УДК 519.2+681.3

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ И МОДЕЛИРОВАНИЕ ЗАВИСИМОСТИ УРОЖАЙНОСТИ ЗЕРНОВЫХ ОТ ЗАТРАТ

Кацко Игорь Александрович
к.т.н., профессор

Кубанский государственный аграрный университет, Краснодар, Россия

В статье проводится сравнительный анализ применения методов интеллектуального анализа данных и регрессионного анализа данных на примере моделирования зависимости урожайности зерновых культур от затрат, по данным о деятельности сельхоз организаций Краснодарского края за 2006г. Рассмотрено два подхода: регрессии с разрывом, основывающейся на классическом варианте и реализованной в пакете Statistica; многопараметрической линейной регрессии, основанной на идеологии эволюционного программирования. На основе результатов анализа делается вывод о том, в анализе данных нельзя ограничиваться только одной точкой зрения – и регрессия в Statistica и аналогичные средства в системе PolyAnalyst взаимно дополняют друг друга в описании изучаемого процесса.

Ключевые слова: ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ МОДЕЛИРОВАНИЕ, ЗАВИСИМОСТЬ, УРОЖАЙНОСТЬ ЗЕРНОВЫХ, ЗАТРАТЫ

UDC 519.2+681.3

INTELLECTUAL ANALYSIS OF DATE AND MODELLING OF DEPENDENCE OF GRAINS YIELDING ON EXPENDITURES

Katsko Igor Alexandrovich
Cand. Tech. Sci., professor

Kuban State Agrarian University, Krasnodar, Russia

Comparative analysis of intellectual data analysis methods application and regressive data analysis on the example of modeling of grains yielding dependence on expenditures, by data on activities of agricultural organizations of Krasnodar region for 2006 is carried out in the article. Two approaches: regression with rupture, based on the classical variant and realized in the packet “Statistica”; multi parametric linear regression, based on the ideology of evolutional programming are carried out in the article. There was made a conclusion on the basis of the analysis results that it was impossible to be restricted only by one point of view – and regression in Statistica and analogous means in system PolyAnalyst mutually supplement each other in the description of investigating process.

Key words: INTELLECTIAL ANALYSIS OF DATA, MODELING, DEPENDENCE, GRAIN YIELDING, EXPENDITURES

Начало XXI века с точки зрения экономического анализа данных характеризуется интенсивным внедрением различных средств анализа данных – начиная от средств анализа в бизнес приложениях (Excel) и статистических пакетах (SAS, Statistica, SPSS и др.) до специализированных программ извлечения знаний из баз и хранилищ данных – Data Mining систем (например, PolyAnalyst, Deductor). Вместе с этим следует отметить – несмотря на рекламные акции с претензией на универсальность указанных выше средств анализа данных и все их плюсы – программные средства и были и остаются инструментарием анализа данных в руках специалиста.

Развитие сельского хозяйства (и других отраслей) требует получения адекватных познавательных моделей для решения задач принятия управленческих решений. Какой из подходов необходимо использовать? Для ответа на поставленный вопрос в настоящей работе рассматривается задача нахожде-

ния зависимости урожайности зерновых от ряда экономических факторов с использованием пакетов Statistica и PolyAnalyst, на примере сельскохозяйственных предприятий северной и центральной зоны Краснодарского края, которые характеризуются близкими свойствами почв, климата и метеоусловий. Следует отметить, что статистическая зависимость не позволяет установить причинность связи. Причинность в экономических исследованиях подтверждается только содержательно и может подкрепляться или не подкрепляться статистически. *Задача исследователя – найти аналитическую функцию, которая наилучшим образом описывает экспериментальные данные в соответствии с предполагаемой связью*.

Большое разнообразие реальных ситуаций служило стимулом эволюции регрессионного анализа, развитию метода в направлении снятия классических ограничений и распространению его принципов на новые явления и процессы. Многие новые веяния внедрялись в пакеты прикладных программ. В столь небольшом обзоре нереально полностью описать даже один пакет. Мы остановимся на кратком обзоре двух из них – Statistica 6.1 и PolyAnalyst 5.0. Обе системы содержат средства интеллектуального анализа данных.

