результаты могут быть получены и при использовании в этих целях факторного анализа.

Факторный анализ предназначен для перехода от исходного пространства предикторов к факторному пространству значительно меньшей размерности без существенной потери информации. Таким образом, по своей сути факторный анализ решает задачу отбора предикторов, что делает его практически идеальным средством достижения указанной цели.

При всех несомненных достоинствах факторный анализ характеризуется и наличием недостатка, принципиально недопустимого именно при решении задачи отбора предикторов. Речь идет о неопределенности ответа на вопрос о количестве факторов, необходимом для адекватной замены исходного перечня признаков без потери информативности. В связи с этим целью данной статьи является разработка методических рекомендаций для получения однозначного ответа на вопрос об оптимальном числе оставляемых факторов. Это позволит использовать факторный анализ для отбора предикторов при построении прогностических метеорологических способов и, в конечном счете, повысить успешность прогнозирования атмосферных параметров.

Для однозначного оценивания требуемого число факторов предлагается рассчитывать частное количество информации, получаемой или теряемой при переходе от пространства исходных признаков к факторному пространству. Для этого следует использовать известное выражение, основанное на том, что получение информации предполагает уменьшение неопределенности:

$$I(z) = H_{pr} - H_{ps}, \qquad (1)$$

где $I(z)$ – количество информации; H_{pr} – априорная энтропия; H_{ps} – апостериорная энтропия [1].

Рассчитать количество информации, получаемое или теряемое при переходе от некоторого количества исходных предикторов к меньшему числу факторов, позволяет использование специального показателя, характеризующего соотношение масштабов «старой» и «новой» координатных систем – якобиана. При переходе от координат x_1, x_2, x_3 к координатам y_1, y_2, y_3 (например, от предикторов к факторам) он обозначается как

$$J\left\{\frac{x_1,\,x_2,\,x_3}{y_1,\,y_2,\,y_3}\right\} \tag{2}$$

и представляет собой следующее выражение:

$$J\left\{\frac{x_1,x_2,x_3}{y_1,y_2,y_3}\right\}=\begin{vmatrix}\frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \frac{\partial x_1}{\partial y_3}\\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_2}{\partial y_3}\\ \frac{\partial x_3}{\partial y_1} & \frac{\partial x_3}{\partial y_2} & \frac{\partial x_3}{\partial y_3}\end{vmatrix} \tag{3}$$

Опуская ряд теоретических положений и преобразований, можно записать, что в ходе преобразования координат энтропия изменяется следующим образом:

$$H(Y)=H(X)-\int P(X)\,\log\left[\left|J\left\{\frac{X}{Y}\right\}\right|\right]dX. \tag{4}$$

Таким образом, энтропия при переходе к новым координатам равна исходной энтропии минус математическое ожидание логарифма модуля якобиана преобразования от исходных координат к новым [2].

Как уже было сказано выше, в ходе проведения факторного анализа происходит переход от одной системы координат к другой. При этом новые переменные – факторы – связаны с исходными величинами линейными зависимостями. Поэтому для оценки изменения энтропии необходимо рассмотреть линейное преобразование координат. В общем случае подобное преобразование описывается формулами:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k,\\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k,\\ &\ldots\ldots\ldots\quad \ldots\ldots\ldots\quad \ldots\ldots\ldots\quad \ldots\ldots\ldots\quad \ldots\\ y_k &= a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k, \end{aligned} \tag{5}$$

Нетрудно вычислить якобианы $J\left\{\frac{Y}{X}\right\}$ и $J\left\{\frac{X}{Y}\right\}$ подобного преобразования:

$$J\left\{\frac{Y}{X}\right\}=\begin{vmatrix}\frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2}\ldots\frac{\partial y_1}{\partial x_k}\\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2}\ldots\frac{\partial y_2}{\partial x_k}\\ \ldots\ldots\ldots\ldots\ldots\ldots\\ \frac{\partial y_k}{\partial x_1} & \frac{\partial y_k}{\partial x_2}\ldots\frac{\partial y_k}{\partial x_k}\end{vmatrix}=\begin{vmatrix}a_{11} & a_{12}\ldots a_{1k}\\ a_{21} & a_{22}\ldots a_{2k}\\ \ldots\ldots\ldots & \ldots\ldots\\ a_{k1} & a_{k2}\ldots a_{kk}\end{vmatrix} \tag{7}$$

