УДК 311:681.3.06

Л.А. Хворова, Н.В. Гавриловская, Н.Н. Лопатин

Применение информационных технологий, математических методов и моделей для обработки и анализа многомерных данных

В статье рассматривается применение информационных технологий, математических методов и моделей для обработки и анализа многомерных статистических данных социально-экологических исследований и в сфере сельскохозяйственного производства, выявление и описание существующих между ними взаимосвязей, выделение причинных связей с целью их интерпретации и получения научных и практических выводов.

Исследования проводились по следующим направлениям: биология, медицина, экология и агрометеорология.

Объекты исследования – регионы Алтайского края (по медико-экологическим и агрометеорологическим показателям).

Объект «регионы Алтайского края» представлен следующими агроклиматическими показателями (табл. 1): суммой эффективных температур за вегетационный период (показатель STEMP) – три варианта расчета показателя, суммой осадков за вегетационный период (OSAD) – три варианта расчета показателя, количеством дней с осадками (KOLOS) – три варианта расчета; суммой осадков за зимний (ZIMA), весенний (APRIL) и осенний периоды (OCEN), урожайностью пшеницы (CROP). Цель исследования – определение года-аналога для прогноза урожайности зерновых культур. Задача прогноза урожайности зерновых культур относится к кругу тех научных и технологических задач, в которых и проблемы, и связанные с ними наборы данных являются многомерными. Так, база агрометеорологических данных, необходимых для прогноза и оценки урожайности по всей территории Алтайского края, содержит информацию с 1928 по 2004 г. Данная информация несет в себе «генетически» заложенную закономерность в природных явлениях и в различных сочетаниях факторов.

Задача *определения года-аналога* состоит в том, чтобы на основе различных сочетаний погодных и биологических факторов классифицировать текущий год, т.е. отнести его к одному из нескольких классов некоторым оптимальным способом для прогноза урожайности текущего года.

Сумма эффективных температур рассчитывалась тремя способами:

– при устойчивом переходе температуры через 10 °С;

– от первого до последнего дня с температурой больше 10 °С;

– за вегетационный период (от начала сева до уборки: в среднем с 15 мая до 1 сентября).

Определение года-аналога проводилось по следующей схеме:

I этап. Применение дисперсионного анализа для установления причинно-следственных отно-

Таблица 1

Результаты обработки агрометеорологических данных

Фактор	SS		Total	Fcrit	BG/T	Fэксп
	BG	WG				
STEMP1	55836555	1105085	56941640	4.026631	0.98059	50.5269
STEMP2	66948297	812383.7	67760680	4.026631	0.98801	82.4097
STEMP3	44156969	269734.9	44426704	4.026631	0.99392	163.705
OSAD1	430622.5	89566.93	520189.4	4.026631	0.82781	4.80782
OSAD2	421862.5	78180.3	500042.8	4.026631	0.84365	5.39602
OSAD3	310325.2	75952.63	386277.8	4.026631	0.80337	4.08577
KOLOS1	1699.045	1699.604	3398.648	4.026631	0.49991	1.00032
KOLOS2	1545.615	1544.048	3089.663	4.026631	0.50025	1.00101
KOLOS3	461.7113	1351.974	1813.685	4.026631	0.25457	2.92818
OCEN	37921.5	21316.13	59237.63	4.026631	0.64015	1.77900
ZIMA	53279.53	12126.25	65405.78	4.026631	0.8146	4.39373
APRIL	187090	64684.57	251774.6	4.026631	0.74308	2.89234
CROP	0	1027.948	1027.948	4.026631	0	0

Обозначения: *BG* – межгрупповая дисперсия; *WG* – внутригрупповая дисперсия; *Total* – общая дисперсия; *Fcrit* – табличное значение F-критерия; *Fэксп* – экспериментальное значение.

шений между признаками, характеризующими каждый объект.

Дисперсионный анализ [1–3] в данном случае был использован для первичной обработки экспериментальных данных, а именно для уменьшения числа факторов-признаков объектов (по наличию и силе связи). С помощью F-критерия была произведена оценка влияния на результативный признак многочисленных факторов. Заключительным этапом применения дисперсионного анализа была оценка силы влияния отдельных факторов на результативный признак.

