УДК 681.513

Ю.В. КОППА, В.С. СТЕПАШКО

СРАВНЕНИЕ ПРОГНОЗИРУЮЩИХ СВОЙСТВ МОДЕЛЕЙ РЕГРЕССИОННОГО ТИПА И МГУА

На примерах применения АСТРИД для построения моделей объема производства легкой промышленности и процесса инфляции выполнено сравнение прогнозирующих свойств моделей, полученных по МНК и по МГУА. Показано, что МНК не позволяет построить модели, отражающие системные закономерности и пригодные для получения надежного прогноза, несмотря на то, что по статистическим оценкам модели оказались значимыми.

В статье выполнено сравнение прогнозирующих свойств регрессионных моделей и моделей метода группового учета аргументов (МГУА) на примерах задач моделирования объема производства легкой промышленности и процесса инфляции. Приведен также способ выбора аргументов, наиболее существенных (информативных) для имеющейся выборки данных, позволяющий существенно уменьшить вычислительные затраты благодаря исключению неинформативных аргументов из процесса моделирования.

Пример 1. Моделирование объема производства легкой промышленности.

Исходная таблица данных для моделирования объема производства легкой промышленности, взятая из [1], содержит 23 исходных аргумента (NA = 23):

X_1 - доходы населения (млрд.грн.), X_2 - индекс потребительских цен (% – 100), X_3 - индексы ВВП (% – 100), X_4 - объем производства промышленности, всего (млрд. грн.), X_5 - розничный товарооборот, всего (млрд. грн.), X_6 - индексы зарплаты (реальная, % – 100), X_7 - средние банковские процентные ставки на кредиты (% –100), X_8 - официальный курс гривны к доллару США (грн.), X_9 - налог на добавленную стоимость (млрд. грн.), X_{10} - налог на прибыль предприятий (млрд. грн.), X_{11} - расходы консолидированного бюджета, всего (млрд. грн.), X_{12} - расходы бюджета на народное хозяйство (млрд. грн.), X_{13} - индекс оптовых цен легкой промышленности (% – 100), X_{14} - занятость в легкой промышленности (млн. человек), X_{15} - общая занятость (млрд. человек), X_{16} - заработная плата (номинальная, грн.), X_{17} - учетная ставка для коммунальных банков (% – 100), X_{18} - денежная масса наличности в обращении (млрд. грн.), X_{19} - дебиторская задолженность (млрд. грн.), X_{20} - кредиторская задолженность (млрд. грн.), X_{21} - акцизный сбор (млн. грн.), X_{22} - налог на доход граждан (млн. грн.), X_{23} - доход консолидированного бюджета (млрд. грн.).

При большом числе исходных аргументов построение модели занимает очень много машинного времени. В то же время практика построения моделей показывает, что среди представленных аргументов много несущественных, которые можно без ущерба для результирующей модели исключить. Ниже рассматривается один из способов решения этой проблемы. Следует отметить, что разделение аргументов на существенные и несущественные справедливо в основном для конкретного набора данных.

В многорядном алгоритме МГУА к каждой i-й переменной оптимальным образом подбираются наилучший для данного случая ансамбль из остальных исходных аргументов [2]. Следо-

ISBN 966-02-0783-2 Моделирование и управление состоянием эколого-экономических систем региона. Киев, 2000.

вательно, можно построить структурную таблицу размером NA*NA, i-тая строка которой соответствует i-той модели, а j-тый столбец - j-тому аргументу (i,j = 1,..., NA), значение ij-я ячейки этой таблицы равно j, если j-й аргумент участвует в формировании i-й модели, или 0, если не участвует. Эта таблица показывает, какие исходные аргументы участвуют в формировании каждой i-й модели. Определим значение "индекса полезности" j-го аргумента $NumX_j$ как число, показывающее, сколько раз данный исходный аргумент участвует в формировании всех моделей (частота использования j-го аргумента). Ясно, что эти числа лежат в интервале от 1 до NA. Если $NumX_j = 1$, то можно утверждать, что j-й аргумент несуществен для данного случая, так как присутствует только в модели, где он включен насильно в соответствии со спецификой данного алгоритма отбора лучших моделей. Если NumXj = NA, то можно утверждать, что j-й аргумент существен для данного случая. Это крайние случаи, поэтому следует ввести некоторый порог для определения существенности j-го аргумента: будем считать аргумент существенным, если $NumX_j > 0.5*NA$. Кроме того, можно упорядочить аргументы по убыванию величины NumXj и, пользуясь дополнительными критериями, выбрать столько аргументов, сколько может быть допустимо в каждом конкретном случае.

