УДК 311:681.3.06

Л.А. Хворова, Н.В. Гавриловская

Построение статистической модели прогноза урожайности яровой пшеницы методом главных компонент

Для построения статистических моделей в большинстве случаев используется аппарат классического регрессионного анализа. Уравнение регрессии переменной y по переменным $(x_1, x_2, \ldots, x_n)$ в матричной форме можно представить в виде

$$y = L^* X = Q^* C^{-1} X , \qquad (1)$$

где L – вектор коэффициентов регрессии; X – вектор независимых величин (предикторов); Q – вектор, составленный из коэффициентов ковариации между предсказуемым y (предиктантом) и составляющими вектора X; C^{-1} – матрица, обратная ковариационной матрице предикторов; (*) – знак транспонирования. Предполагается, что все переменные предварительно центрированы [1, 2].

Важным обстоятельством, затрудняющим применение обычного регрессионного анализа, является сильная корреляция между переменными, описывающими метеорологические условия вегетационного периода.

Температура, влажность воздуха, фотосинтетически активная радиация, осадки, влагозапасы почвы, взятые в виде сумм или средних за отрезки вегетационного периода любой продолжительности, связаны между собой, и эта корреляция легко объяснима с физической точки зрения.

Кроме «синхронной» корреляции между различными параметрами за один и тот же отрезок времени, не менее существенно влияние «асинхронной» корреляции между параметрами, относящимися к разным временным интервалам. Например, корреляция между температурой воздуха или влагозапасами почвы за смежные декады, корреляция между температурой и дефицитом влажности за смежные декады и т.д. Эта корреляция вызывается инерцией метеорологических процессов, а также инерцией параметров, характеризующих состояние посевов и корнеобитаемого слоя почвы.

Мы считаем, что применение регрессионного анализа для исследования влияния метеорологических условий на формирование урожая и для построения соответствующих многомерных прогностических схем не может дать должного результата. В работе [2] предлагается использовать в этих случаях компонентный анализ или метод главных компонент.

Пусть $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ – собственные значения, матрицы C; $U_1, U_2, \ldots, U_n$ – соответствующие ортонормированные собственные векторы. Матрицу, составленную из собственных векторов $U_1, U_2, \ldots, U_n$, обозначим U, тогда

$$U^* U = U U^* = E_n , \qquad (2)$$

где E_n – единичная матрица порядка n.

Перейдем от исходного набора переменных $X^* = (x_1, x_2, \ldots, x_n)$ к новому набору переменных $A^* = (a_1, a_2, \ldots, a_n)$ посредством преобразования

$$A = UX . \qquad (3)$$

Переменные a_i называются главными компонентами переменных x. Дисперсии переменных a_i равны соответствующим собственным числам матрицы

$$\sigma^2_{a_i} = \lambda_i . \qquad (4)$$

Новые переменные обладают следующим экстремальным свойством. Дисперсия a_1 является максимально возможной дисперсией для любой переменной, представляющей собой линейную комбинацию вида

$$a_i = \sum_{i=1}^{n} u_i x_i \qquad (5)$$

при условии нормировки весов

$$\sum_{i=1}^{n} u_i^2 = 1 . \qquad (6)$$

Дисперсия a_k является максимальной среди всех комбинаций вида (5), не коррелированных с a_{k-1}, и т.д. Благодаря этому свойству главные компоненты являются в некотором смысле наилучшими линейными функциями для описания изменений случайного вектора X от реализации к реализации или, имея в виду нашу задачу, наилучшими линейными функциями для описания изменений условий произрастания сельскохозяйственных культур от года к году.

Первые q главных компонентов учитывают из полной вариации переменных $\sum_{i=1}^{n} \sigma^2_{x_i}$ долю, равную

$$\xi_q = \sum_{i=1}^{q} \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i , \qquad (6)$$

с возрастанием q ξ_q также растет и в случае сильной коррелированности факторов $x_1, x_2, \ldots, x_n$ быстро приближается к единице. Это дает возможность предположить, что разность $(1 - \xi_q)$, начиная с некоторого $q < n$, несущественна, что позволяет вместо исследования n коррелированных перемен-

ных заняться анализом в несколько раз меньшего числа q некоррелированных переменных. Применим этот метод для построения прогностических зависимостей.