После обработки агрометеорологических данных дисперсионным анализом (табл. 1) по значимости и силе влияния были отобраны следующие факторы: STEMP2, OSAD2, ZIMA, CROP.

В результате применения этих процедур исходная совокупность объектов разделяется на кластеры или группы (классы) схожих между собой объектов. Под кластером понимают группу объектов, обладающую свойством плотности (плотность внутри кластера выше, чем вне его), дисперсией, отделимостью от других кластеров, формой, размером.

II этап исследования. Цель его состоит в том, чтобы на основе измерения различных характеристик объекта классифицировать его, т.е. отнести к одному из нескольких классов некоторым оптимальным способом.

Для этой цели был использован кластерный анализ в связи с отсутствием обучающих выборок. Процедуру кластеризации проводили несколько раз с помощью программы статистической обработки данных Statistica [4] при различных значениях числа кластеров (3–7), после чего выбирали лучшую группировку в смысле критерия минимума отношений средних внутрикластерных и межкластерных расстояний (табл. 2):

$$F = \frac{d_w / f_w}{d_b / f_b},$$

где d_w, d_b – суммы внутрикластерных и межкластерных расстояний; f_w, f_b – количество внутрикластерных и межкластерных расстояний.

Лучший вариант – разбиение на пять кластеров (рис. 1).

Зависимость отношения
средних внутрикластерных и межкластерных
расстояний от количества кластеров

Так, например, годами-аналогами 1997 г. (соответствует номеру 27) могут служить следующие годы: 1974 (4), 1976 (6), 1981 (11), 1982 (12), 1988 (18). Классификация была проведена дважды: с учетом данных по 1997 г. и без него. Результаты показали, что 1997 г. при двух классификациях относится к одному и тому же первому кластеру.

Данные оформлены в таблицах 3 и 4, где указаны диапазон изменения соответствующих параметров, их средняя величина, годы, отнесенные к определенному кластеру. Цвет клеток характеризует заложенную в кластере закономерность.

Так, например, в таблице 3 первый кластер характеризуется минимальными осадками зимой и максимальными температурами в вегетационный период и, как следствие, – низкая урожай-

Таблица 2

Результаты обработки кластерным анализом

STEMP 2, OSAD 2, ZIMA, CROP							
Количество кластеров	d(w)	f(w)	d(b)	f(b)	d(w)/f(w)	d(b)/f(b)	F
2	19.83079	27	0.971112	1	0.734474	0.971112	0.756322
3	17.22645	27	3.314948	3	0.638017	1.104983	0.5774
4	14.52467	27	7.902046	6	0.537951	1.317008	0.408464
5	13.51582	27	12.49379	10	0.500586	1.249379	0.400668
6	12.3116	27	18.72878	15	0.455985	1.248585	0.365201
7	11.27899	27	27.43389	21	0.41774	1.306376	0.31977
8	10.17189	27	39.70294	28	0.376737	1.417962	0.265689
9	9.346788	27	52.80908	36	0.346177	1.466919	0.235989
10	8.324905	27	63.88817	45	0.30833	1.419737	0.217174

Таблица 3

Разбиение на кластеры без 1997 г.

	Номер кластера				
	1	2	3	4	5
ГОДЫ	4,6,11,12,18	7,9,20,24,25	1,2,5,8,13,14,15,16,17,19	22,23,26	3,10,21
STEMP2	(2234,1; 2525,2) 2379,65	(2276,3; 2449) 2362,65	(1892; 2221,3) 2056,65	(2046,1; 2159,1) 2102,6	(2309,5; 2427,7) 2368,6
OSAD 2	(128; 189,3) 158,65	(217; 316,8) 266,9	(116,9; 251,4) 184,15	(239,6; 291,8) 265,7	(119,8; 215,5) 167,65
ZIMA	(35,3; 81,2) 58,25	(60,6; 79,1) 69,85	(58,7; 99,8) 79,25	(76,6; 88,5) 82,55	(115,7; 140,4) 128,05
CROP	(11,3; 16,3) 13,8	(16,3; 20,2) 18,25	(20,3; 30,3) 25,3	(15,2; 19) 17,1	(17,5; 22,8) 20,15

Таблица 4

Разбиение на кластеры с 1997 г.