По многорядному алгоритму была получена следующая структурная таблица:

1	2	0	4	0	6	0	8	0	10	11	0	13	0	0	0	0	0	0	20	0	0	0
1	2	3	4	0	6	0	0	0	10	0	12	0	14	0	16	17	0	0	0	0	0	0
1	2	3	4	0	6	7	0	0	10	0	12	0	14	0	16	17	0	0	20	21	0	0
1	2	0	4	0	6	0	0	0	0	11	12	13	14	0	0	17	18	19	0	0	0	0
1	2	3	4	5	6	7	0	0	0	11	12	13	14	0	16	17	0	19	0	0	0	0
1	0	0	4	0	6	0	8	0	10	11	12	13	14	0	16	0	0	0	20	0	0	0
1	0	3	4	0	6	7	8	0	10	11	12	13	0	0	0	17	0	19	0	21	22	0
1	0	0	4	5	6	7	8	0	10	11	12	13	14	0	16	17	18	19	0	0	0	0
1	0	3	4	0	6	7	8	9	10	11	12	13	14	0	16	17	0	19	0	0	0	0
1	2	0	4	0	6	7	0	0	10	0	0	0	14	0	16	17	0	19	20	0	0	0
1	2	3	4	0	6	7	0	0	10	11	12	0	0	0	16	0	0	19	20	0	0	0
1	0	3	4	0	6	0	8	0	10	11	12	13	14	0	16	17	0	19	0	0	22	0
1	2	0	4	5	6	7	0	0	0	11	0	13	14	0	0	17	18	19	0	0	0	0
1	0	3	4	0	6	7	0	0	10	0	12	13	14	0	16	17	0	0	0	0	0	0
1	2	3	4	0	6	7	8	0	10	11	0	13	14	15	16	17	18	19	0	0	22	0
1	2	0	4	0	6	0	8	0	10	11	0	13	14	0	16	17	0	19	0	0	0	0
1	0	0	4	0	6	7	8	0	10	11	12	13	14	15	16	17	0	19	20	0	0	0
0	2	0	4	0	6	7	0	0	0	11	0	0	14	0	0	17	18	0	20	0	0	0
1	0	0	4	0	6	7	8	0	10	11	12	13	14	15	16	17	0	19	0	21	0	0
1	0	0	4	0	6	7	8	0	10	11	12	13	14	15	16	17	0	19	20	21	22	0
1	2	3	4	0	6	7	8	0	10	11	12	13	0	0	16	17	0	19	0	21	0	0
1	2	0	4	0	6	7	0	0	0	11	12	0	14	0	16	0	18	0	0	0	22	0
1	2	0	4	0	6	0	0	0	10	0	0	0	0	0	16	17	0	19	20	0	0	23
22	**14**	**10**	**23**	**3**	**23**	**16**	**12**	**1**	**18**	**18**	**16**	**16**	**18**	**4**	**18**	**19**	**6**	**16**	**9**	**5**	**5**	**1**

Последняя строка таблицы показывает частоту участия каждого аргумента в формировании всех моделей. Если воспользоваться порогом $NumX_j > 0.5*NA$, получим, что для моделирова-

ния достаточен набор из следующих 13 аргументов: X_1, X_2, X_4, X_6, X_7, X_{10}, X_{11}, X_{12}, X_{13}, X_{15}, X_{16}, X_{17}, X_{19}.

Выходной величиной Y_1 в данной задаче является объем производства легкой промышленности (млрд.грн.). В [1] приведены значения вышеперечисленных аргументов и выходной величины за период времени с ноября 1995 по июнь 1997 гг., т.е. длина исходной выборки – 20 точек. Отметим, что показатели X_1, X_2 и X_6 характеризуют уровень жизни населения, X_7, X_{10}–X_{12} являются бюджетными, X_{13} – отраслевой, X_{17} – финансовый, а остальные являются макроэкономическими показателями.