Statistica в основном ориентируется на классические методы математической статистики, многомерного статистического анализа и нейронных сетей. Так, например, статистический модуль *Общая линейная модель*, является современным обобщением линейной регрессионной модели и позволяет включать в планы категориальные предикторные переменные наряду с непрерывными и многомерные зависимые переменные. Основная идеология методов многомерного статистического анализа используемых в системе Statistica сводится к использованию теории алгебраических инвариантов не изменяющихся при линейных преобразованиях (например, собственные значения, собственные вектора, определители, декомпозиция матриц, корреляция между переменными и т.д.).

В нашем случае рассматривалось 547 сельскохозяйственных предприятий из которых 169 принадлежат северной и центральной зоне Краснодарского края – основным производителям зерновых. Рассматривались следующие факторы: затраты на 1га - x_1 (тыс.руб.); оплата труда на 1га (тыс.руб.) - x_2; затраты на семена на 1га (тыс.руб.) - x_3; затраты на удобрения на 1 га (тыс.руб.) - x_4; затраты на ГСМ на 1 га (тыс.руб.) - x_5; амортизация (тыс.руб.) - x_6; урожайность ц/га - у.

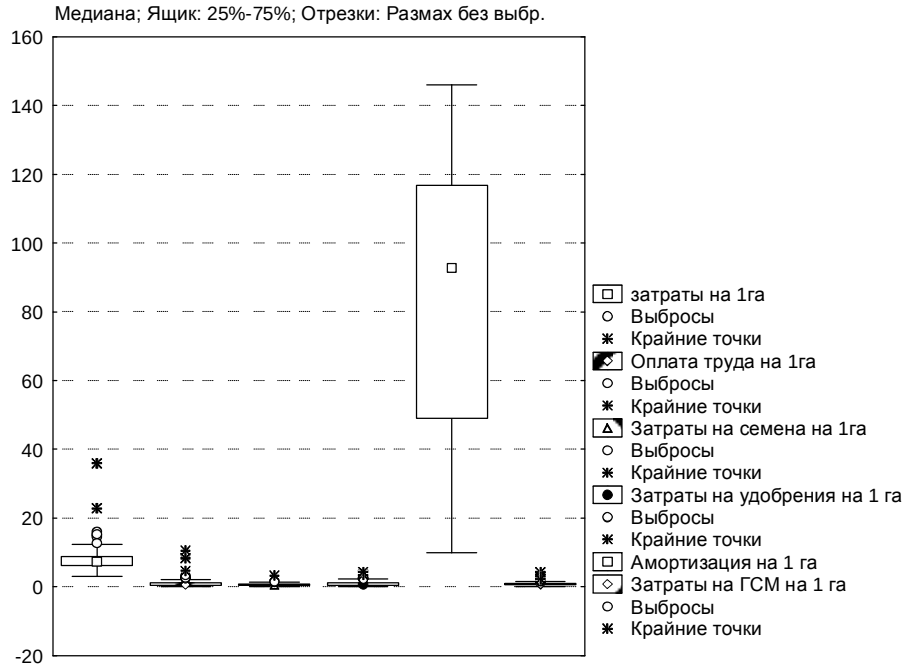

Рис. 1. Диаграмма размаха

Графическое изображение анализируемых данных в виде «ящика с усами» (рис. 1) показывает, что наибольший разброс имеют переменные амортизации на 1 га и затрат на 1га остальные переменные различаются незначительно.

Регрессионный анализ с использованием Statistica 6.0 показал, что линейная модель объясняет всего 16,2% вариации урожайности и кроме свободного члена и затрат на удобрения на 1 га других значимых переменных нет (табл.1).