	Номер кластера				
	1	2	3	4	5
ГОДЫ	4, 11, 12, 18, 27	7, 9, 20, 23, 24, 25	6, 22, 26	1, 2, 5, 8, 13, 14, 15, 16, 17, 19	3, 10, 21
STEMP 2	(2286,3; 2704,9) 2495,6	(2159,1; 2449) 2304,05	2046,1; 2234,1) 2140,1	(1892; 2221,3) 2056,65	(2309,5; 2427,7) 2368,6
OSAD 2	(128; 189,3) 158,65	(217; 316,8) 266,9	(173; 241,3) 207,15	(116,9; 251,4) 184,15	(119,8; 215,5) 167,65
ZIMA	(35,3; 95,3) 65,3	(60,6; 88,5) 74,55	(76,6; 87,5) 82,05	(58,7; 99,8) 79,25	(115,7; 140,4) 128,05
CROP	(11,3; 16,3) 13,8	(16,3; 20,2) 18,25	(13,7; 17) 15,35	(20,3; 30,3) 25,3	(17,5; 22,8) 20,15

ность. Во второй кластер попали годы со средними значениями всех показателей. Третий кластер характеризуется средними осадками и средней температурой за вегетационный период, и как следствие – высокая урожайность.

Таким образом, результатом работы является построение классификатора для определения года-аналога на каждом этапе предварительного прогноза урожайности. Результаты классификации показывают, что группировка по годам соответствует закономерностям, которые проявляются при определенном сочетании факторов.

Прогноз урожайности также осуществляется в несколько этапов.

1. Предварительный прогноз осенью.

Расчет теплообеспеченности вегетационного периода следующего года осуществляется по тригонометрическому тренду:

$$TEMP_j = \sum_{i=1}^{(m/2)-1}\left[a_{2i-1}\cdot\cos\left(\frac{2\pi i}{m}\cdot j\right)+a_{2i}\cdot\sin\left(\frac{2\pi i}{m}\cdot j\right)\right]+ \\ +a_{m-1}\cdot(-1)^j+a_m,$$

где $TEMP_j$ – рассчитываемая сумма эффективных температур; T_j – ряд сумм эффективных температур; T – длина временного ряда сумм эффективных температур; m – длина цикла; a_j – коэффициенты тригонометрического тренда, рассчитываемые по следующим формулам:

$$a_{2i-1}=\frac{2}{T}\cdot\sum_{j=1}^{T}\left(T_j\cdot\cos\left(\frac{2\pi i}{m}\cdot j\right)\right), i=1,\ldots,\frac{m}{2}-1;$$

$$a_{2i}=\frac{2}{T}\cdot\sum_{j=1}^{T}\left(T_j\cdot\sin\left(\frac{2\pi i}{m}\cdot j\right)\right), i=1,\ldots,\frac{m}{2}-1;$$

$$a_{m-1}=\frac{1}{T}\cdot\sum_{j=1}^{T}\left((-1)^j\cdot T_j\right);\ a_m=\frac{1}{T}\cdot\sum_{j=1}^{T}T_j.$$

Возможность использования данной зависимости основана на цикличности солнечной активности.

По году-аналогу определяется возможный диапазон изменения суммы осадков.

По эмпирической модели ожидаемой урожайности осуществляется прогноз:

$$Y_{j+1}=\begin{cases}Y_{\min}+\left(Y_j-Y_{\min}\right)\cdot H(P)\cdot H(T)\cdot H(N), если Y_j\ge\overline{Y};\\ Y_{\min}+\left(Y_{\max}-Y_j\right)\cdot H(P)\cdot H(T)\cdot H(N), если Y_j<\overline{Y},\end{cases}$$

где Y_{j+1} – урожайность текущего года (ожидаемая); Y_j – урожайность предыдущего года; $Y_{\max}$ – максимальная; $Y_{\min}$ – минимальная; $\overline{Y}$ – средняя урожайности по всему временному ряду урожайностей; P – сумма осадков за вегетационный период с пороговыми значениями функции отклика p_1, p_2, p_3, p_4, – сумма эффективных температур за вегетационный период с пороговы-

ми значениями функции отклика t_1, t_2, t_3, t_4; N – число дней с осадками за вегетационный период с пороговыми значениями функции отклика n_1, n_2, n_3, n_4; $H(P), H(T), H(N)$ – нормированные функции отклика.