В рассматриваемом ниже примере в качестве экзамена, т.е. для проверки прогнозных свойств модели, оставим последние 5 точек. Это связано с тем, что для оценки 14 коэффициентов (13 аргументов и свободный член) по МНК и для вычисления их статистических оценок требуется по крайней мере 15 точек.

Модель, полученная по МНК и включающая все тринадцать исходных аргументов, имеет вид (М1):

$$V_{1,t} = 1.0849 + 0.0315X_{1,t} + 0.0015X_{2,t} + 0.0683X_{4,t} + 0.0015X_{6,t} - 0{,}0144X_{7,t} + 0.1437X_{10,t} - 0.0064X_{11,t} + 0.0977X_{12,t} - 0.0046X_{13,t} - 0.0457X_{15,t} - 0.0111X_{16,t} + 0.0069X_{17,t} - 0.0019X_{19,t}, \quad (1)$$

где V_1 – модельная оценка выходной переменной Y_1, $X_{i,t}$ - значение i-го аргумента в t-й точке (i = 1,2,3,4,6,7,10,11,12,13,15,16,17,19; t = 1,...,20).

Модель имеет следующие характеристики:

$$\text{СКО} = 0.03168; \quad R = \max_{t=1,20} |V_{1,t} - Y_{1,t}| = 0.1102; \quad S = R/(Y_{1,\max} - Y_{1,\min}) * 100\% = 118.5\%,$$

где СКО – величина среднеквадратичного отклонения оценок V_t от табличных значений $Y_{1,t}$, R – наибольшая абсолютная ошибка, S – относительная максимальная ошибка в процентах от наибольшего “размаха” значений $Y_{1,t}$.

Статистические оценки модели: $R^2 = 0.984$; $R^2_y = 0.782$ (R^2 - коэффициент детерминации, R^2_y - скорректированный коэффициент детерминации).

Значения t-статистик t_i : $t_0 = 0.388$; $t_1 = 0.631$; $t_2 = 0.228$, $t_4 = 1.948$; $t_6 = 0.676$; $t_7 = 0.681$, $t_{10} = 0.502$; $t_{11} = 0.262$, $t_{12} = 0.784$, $t_{13} = 0.307$, $t_{15} = 0.339$, $t_{16} = 0.666$, $t_{17} = 0.657$; $t_{19} = 0.386$. Табличное значение $t_{(0.05,15)} = 1.753$.

В таблице 1 приведены оценки объема производства легкой промышленности М1, полученные по модели (1), на рис. 1 представлен график их изменения. Видно, что модель довольно плохая, причем наибольшие ошибки относятся к экзаменационным точкам. Несмотря на то, что только один коэффициент оказался значимым ($t_4 > 1.753$), приравнять нулю остальные коэффициенты без дополнительных исследований нельзя [3]. Таким образом, МНК не позволяет построить модель, отражающую системные закономерности и пригодную для получения надежного прогноза, хотя по статистическим оценкам модель оказалась значимой.

Модель по МГУА, построенная при тех же условиях, имеет вид (М2):

$$V_{1,t} = -0.1425 + 0.0046X_{2,t} + 0.0491X_{4,t} + 0.0022X_{6,t} - 0.0826X_{10t} + 0.0044X_{13,t}. \quad (2)$$

Статистические оценки модели:

$R^2 = 0.958$; $R^2_y = 0.934$; $F = 40.82$; $m1 = 5$; $m2 = 9$; $F_{(0.05,5,9)} = 3.48$,

где F, $F_{(\alpha,m1,\ m2)}$ – расчетное и табличное значения критерия Фишера с уровнем значимости α и степенями свободы m1 и m2 .

Значения t-статистик t_i (i = 0,2,4,6,10,13): $t_0 = 43.434$, $t_2 = 5.899$, $t_4 = 6.132$, $t_6 = 5.216$, $t_{10} = 3.573$; $t_{13} = 2.769$. Табличное значение $t_{(0.05,15)} = 1.753$.