После выбора опции *Кусочно-линейная регрессия* во вкладке стартовой панели модуля *Нелинейное оценивание*, *STATISTICA* производит оценивание по методу наименьших квадратов следующей модели:

$$y = (b_{01} + b_{11}*x_1 + ... + b_{m1}*x_m)*(y <= b_n) + (b_{02} + b_{12}*x_1 + ... + b_{m2}*x_m)*(y > b_n)$$

Таким образом, производится оценивание с использованием двух различных уравнений линейной регрессии; одно для значений *y*, которые меньше или равны точки разрыва (b_n) и одно для значений *y* больше точки разрыва.

Таблица 1. – Итоги анализа регрессионной зависимости урожайности от затрат.

N=165	R= .40294152 R2= .16236187 Скорректир. R2= .13055282 F(6,158)=5.1043 p<.00008 Станд. ошибка оценки: 10.703					
	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(158)	p-уров.
Св.член			35.29484	3.472056	10.16540	0.000000
затраты на 1га	0.049511	0.079294	0.17096	0.273795	0.62440	0.533268
Оплата труда на 1га	0.111457	0.082549	0.97442	0.721690	1.35019	0.178886
Затраты на семена на 1га	-0.037289	0.074332	-1.20772	2.407488	-0.50165	0.616611
Затраты на удобрения на 1 га	0.334587	0.076580	7.01291	1.605102	4.36914	0.000023
Амортизация на 1 га	0.082921	0.075001	0.02574	0.023285	1.10560	0.270580
Затраты на ГСМ на 1 га	0.133090	0.082694	2.92407	1.816837	1.60943	0.109518

Для оценки параметров модели использовалось несколько численных методов оптимизации (табл. 2).

Таблица 2. – Результаты моделирования с помощью кусочно-линейной регрессии.

Методы оптимизации	Точки разрыва	B0	Затраты на 1га, x_1	Оплата труда на 1га, x_2	Затраты на семена на 1га, x_3	Затраты на удобрения на 1 га, x_4	Затраты на ГСМ на 1 га, x_5	Амортизация, x_6	Объясненная доля дисперсии:	R
Квазиньютоновский	Y<=47.199	30.358	0.582	0.207	-1.694	1.726	-0.002	3.932	0.707	0.841
	Y>47.199	45.686	1.206	1.828	1.029	-0.225	-0.015	-0.885		
Хукадживиса	Y<=49.139	28.312	0.799	0.375	-1.998	2.252	0.022	3.347	0.702	0.838
	Y> 49.139	43.299	1.517	1.702	1.217	-0.119	-0.007	-0.538		
Хукадживиса и квазиньютоновский	Y<=49.139	28.312	0.799	0.375	-1.998	2.252	0.022	3.347	0.702	0.838
	Y> 49.139	43.299	1.517	1.702	1.217	-0.119	-0.007	-0.538		

Розен-брока	Y<=49.142	20.126	0.264	0.227	5.316	2.697	0.099	3.591	0.547	0.74
	Y> 49.142	26.990	3.384	1.461	-9.231	-0.428	0.081	0.884		
Розен-брока и ква-зиньюто-новский	Y<=49.053	28.320	0.799	0.375	-1.998	2.252	0.022	3.348	0.702	0.838
	Y> 49.053	43.309	1.517	1.702	1.221	-0.121	-0.007	-0.538		

Из таблицы 2 следует, что лучше всего - на 70,7% вариацию урожайности объясняет зависимость найденная с помощью квази-ньютоновского метода оптимизации (хотя и в этом случае практически все факторы не являются статистически значимыми). Анализ коэффициентов проводится стандартно. Результаты пошаговой регрессии, несмотря значимость факторов включенных в модель, объясняют всего 13% вариации урожайности (табл.3).

Таблица 3. – Итоги пошаговой регрессии

N=165	R= .37444946 R2= .14021240 Скорректир. R2= .12959773 F(2,162)=13.209 p<.00000 Станд. ошибка оценки: 10.709					
	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(162)	p-уров.
Св.член			37.91840	2.058233	18.42279	0.000000
Затраты на удобрения на 1 га	0.325294	0.072874	6.81813	1.527428	4.46380	0.000015
Затраты на ГСМ на 1 га	0.193716	0.072874	4.25606	1.601088	2.65823	0.008643

Таким образом, полученные с использованием пакета Statistica модели, неудовлетворительно описывают зависимость урожайности зерновых от затрат.