2. Уточняющий прогноз урожайности весной.

Используется метод прогноза теплообеспеченности, разработанный Ф.Ф. Давитая, основанный на связи сумм активных температур с датой весеннего перехода средней суточной температуры воздуха через 10 °С. Эта связь установлена в результате обработки данных многолетних метеорологических наблюдений основных метеостанций. Уравнение связи имеет вид

$$\sum t = K_1 \cdot D + K_2 ,$$

где $\sum t$ – сумма температур за период со средней суточной температурой воздуха более 10 °С; D – дата весеннего перехода температуры через 10 °С, выраженная числом дней от 1 апреля. Находится год-аналог, диапазон изменения суммы осадков, урожайность по году-аналогу или формуле (*).

3. Окончательный прогноз летом текущего года.

Осуществляется по ежесуточным данным уточненного года-аналога по динамической модели продуктивности сельскохозяйственных культур EPIC (Erosion-Productivity Impact Calculator, Департамент сельского хозяйства Соединенных Штатов, Сельскохозяйственная научно-исследовательская служба).

Для объекта «регионы Алтайского края» собрана также медико-экологическая информация со следующими показателями:

1. Общая заболеваемость (всего заболеваний) на 100 тыс. населения.

2. Общая заболеваемость (инфекционные и паразитарные болезни).

3. Общая заболеваемость (болезни органов дыхания) на 100 тыс. населения.

4. Общая заболеваемость (болезни эндокринной системы) на 100 тыс. населения.

5. Общая заболеваемость (болезни костно-мышечной системы) на 100 тыс. населения.

6. Общая заболеваемость (болезни системы кровообращения) на 100 тыс. населения.

7. Общая заболеваемость (болезни мочеполовой системы) на 100 тыс. населения.

8. Общая заболеваемость (болезни нервной системы) на 100 тыс. населения.

9. Общая заболеваемость (болезни органов пищеварения) на 100 тыс. населения.

10. Общая заболеваемость (психические расстройства) на 100 тыс. населения.

11. Число больных, состоящих на учете (психические расстройства) на 1000 населения.

12. Общая заболеваемость населения (шизофрения) на 100 тыс. населения.

13. Общая заболеваемость населения (психические расстройства непсихотического характера) на 100 тыс. населения.

14. Число больных активным туберкулезом, состоящих на диспансерном учете (на 100 тыс. населения).

15. Число больных с впервые в жизни установленным диагнозом активного туберкулеза (на 100 тыс. населения).

16. Обеспеченность населения врачами на 10 тыс. населения.

17. Заболеваемость алкоголизмом, наркоманиями и токсикоманиями (на 100 тыс. населения).

18. Общее число абортов (включая мини-аборты) на 1000 женщин.

19. Среднее число детей, находящихся в школах-интернатах и детских домах (абс. числа).

Всего более 100 показателей. Цель исследования – анализ медико-экологической обстановки в крае с учетом социальных факторов; прогноз уровня заболеваемости по районам и краю в целом.

В настоящее время проведена классификация регионов по психическим заболеваниям и различным показателям заболевания туберкулезом (табл. 5).

Лучшая классификация объектов наблюдается при разбиении на три кластера. Выделились районы с интенсивными показателями заболевания и благоприятной эколого-медицинской обстановкой.