Таблица 1 Значения оценок объема производства легкой промышленности

t	Y_1	V_1	
		M1	M2
1	0.1623	0.1632	0.1604
2	0.1468	0.1487	0.1524
3	0.1106	0.1108	0.1118
4	0.1279	0.1225	0.1267
5	0.1378	0.1404	0.1389
6	0.1417	0.1402	0.1389
7	0.1096	0.1135	0.1114
8	0.0995	0.0979	0.0976
9	0.0938	0.0943	0.0991
10	0.0979	0.1025	0.0969

t	Y_1	V_1	
		M1	M2
11	0.1022	0.0995	0.0927
12	0.1200	0.1155	0.1159
13	0.1065	0.1037	0.0997
14	0.1093	0.1108	0.1144
15	0.0693	0.0718	0.0782
16	0.0917	0.1056	0.0870
17	0.0964	0.1541	0.1157
18	0.0941	0.2043	0.0932
19	0.0882	0.0733	0.0914
20	0.0866	0.0313	0.0905

В таблице 1 приведены оценки объема производства легкой промышленности M2, полученные по модели (2), на рис.1 представлен график их изменения. Из приведенных данных видно, что модель МГУА значительно лучше МНК-модели как по прогнозирующим свойствам, так и в смысле значимости коэффициентов и всей модели в целом. При этом важно отметить, что модель МГУА значительно проще, т.е. включает меньше аргументов и, соответственно, меньше оцениваемых параметров.

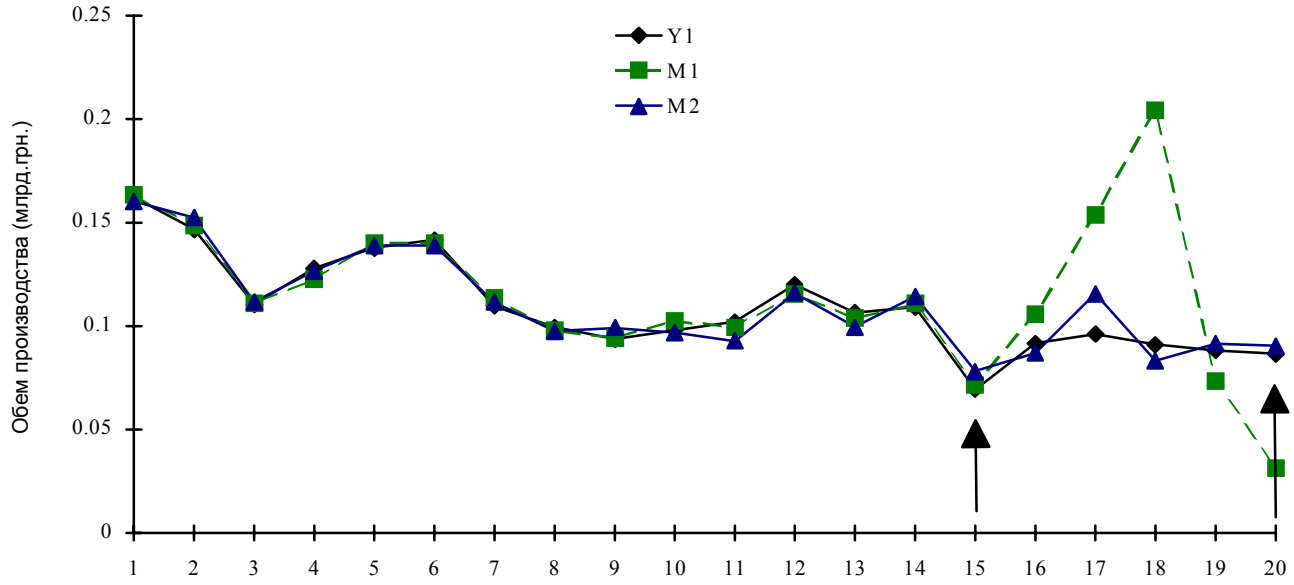

Рис. 1 Сравнение моделей объема производства легкой промышленности, полученных по МНК

(М1) и по МГУА (М2) (стрелками отмечены начало и конец экзаменационной последовательности).

Пример 2. Моделирование инфляции. Для построения модели по данным, взятым из [4], по описанному выше способу анализа структур моделей, построенных по многорядному алгоритму МГУА, были отобраны наиболее существенные аргументы. Ими оказались: X_1 – накопления личные ($ млн.); X_2 – число безработных всего; X_3 – процентные ставки (по Муди); X_4 – потребление личное ($ млн.); X_5 – доходы личные ($ млн.); X_6 – валовой национальный продукт. Выходной величиной является инфляция Y_2 (рассчитывалась по формуле, приведенной в [4] на стр.186). Соответствующие данные приведены в [5].