PolyAnalyst формулирует и проверяет гипотезы о виде регрессионной зависимости на внутреннем языке программирования с помощью функциональных примитивов (простейших программ). Результаты анализа представляются в виде понятном пользователю - таблиц, графиков и формул. Следует отметить, что авторы рассматривают методы линейной регрессии, поиска зависимостей в системе PolyAnalyst как дальнейшее развитие методов классического регрессионного анализа.

Следует отметить, что все описанные выше модули (как в принципе и все методы ИАД) используют классические статистические методы на этапах

поиска и оценки моделей, и оценки её адекватности [1]. Например, надёжность полученных результатов основывается на стандартном отклонении

$$s_{dev} = \sqrt{\sum_i (y_i - \hat{y}_i)^2 / (m-1)}$$

и стандартной ошибке

$$s_{err} = \sqrt{\sum_i (y_i - \hat{y}_i)^2 / ((n-1)s_y^4)},$$

где y_i – зависимая переменная; $\hat{y}_i$ – соответствующее значение, предсказанное моделью; n – число наблюдений; s_y^4 – квадрат дисперсии переменной y. Значимость найденной зависимости оценивается с помощью *индекса значимости*

$$I_z = -k \lg(s_{real} / s_{rand}).$$

Здесь s_{real} – стандартное отклонение, полученное на реальных данных; s_{rand} – стандартное отклонение случайных данных, в которых значение результативной переменной случайно перемешано для разных наблюдений, k=const. Считается, что результат моделирования значим, если значение I_z>2,0.

Стандартный подход к оценке значимости модели – коэффициент детерминации R squared (чем ближе он к единице, тем лучше модель).

Визуальный подход к оценке значимости модели заключается в изображении зависимости предсказанных значений ($y_{predicted}$), от реальных (y_{real}): чем ближе точки лежат к прямой $y_{predicted}=y_{real}$, тем точнее модель описывает данные.

Оценка значимости линейных регрессионных моделей основывается на известной статистике Фишера-Снедекора F-ratio: $\left[b_j / m_{b_j}\right]^2$, где b_j – j-й коэффициент модели; m_{b_j} – стандартное отклонение коэффициента, который лежит в

основе отбора наилучших переменных в уравнение регрессии (обычно переменная включается, если F-ratio>2,0).

Линейная регрессия в системе PolyAnalyst позволила найти зависимость:

Урожайность = +38.1469 +0.794892*"затраты на 1га" +2.68959*"Затраты на ГСМ на 1 га".

стандартная ошибка	0.9512
R-squared	0.09516

Поиск законов (в виде формул) позволил найти:

Лучшее по значимости правило:

Урожайность = (58.9852 *"Оплата труда на 1га"*"затраты на 1га"*"затраты на 1га"*if(NewVar,1 ,0.761038)+120.481 *"Оплата труда на 1га"*"затраты на 1га")/("Оплата труда на 1га"*"затраты на 1га"*"затраты на 1га"+25.472 *"Оплата труда на 1га"+0.0620379 *"затраты на 1га"*"затраты на 1га")

Лучшее по точности правило:

Урожайность = (59.7531 *"Оплата труда на 1га"*"Оплата труда на 1га"*"затраты на 1га"*"затраты на 1га"*if(NewVar,1 ,0.761038)+119.49 *"Оплата труда на 1га"*"Оплата труда на 1га"*"затраты на 1га"-1.51034 *if(NewVar,1 ,0.761038))/("Оплата труда на 1га"*"Оплата труда на 1га"*"затраты на 1га"*"затраты на 1га"+25.472 *"Оплата труда на 1га"*"Оплата труда на 1га"+0.0754276 *"Оплата труда на 1га"*"затраты на 1га"*"затраты на 1га"-0.11112).