В результате исследования была проведена оценка социально-экологической ситуации в Алтайском крае. Для этого был применен метод кластерного анализа. В основу были взяты следующие социально значимые факторы: заболеваемость туберкулезом; смертность от туберкулеза; заболеваемость алкоголизмом, наркоманиями и токсикоманиями и заболеваемость сифилисом. При разделении на три группы четко выделились районы с высокими показателями данных факторов (8), со средними показателями (13) и районы с благоприятными значениями факторов (45). При сравнении с ранее проведенным кластерным анализом по такому социальному фактору, как психические расстройства (в том числе психозы, шизофрения, расстройства непсихотического характера), определились некоторые закономерности, а именно: большое количество совпадений районов с благоприятной обстановкой по всем факторам. С другой стороны, среди районов с высокими неблагоприятными показателями по всем социальным факторам, четко выделились три района (в порядке убывания интенсивности фактора): Тальменский, Кал-

Таблица 5

Разбиение на кластеры районов Алтайского края по интенсивности заболеваний

№	Районы	Заболеваемость (средний)				
		A	B	C	D	E
КЛАСТЕР №1						
2	Белокуриха	231,65	79,00	23,48	1908,12	319,56
4	Заринск	286,64	73,77	16,61	1987,12	423,98
7	Алтайский	324,28	77,18	23,95	1250,23	195,50
9	Баевский	271,53	53,49	20,96	2213,12	117,72
10	Бийский	358,73	61,56	20,02	860,52	328,48
11	Благовещенский	251,73	61,06	10,53	1017,03	127,96
12	Бурлинский	208,89	65,98	18,77	2456,30	100,76
14	Волчихинский	268,41	77,94	14,20	745,60	131,08
15	Ельцовский	194,13	62,64	34,73	2225,57	244,54
16	Егорьевский	319,88	64,53	18,81	973,73	278,28
19	Заринский	229,30	49,41	21,42	640,12	304,88
24	Ключевский	212,48	56,27	11,66	1699,57	147,36
25	Косихинский	319,00	84,26	17,81	1398,97	241,72
26	Красногорский	343,30	79,79	20,67	1422,18	191,60
27	Краснощековский	343,08	53,44	17,90	1394,60	146,24
28	Крутихинский	275,64	82,13	16,56	1080,58	136,78
29	Кулундинский	207,68	69,75	21,99	1894,65	215,46
30	Курьинский	241,87	79,25	29,92	1909,75	174,12
31	Кытмановский	292,58	65,54	18,99	1328,62	216,38
33	Мамонтовский	239,09	50,77	13,69	1104,98	192,30
34	Михайловский	332,21	56,62	13,55	1390,15	100,60
35	Немецкий	169,67	76,68	8,71	230,75	132,22
36	Новичихинский	263,32	65,03	16,26	1252,43	146,00
37	Павловский	292,94	88,70	27,04	1857,48	239,18
38	Панкрушихинский	388,05	101,94	18,64	760,25	92,98
40	Петропавловский	242,28	57,12	22,62	2618,45	150,20
41	Поспелихинский	210,77	66,69	22,22	2066,72	229,66
42	Ребрихинский	280,95	73,16	22,85	1760,20	247,16
43	Родинский	206,55	61,32	13,29	1453,82	196,54
44	Романовский	227,60	67,85	14,38	984,12	129,32
45	Рубцовский	257,59	73,71	25,78	1003,87	199,22
46	Славгородский	298,89	72,47	27,03	2030,48	134,60
47	Смоленский	250,93	65,66	25,18	2012,55	213,30
48	Советский	200,13	57,24	18,43	1332,82	251,38
49	Солтонский	230,54	61,48	18,93	1426,02	221,96
50	Солонешенский	226,12	61,58	16,25	2878,97	319,82
52	Табунский	159,21	48,34	19,75	755,47	193,16
54	Тогульский	321,98	62,43	24,32	1947,72	266,68
56	Третьяковский	295,22	51,08	16,81	2385,65	159,62
58	Тюменцевский	317,51	43,43	17,10	2311,21	186,08
59	Усть-Калманский	298,80	60,10	20,83	2005,57	131,86
60	Усть-Пристанский	216,84	65,63	15,73	2752,62	260,38
62	Хабарский	215,98	53,94	17,84	1761,63	43,88
63	Целинный	270,38	68,03	22,59	682,60	206,24
66	Шипуновский	351,82	77,71	20,67	741,72	155,00
КЛАСТЕР №2						
5	Новоалтайск	347,26	118,78	230,56	2692,93	369,48
6	Рубцовск	291,88	140,68	39,14	2848,33	328,54
21	Зональный	425,59	97,43	32,03	1368,70	324,62
39	Первомайский	533,43	137,96	31,40	1650,23	268,40
51	Суетский	422,27	104,90	11,96	495,00	692,04
53	Тальменский	454,12	119,93	44,48	1829,77	338,82
64	Чарышский	192,05	56,23	68,02	1194,78	1182,70
65	Шелаболихинский	401,11	102,18	43,03	810,95	271,62