По имеющейся выборке данных строились модели зависимости инфляции от текущих значений аргументов, причем три последние точки выборки составляли экзаменационную последовательность, т.е. для получения модели использовались только первые пятнадцать точек. Выбор такого варианта расчетов связан с резким изменением характера развития процесса. Ставилась задача: можно ли, используя данные, относящиеся к периоду монотонного развития инфляции, предсказать это резкое изменение? Другими словами, следовало проверить, заложено ли это изменение в предыстории и можно ли его выявить с помощью моделирования.

Модель, полученная по МНК и включающая все шесть исходных аргументов, имеет вид:

$$V_{2,t} = 0.1646 + 0.00015X_{1t} - 0.00702\times X_{2t} - 0.00996X_{3t} + 00143\times X_{4t} + \\ + 0.00031\times X_{5t} - 0.00056\times X_{6t}\ , \qquad (3)$$

где V_2 – модельная оценка выходной переменной Y_2, X_{it} - значение i-го аргумента в t-й точке (i = 1,...,6; t = 1,...,18).

Модель имеет следующие характеристики качества:

СКО = 0.0297; R = 0.0854 ; S =157.9%;

и следующие статистические характеристики:

$R^2 = 0.864$; $R^2_y = 0.762$; F = 8.475; m1 = 6; m2 = 8; $F_{(0.05,6,8)} = 3.58$.

Значения t-статистик t_i (i = 0,1,..,6): $t_0 = 2.9155$; $t_1 = 0.296$; $t_2 = 1.619$; $t_3 = 1.005$; $t_4 = 0.685$; $t_5 = 1.801$; $t_6 = 2.927$; $t_{(0.05,15)} = 1.753$.

В таблице 3 приведены оценки М3 инфляции (МНК), полученные по модели (3), на рис. 2 представлен график их изменения. Видно, что модель довольно плохая, причем наибольшие ошибки относятся к трем экзаменационным точкам, когда тенденция развития инфляции резко изменилась. Таким образом, МНК не позволяет построить модель, отражающую системные закономерности и пригодную для получения надежного прогноза, несмотря на то, что по статистическим оценкам она значима $F > F_{(0.05,6,8)}$, $t_0, t_5, t_6 > t_{(0.05,15)}$.

Модель инфляции, полученная при тех же условиях по МГУА (М4), имеет вид:

$$V_{2,t} = 0.00036X_{1,t} + 0.0109X_{3,t} - 0.00046X_{4,t} + 0.00019X_{5,t} - 0.0001156X_{6,t}\ , \qquad (4)$$

характеризуется такими показателями качества:

СКО = 0.00873; R = 0.0194; S = 35.9%

и имеет следующие статистические характеристики:

$R^2 = 0.718$; $R^2_y = 0.604$; F = 6.796; $F_{(0.05,4,10)} = 3.48$;

Значения t-статистик t_i (i = 1,3,4,5,6):
t_1= 0.631; t_3 = 1.252; t_4 = 0.816; t_5 = 1.299; t_6 = 2.646. $t_{(0.05,15)}$ = 1.753.

Таблица 3 Оценки процесса инфляции

t	Y_2	V_2	
		M3	M4
1	0.0145	0.025	0.0125
2	0.0357	0.024	0.0163
3	0.0267	0.018	0.0107
4	0.0081	0.016	0.0187
5	0.0160	0.014	0.0205
6	0.0092	0.009	0.0174
7	0.0117	0.016	0.0147
8	0.0116	0.014	0.0152
9	0.0140	0.017	0.0192

t	Y_2	V_2	
		M3	M4
10	0.0175	0.020	0.0227
11	0.0283	0.021	0.0307
12	0.0287	0.031	0.0305
13	0.0419	0.040	0.0382
14	0.0536	0.051	0.0451
15	0.0593	0.063	0.0618
16	0.0430	0.090	0.0483
17	0.0326	0.118	0.0421
18	0.0622	0.132	0.0736

В таблице 3 приведены оценки М4 инфляции (МГУА), полученные по модели (4), качество этой модели наглядно характеризует также рис. 2. Видно, что она четко отражает изменение тенденции процесса, не очевидное из предыдущей информации, т.е. из предыстории (до 16-й точки) процесса.