От исходного набора предикторов перейдем к новому набору $a_1, a_2, \ldots, a_n$ предикторов – коэффициентов разложения случайного вектора X по собственным векторам его корреляционной матрицы. Отберем среди $a_1, a_2, \ldots, a_n$ несколько коэффициентов (p), наиболее информативных для прогнозирования y, и построим соответствующее уравнение регрессии

$$\hat{y} = l_1 a_1 + l_2 a_2 + \ldots + l_p a_p . \quad (7)$$

Для получения коэффициентов уравнения (7) нет необходимости непосредственно рассчитывать значения переменных $a_1, a_2, \ldots, a_p$. Коэффициенты $l_1, l_2, \ldots, l_p$ можно найти непосредственно по формуле:

$$l_i = \frac{1}{\lambda_i} Q^* U_i . \quad (8)$$

Соответствующий l_i парный коэффициент корреляции равен

$$r_{ya_i} = \frac{1}{\sqrt{\lambda_i}} Q^* U_i . \quad (9)$$

Множественный коэффициент корреляции, благодаря некоррелированности главных компонентов, определяется особенно просто:

$$R_{ya_1, a_2, \ldots, a_p} = \sqrt{\sum_{i=1}^{p} r_{ya_i}^2} . \quad (10)$$

Возвратимся к исходным переменным $x_1, x_2, \ldots, x_n$, тогда уравнение (10) следует записать так:

$$\hat{y} = \left(\sum_{i=1}^{p} \frac{1}{\lambda_i} Q^* U_i U_i^* \right) X . \quad (11)$$

Выражение в скобках представляет собой вектор коэффициентов регрессии, т.е.

$$\sum_{i=1}^{p} \frac{1}{\lambda_i} Q^* U_i U_i^* = L_p^* . \quad (12)$$

Компонентный анализ вместо одного «классического» уравнения регрессии позволяет на том же материале наблюдений построить, по крайней мере, n прогностических зависимостей. Действительно, учитывая лишь один наиболее информативный коэффициент разложения a_1, получаем

$$\hat{y} = \left(\frac{1}{\lambda_1} Q^* U_1 U_1^* \right) X = L_1^* X .$$

Учтем, кроме a_1, еще и a_2, получим

$$\hat{y} = \left(\frac{1}{\lambda_1} Q^* U_1 U_1^* + \frac{1}{\lambda_2} Q^* U_2 U_2^* \right) X = L_2^* X .$$

При введении последнего коэффициента разложения a_n уравнение примет вид

$$\hat{y} = \left(\frac{1}{\lambda_1} Q^* U_1 U_1^* + \frac{1}{\lambda_2} Q^* U_2 U_2^* + \ldots + \frac{1}{\lambda_n} Q^* U_n U_n^* \right) X = L_n^* X , \quad (13)$$

совпадающее с обычным уравнением регрессии (1), т.е. $L_p = L$ при $p = n$.

Уравнение регрессии (1) можно рассматривать как частный случай уравнения (11). Следовательно, суть преимуществ, которые может дать метод главных компонент по сравнению с обычным подходом, заключается в возможности отбрасывать часть коэффициентов разложения вектора-предиктора. Но какие из коэффициентов разложения $a_1, a_2, \ldots, a_n$ отбросить, а какие ввести в прогностическую зависимость? Предложено несколько принципов отбора коэффициентов разложения для построения прогностических зависимостей. Наиболее целесообразной является процедура отбора, основанная на ранжировании парных коэффициентов корреляции, связывающих y с $a_1, a_2, \ldots, a_n$ [2].

Для перехода от стандартизованного к естественному масштабу представления переменных необходимо каждый коэффициент регрессии умножить на отношение σ_y / σ_{x_i}, а свободный член рассчитать по формуле

$$l_0' = \bar{y} - \sum_{i=1}^{p} \frac{\sigma_y}{\sigma_{x_i}} l_i' , \quad (14)$$

где $l_i' = L_p \cdot \bar{x}_i$.

Выполнив эти операции, получаем окончательное уравнение для прогноза урожайности.

Рассмотренный вариант многомерного регрессионного анализа – метод главных компонент, или метод разложения по «естественным» ортогональным составляющим – уже используется для решения агрометеорологических задач [1, 2].

Проведем прогноз урожайности яровой пшеницы по статистической модели на основе метода главных компонент. Все вычисления производились в математическом пакете SCILAB, используя соответствующие функции.