Если записать определитель из коэффициентов сокращённо в виде $\left|a_{ij}\right|$, можно получить более краткую форму записи (6):

$$J\left\{\frac{X}{Y}\right\} = \left|a_{ij}\right|^{-1}. \tag{8}$$

Следовательно, согласно (4),

$$H(Y) = H(X) - \int P(X) \log\left[\left|J\left\{\frac{X}{Y}\right\}\right|\right] dX = \\ = H(X) + \log\left|a_{ij}\right|. \tag{9}$$

Необходимо заметить, что при таких преобразованиях координат, при которых масштаб не изменяется, например, при повороте координатной системы или при смещении её начала, $J = 1$ и, следовательно, $H(Y) = H(X)$ [2]. При линейном преобразовании координат энтропия, согласно (9), изменяется на величину логарифма определителя матрицы коэффициентов данного преобразования. Таким образом, поскольку в факторном анализе рассматривается линейное преобразование пространства исходных признаков в факторное пространство, появляется возможность определить изменение энтропии при данном преобразовании. С учётом (1) и (9) легко получить выражение для вычисления количества информации, равное величине «исчезнувшей» неопределённости:

$$I = H(X) - H(Y) = -\log\left|a_{ij}\right| \tag{10}$$

Разумеется, количество информации будет положительным только в тех случаях, когда $H(Y) > H(X)$, т.е. когда неопределённость уменьшается. Для этого в выражении (10) величина логарифма должна быть отрицательной, что будет иметь место при значениях определителей матриц факторных нагрузок от 0 до 1.

Предлагаемая методика использования информационных характеристик в ходе факторного анализа, применяемого для отбора предикторов, предполагает проведение следующих операций:

- уяснение требований к прогностическому способу (прогнозируемый атмосферный параметр, заблаговременность, точность, физико-географический район, сезон года и т.д.);
- составление предварительного перечня предикторов;
- нормирование и центрирование значений предикторов для нивелирования их различной размерности;
- проведение обычной процедуры факторного анализа;
- вычисление количества информации, получаемой или теряемой при переходе от *n* исходных предикторов к *n* факторам;
- проведение аналогичных вычислений при переходе к (*n*-1), (*n*-2) и т.д. до одного оставляемого фактора включительно;
- определение оптимального числа факторов, которому соответствует максимальный выигрыш или минимальные потери исходной информации.

Полученное число факторов используется для построения прогностических способов, например, методами регрессионного или дискриминантного анализа.

Полученные результаты позволяют сделать определенные выводы.

1. Использование факторного анализа в целях отбора предикторов способно существенно повысить качество проведения указанной процедуры.
2. Для устранения существующей неопределенности при оценивании количества факторов, необходимого для адекватной замены исходного перечня признаков без существенной потери информации, предлагается использовать информационные показатели и, в частности, количество

информации, получаемой или теряемой при переходе от пространства исходных предикторов к факторному пространству.

3. Предложенный подход не противоречит требованиям руководящих документов, не требует существенных временных или материальных затрат и относительно легко может быть использован различными специалистами гидрометеорологических подразделений при разработке прогностических метеорологических способов.

METHODICAL ASPECTS OF THE NECESSARY NUMBER OF THE FACTORS DETERMINATION ON THE BASIS OF THE INFORMATION QUANTITY

I. I. Ulshin, V. S. Ivanov

The important direction of the enhancement of meteorological data provision of the different customers is development of new prognostic meteorological methods. The most difficult stage of similar development is selection of predictors. For these purposes use of the factor analysis is offered.

Factor analysis is intended for transition from the initial space of predictors to the factor space considerably smaller dimensionality without essential loss of the information. However it is characterized by existence of an essential shortcoming. It consists in uncertainty of the response to a question of the quantity of factors necessary for adequate changeover of the initial list of signs.

In article it is shown that for elimination of existing uncertainty in estimation of sufficient number of factors it is necessary to calculate the amount of information, received upon transition from initial space to the factor space. The technique of use of the information characteristics during the factor analysis applied to selection of the predictors is offered.

KEYWORDS: INFORMATION, THE FACTOR SPACE, FACTOR ANALYSIS, METEOROLOGY.

ЛИТЕРАТУРА

1. Шеннон К. Работы по теории информации и кибернетике. М.: Изд-во иностранной литературы, 1963. 832 с.
2. Голдман С. Теория информации. М.: Изд-во иностранной литературы, 1957. 446 с.