Важно отметить, что улучшение прогноза состоялось за счет у*прощения прогнозирующей модели* (в данном случае за счет исключения из нее аргумента X_2), что характерно именно для применения МГУА (эффект исключения “лишних”, неинформативных факторов).

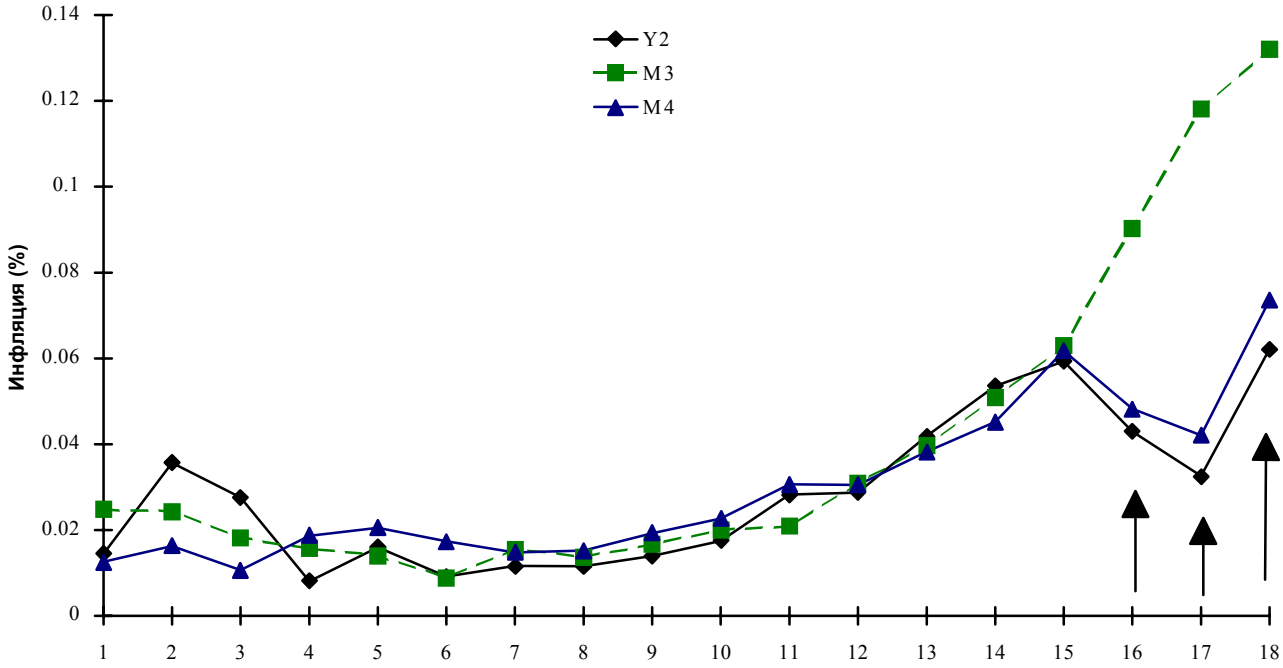

Рис.2 Сравнение моделей инфляции, полученных по МНК (М3) и по МГУА (М4) (стрелками отмечены точки экзаменационной последовательности)

Как следует из изложенного выше, регрессионные модели, даже если они по статистическим характеристикам являются значимыми, мало пригодны для целей прогнозирования. Мо-

дели, построенные по алгоритмам МГУА, по своим прогнозирующим свойствам значительно превосходят регрессионные модели в силу того, что по этим алгоритмам автоматически (за счет применения внешнего дополнения) отбираются аргументы (факторы), наиболее информативные для данного объекта моделирования.

Список литературы

1. Бюлетень економічної кон'юнктури України. – Київ: НДІ статистики Мінстату України. – 1997. - випуск № 3. - 134 с.
2. Справочник по типовым программам моделирования / Под ред. А.Г.Ивахненко. – К.: Техніка, 1980.-184 с.
3. Вучков И., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. – М.: Финансы и статистика, 1987. – 239с.
4. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия: В 2-х вып. Вып. 2. – М.: Финансы и статистика, 1982. – 239с.
5. Степашко В. С., Коппа Ю. В. Опыт применения системы АСТРИД для моделирования экономических процессов по статистическим данным // Кибернетика и выч. Техника, 1999. – Вып. 117. – С. 23-29.

Получено 11.05.2000