На основании данных о количестве осадков за осенний период, количестве осадков за зимний период, сумме температур >5 ° С за первые две декады вегетационного периода, количестве осадков за две декады, количестве дней с осадками за две декады, числе Вольфа, урожайности рассчитаем корреляционную матрицу C и вектор Q (исходные данные предварительно центрируем):

$$C = \begin{pmatrix} 1 & -0{,}267 & 0{,}136 & 0{,}049 & 0{,}016 & 0{,}099 \\ -0{,}267 & 1 & 0{,}003 & 0{,}286 & 0{,}147 & 0{,}112 \\ 0{,}136 & 0{,}003 & 1 & -0{,}449 & -0{,}273 & 0{,}318 \\ 0{,}049 & 0{,}286 & -{,}0449 & 1 & 0{,}541 & -0{,}304 \\ 0{,}016 & 0{,}147 & -0{,}273 & 0{,}541 & 1 & -0{,}102 \\ 0{,}098 & 0{,}112 & 0{,}318 & -0{,}304 & -0{,}102 & 1 \end{pmatrix},$$

$$Q=\begin{pmatrix}0,317\\0,324\\-0,150\\0,555\\0,203\\0,019\end{pmatrix}.$$

Элементы матрицы C и вектора Q составлены из коэффициентов корреляции. Найдем собственные значения матрицы C, используя функцию «princomp()» в математическом пакете SCILAB, получим

$$U_1=\begin{pmatrix}-0,1116\\0,2141\\-0,4859\\0,5945\\0,4826\\-0,3454\end{pmatrix},\ U_2=\begin{pmatrix}0,5411\\-0,7348\\-0,1865\\0,0135\\-0,0135\\-0,3635\end{pmatrix},\ U_3=\begin{pmatrix}-0,6466\\-0,1546\\-0,2698\\-0,2459\\-0,4114\\-0,5055\end{pmatrix},$$

$$U_4=\begin{pmatrix}0,2888\\0,4008\\0,5017\\0,1933\\-0,2876\\-0,6198\end{pmatrix},\ U_5=\begin{pmatrix}0,3388\\0,2952\\-0,5679\\0,1412\\-0,6353\\0,2279\end{pmatrix},\ U_6=\begin{pmatrix}0,2801\\0,3776\\-0,2864\\-0,7271\\0,3336\\-0,2387\end{pmatrix}.$$

Собственные числа соответственно равны $\lambda_1=2,0763$, $\lambda_2=1,2503$, $\lambda_3=1,0993$, $\lambda_4=0,6710$, $\lambda_5=0,5957$, $\lambda_6=0,3075$. Сумма собственных чисел равна шести, т.е. сумме диагональных элементов матрицы C.

По формуле (9) вычисляем коэффициенты корреляции между y и коэффициентами разложения $a_1,a_2,\ldots,a_n$:

$r_{ya_1}=0,3669$, $r_{ya_2}=-0,0366$, $r_{ya_3}=-0,4234$,

$r_{ya_4}=0,2229$, $r_{ya_5}=0,3145$, $r_{ya_6}=-0,1558$.

Выберем наиболее информативные коэффициенты разложения для построения зависимости. Для этого рассчитаем $\xi_q=\sum_{i=1}^{q}\lambda_i\Big/\sum_{i=1}^{n}\lambda_i$, выбирая признаки с самыми высокими коэффициентами корреляции, получим

$$\xi_{1345}=\frac{\lambda_1+\lambda_3+\lambda_4+\lambda_5}{\sum_{i=1}^{n}\lambda_i}=\frac{4,44}{6}\approx 74.$$

Первый, третий, четвертый и пятый коэффициенты разложения несут около 74% всей информации о независимых переменных.

Для того чтобы получить уравнение регрессии, связывающее y с коэффициентами a_1, a_3, a_4 и a_5, используем формулу (8); получим:

$$l_1=0,2548,\ l_3=-0,4039,$$

$$l_4=0,2721,\ l_5=0,4074.$$

Уравнение регрессии запишется в виде

$$y=0,2548\cdot l_1-0,4039\cdot l_3+0,2721\cdot l_4+0,4074\cdot l_5.$$

Свободный член этого уравнения равен нулю, поскольку $\overline{a}_1=0$, $\overline{a}_3=0$, $\overline{a}_4=0$ и $\overline{a}_5=0$, а y представлен в стандартизованном масштабе, т.е. $\overline{y}=0$, $\sigma=1$. Для перехода от переменных $a_1,a_2,\ldots,a_p$ к $x_1,x_2,\ldots,x_n$ подставим в это уравнение выражение для коэффициентов разложения $a_i=U_i^*X$ и рассчитаем вектор коэффициентов регрессии $L_i=l_i\cdot U_i$.