Продолжение таблицы 5

№	Районы	Заболеваемость (средний)				
		A	B	C	D	E
КЛАСТЕР №3						
1	Барнаул	307,77	95,05	31,42	2650,72	413,20
3	Бийск	396,73	99,85	31,07	3109,62	242,76
8	Алейский	368,24	70,90	20,44	2308,07	231,34
13	Быстроистокский	336,74	48,93	19,58	3106,33	263,08
17	Заньяловский	342,49	76,78	19,83	4956,43	157,96
18	Залесовский	358,43	112,08	23,23	6208,38	547,12
20	Змеиногорский	537,77	103,35	26,54	2393,87	175,54
22	Калманский	400,50	81,65	23,35	4474,23	193,82
23	Каменский	297,31	78,08	24,61	2578,77	260,16
32	Локтевский	321,13	83,35	24,27	2430,55	184,08
55	Топчихинский	407,91	79,19	26,03	1717,97	267,70
58	Тюменцевский	317,51	43,43	17,10	2311,21	186,08
61	Угловский	424,41	95,06	29,52	2923,88	155,14

A – число больных туберкулезом, состоящих на диспансерном учете; *B* – число больных с впервые в жизни установленным диагнозом активного туберкулеза; *C* – смертность от туберкулеза; *D* – заболеваемость алкоголизмом; *E* – число больных сифилисом.

манский и Угловский. Возможно, причиной такого разделения является неблагополучная экологическая обстановка в этих районах, которая, по данным ИВЭП СО РАН, оценивается как «критическая» (Тальменский район) и «напряженная» (Калманский и Угловский) (Материалы карты экологических ситуаций Алтайского края, ИВЭП СО РАН, 1995). С другой стороны большое негативное влияние оказывает то, что данные районы входят в зону сильного атмосферного загрязнения, а Угловский и Калманский районы попадают в границы следа взрыва на Семипалатинском полигоне в 1949 г. Кроме того, в Угловском районе наблюдается загрязнение почв цезием 137, превышающее фон в 1,5 и более раз (Материалы ландшафтно-экологической карты Алтайского края, ИВЭП СО РАН, 1993).

Таким образом, кластерный анализ позволил выявить районы с наиболее неблагоприятной социально-экологической обстановкой, на основе чего необходима разработка специальных рекомендаций по сохранению и повышению потенциала здоровья населения Алтайского края.

Исследования в области агрометеорологии, медицины и экологии продолжены с использованием геоинформационных технологий (ArcView), которые позволят выявить с учетом пространственных особенностей объектов наиболее существенно влияющие факторы на процессы заболевания, тенденцию изменения взаимосвязей между факторами, получить более полную информацию о причинно-следственных связях, что в свою очередь позволит управлять ситуацией и вместе со специалистами выдавать рекомендации.

Литература

1. Айвазян С.А. Прикладная статистика и основы эконометрики: Учебник для вузов / С.А. Айвазян, В.С. Мхитарян. М., 1998.

2. Шмидт В.М. Математические методы в ботанике: Учеб. пособие. Л., 1984.

3. Боровиков В.П. STATISTICA – Статистический анализ и обработка данных в среде Windows / В.П. Боровиков, И.П. Боровиков. 2-е изд., стереотип. М., 1998.

4. Боровиков В.П. Популярное введение в программу STATISTICA. М., 1998.