Уровень	Стд.ош.	Стд.откл.	Значим..	R-squared
наиб. знач.	0.7519	10.58	> 100	0.4346
наиб. точн.	0.7454	10.49	> 100	0.4444

Получено несколько моделей. Какую следует выбрать? Простая линейная регрессия и пошаговая регрессии описывают дисперсию урожайности

всего на 16% и 14% соответственно. Кусочно-линейная регрессия описывает свыше 70% вариации урожайности, но практически все переменные не являются статистически значимыми – значим только свободный член. Линейная регрессия в системе PolyAnalyst возможно более приемлема хотя коэффициент корреляции очень мал, но значение индекса значимости 5,142 говорит о том, что модель достаточно хорошо описывает данные по случайной выборке (что соответствует идеологии бутстреп-метода). Наличие числовой информации о деятельности объекта предполагает эконометрический подход. Важнейшим моментом скептического отношения практиков к эконометрическим исследованиям является уверенность в том, что данные, которые являются основой моделирования часто содержат неточности, либо вообще фальсифицированы. Современная прикладная статистика рекомендует в этом случае обращаться к робастным методам - устойчивым к всевозможным ошибкам. Между тем для практика важна оценка зависимостей между факторами, возможность прогнозирования и управления, а не сам факт получения устойчивых моделей.

В рамках новой экономической парадигмы, сформировавшейся в мире после экономических кризисов, статистические результаты и измерения, полученные на предшествующих этапах развития региональной и мировой экономики не имеют научной силы, в связи с возможностью попадания в точки структурных изменений системы (точки бифуркаций). Поэтому оспариваются и традиционные способы прогнозирования и научные результаты полученные с помощью этих методов.

Таким образом, имеющиеся эконометрические методы, преимущественно ориентированные на вероятностную парадигму данных и имеющейся неопределенности, недостаточны для построения адекватных моделей функционирования и прогнозирования АПК.

Результаты оптимизация структурных (и других) параметров деятельности предприятия и всего АПК также часто не удовлетворяет практиков.

Одна из причин – предположение детерминируемости оптимизируемых переменных (в крайне редком случае – стохастичности). В силу этого исследователи настроены на разработку принципиально новой теории и методологии построения моделей функционирования и прогнозирования сельскохозяйственных предприятий в условиях данных не имеющих детерминированной или вероятностной природы, что подтверждает актуальность настоящей статьи. Однако приведенный выше анализ показал, что в анализе данных нельзя ограничиваться только одной точкой зрения – и регрессия в Statistica и аналогичные средства в системе PolyAnalyst взаимно дополняют друг друга в описании изучаемого процесса. Рассмотрение данных с двух альтернативных точек зрения позволяет лучшее проникнуть в суть проблемы. Таким образом «интеллектуализация» методов математической и прикладной статистики – это свершившийся факт, который можно и нужно использовать.

Потенциально с помощью этой модели (и вообще подобных моделей) можно прогнозировать урожайность в разные моменты времени. Даже если точный прогноз не всегда достижим, то хотя бы тенденцию к росту или спаду урожайности. Это необходимо для оценки государственными органами потенциальных возможностей сельского хозяйства. В последние годы сокращается количество многофакторных опытов, падает авторитет прикладной статистики и эконометрики. Между тем, невозможно прогнозировать деятельность АПК, изучать его ресурсы без математических моделей различного рода. Модели должны быть адаптивными и разными для разных зон с/х деятельности. Так полученные модели могут с успехом использоваться в северной и центральной зоне Краснодарского края, но сама идеология применения интеллектуального анализа для обработки данных конечно применима везде. Необходимость подобных моделей подтверждается опытом развития аграрной науки как у нас в стране, так и за рубежом.

Литература

1. Арсеньев С.Б. Извлечение знаний из медицинских баз данных. Москва, Мегапьютер. WEB: http://www.megaputer.ru/

2. Боровиков В.П. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688с.:илл.