$$L_1=0,2548\cdot\begin{pmatrix}-0,1116\\0,2141\\-0,4859\\0,5945\\0,4826\\-0,3454\end{pmatrix}=\begin{pmatrix}-0,0284\\0,0545\\-0,1237\\0,1514\\0,1229\\-0,0880\end{pmatrix},$$

$$L_3=-0,4039\cdot\begin{pmatrix}-0,6466\\-0,1546\\-0,2698\\-0,2459\\-0,4114\\-0,5055\end{pmatrix}=\begin{pmatrix}0,2611\\0,0624\\0,1090\\0,0993\\0,1662\\0,2041\end{pmatrix},$$

$$L_4=0,2721\cdot\begin{pmatrix}0,2888\\0,4008\\0,5017\\0,1933\\-0,2876\\-0,6198\end{pmatrix}=\begin{pmatrix}0,0786\\0,1091\\0,1365\\0,0526\\-0,0782\\-0,1687\end{pmatrix},$$

$$L_5=0,4074\cdot\begin{pmatrix}0,3388\\0,2952\\-0,5679\\0,1412\\-0,6353\\0,2279\end{pmatrix}=\begin{pmatrix}0,1381\\0,1203\\-0,2314\\0,0576\\-0,2589\\0,0928\end{pmatrix}.$$

Сложив $L_1+L_3+L_4+L_5$, в итоге получаем уравнение регрессии

$$y=0,4494x_1+0,3463x_2-0,1096x_3+$$
$$+0,3609x_4-0,0481x_5+0,0403x_6.$$

Для перехода от стандартизованного к естественному масштабу представления переменных необходимо каждый коэффициент регрессии умножить на отношение $\sigma_y\big/\sigma_{x_i}$, а свободный член рассчитать по формуле (14), где средние квадратиче

Результаты прогноза урожайности

Годы	Урожайность (фактическая)	Урожайность (расчетная)	Ошибка
1971	21,3	22,22	0,04
1972	28,6	22,72	0,21
1973	17,5	16,36	0,06
1974	11,3	16,10	0,42
1975	20,3	21,15	0,04
1976	13,7	13,66	0,00
1977	20,2	17,66	0,13
1978	20,4	18,66	0,09
1979	20,1	20,82	0,04
1980	22,8	16,44	0,28
1981	13,4	15,39	0,15
1982	16,3	20,79	0,28
1983	24	23,84	0,01
1984	23	21,77	0,05
1985	22,7	25,80	0,14
1986	25,3	23,42	0,07
1987	24,2	19,76	0,18
1988	15,9	17,90	0,13
1989	25,3	20,53	0,19
1990	16,3	21,35	0,31
1991	17,8	18,81	0,06
1992	17	16,77	0,01
1993	19	21,19	0,12
1994	20	19,53	0,02
1995	20	17,37	0,13
1996	15,2	17,08	0,12
1997	12,8	17,25	0,35

ские отклонения равны соответственно $\sigma_y = 4,3$; $\sigma_{x_1} = 23,29$; $\sigma_{x_2} = 27,37$; $\sigma_{x_3} = 43,06$; $\sigma_{x_4} = 16,04$; $\sigma_{x_5} = 3,66$; $\sigma_{x_6} = 52,41$, а средние значения – $\bar{y} = 19,42$; $\bar{x}_1 = 50,65$; $\bar{x}_2 = 112,34$; $\bar{x}_3 = 266,22$; $\bar{x}_4 = 26,04$; $\bar{x}_5 = 8,00$; $\bar{x}_6 = 69,19$.

Выполнив эти операции, получаем окончательное уравнение

$$y' = 0,083x_1' + 0,054x_2' - 0,011x_3' + 0,097x_4' - \\ -0,056x_5' + 0,003x_6' + 9,730. \quad (15)$$

Результаты прогноза урожайности по модели (15) представлены в таблице. Средняя относительная ошибка прогноза составила 13%.

Оценка существенности различий между средней фактической урожайностью и прогнозируемой показала, что фактическая и прогнозируемая урожайность не отличается статистически значимо.

Библиографический список

1. Дронов, С.В. Многомерный статистический анализ : учеб. пособие / С.В. Дронов. – Барнаул, 2006.

2. Сиротенко, О.Д. Математическое моделирование водно-теплового режима и продуктивности агроэкосистем / О.Д. Сиротенко. – Л., 1981.