УДК 519.23, 519.25

МЕТОД СТРУКТУРНОЙ ИДЕНТИФИКАЦИИ МНОГОФАКТОРНОЙ МОДЕЛИ ЦЕНЫ НА ЖИЛУЮ НЕДВИЖИМОСТЬ

Д.В. БУЛАЕНКО
О. И. СИНЕЛЬНИКОВА

Харьковский национальный университет радиоэлектроники

e-mail:
dbulaenko@gmail.com
ol.sinelnikova @gmail.com

Работа посвящена проблеме разработки многофакторной математической модели цены за один квадратный метр на жилую недвижимость города Харькова, с учетом экономического состояния города и страны. Данная модель позволит моделировать значение ценовых показателей, исходя из различных сценариев, которые возможны в экономике.

Ключевые слова: многофакторная модель, МГУА, комбинированная модель, методы структурной и параметрической идентификации многофакторной модели.

Введение

Современный этап развития информационных технологий характеризуется широким применением математических методов для анализа, моделирования, прогнозирования и проектирования многофакторных систем [3]. Не исключением является и сфера анализа рынка недвижимости. Оценка цен на рынке недвижимости является одной из самых актуальных задач для агентств недвижимости, строительных компаний, а также множества других организаций, деятельность которых связана с инвестициями в объекты недвижимости. Одна из задач, решаемых при этом - построение модели ценообразования для жилья, другими словами, количественная зависимость цены жилья от ценообразующих факторов.

Одна из проблем современного анализа данных на рынке недвижимости - автоматизированный поиск ведущих факторов, определяющих поведение системы. Актуальной и практически значимой является задача определения ведущих факторов, оказывающих максимальное влияние на рынок недвижимости и разработка методов многофакторного статистического анализа, которые позволяют учитывать одновременное влияние на рынок большого числа ценообразующих факторов. Такие методы позволяют разрабатывать новые методики и алгоритмы построения новых многофакторных моделей системы и, на их основе, интерпретировать поведение цен на рынке недвижимости.

На уровень ценовых индексов на недвижимость оказывает влияние огромное количество факторов, например, начиная от показателей состояния экономики страны до уровня развития некоторой территории. Возникает потребность автоматизации обработки данных, так как человек не в состоянии переработать такое количество сведений. Это, в свою очередь, позволит исключить влияние человеческого фактора на результаты эксперимента, сократить время на его проведение. Таким образом, основной целью данных исследований является разработка автоматизированного метода структурной идентификации многофакторной модели цены на рынке недвижимости.

Постановка задачи

Исходя из вышесказанного, приведем формальную постановку задачи построения многофакторной модели цены на рынке недвижимости, исходя из некоторого набора независимых факторов.

Пусть есть исходная выборка:

$$X_1 = \{x_1(t), t = \overline{1,T}\},$$

$$X_2 = \{x_2(t), t = \overline{1,T}\},$$

$$x_m = \{x_m(t), t = \overline{1,T}\},$$

$$y = \{y(t), t = \overline{1,T}\},$$

где: $x_i, i = \overline{1,m}$ - независимые факторы, y - зависимая переменная (средняя цена жилья, у.е), T - максимальная длина ряда. В качестве независимых факторов были выбраны: валовый внутренний продукт (ВВП), инфляция, учетная ставка национального банка Украины (НБУ), курс валют НБУ сред., минимальная заработная плата, наличные деньги в обращении (М0), М0 + чеки, вклады до востребования (М1), М1 + срочные вклады (М2), М2 + сберегательные вклады (М3), средневзвешенные процентные ставки по кредитам Хар. Области, средневзвешенные процентные ставки по депозитам Украины, динамика индекса первой фондовой торговой системы (ПФТС) [2, 8].

Задача многофакторного моделирования и прогнозирования сводится к идентификации структуры и параметров модели:

$$y(t) = F(x_1(t), x_1(t-1), \ldots, x_1(t-T), \ldots, x_m(t), x_m(t-1), \ldots, x_m(t-T), a, s), \quad (1)$$

где: a - вектор параметров модели, s - структура на классе моделей вида:

$$y(t) = a_0 + a_1 x_1(t-l_1) + a_2 x_2(t-l_2) + \ldots + a_m x_m(t-l_m) +$$
$$+ \sum_{i=1}^{m} \sum_{k=1}^{m} c x_i^k(t) + \sum_{i=1}^{m} \sum_{j=1}^{m} b_{ij} x_i(t) x_j(t) +$$

где: l_1, l_m - лаги, максимальной кросс-корреляции между соответствующей независимой переменной и зависимой; a - вектор параметров линейной модели с учетом кросс-корреляционного анализа, c - вектор параметров полиномиальной модели с учетом дисперсионного анализа, b - вектор параметров линейной модели, полученной методом группового учета аргументов.

Первая часть модели получается в виде линейной модель с учетом кросс-корреляционного анализа, полученная методом регрессионного анализа.

Вторая часть модели - полиномиальная модель, полученная в результате структурной и параметрической идентификации моделей отдельно для каждой независимой переменной.

Третья часть модели - полиномиальная модель, учитывающая сочетания переменных, полученная методом группового учета аргументов, при использовании частичных описаний вида:

$$f_{ik}(t) = b_{kl} x_i(t) x_k(t) + b_{ll} x_l(t) + b_{kk} x_k(t).$$

Объединив данные три метода в одну комбинированную модель, мы сможем недостатки отдельных методов скомпенсировать достоинствами других и в результате получить более точную и эффективную модель.

Достоинство модели, полученной методом группового учета аргументов [6, 7], в отличие от кросс-корреляционного анализа и линейных регрессионных моделей, что даже при коротких, неточных или зашумленных данных, метод выдает довольно точную многофакторную модель с учетом парных взаимодействий факторов и с высокой точностью прогноза. Однако, при использовании метода группового учета аргументов (МГУА) часто отмечается опасность потери существенного фактора, и тут поможет использование моделей с учетом кросс-корреляционного и регрессионного анализа, которые помогут выбрать вид частного описания с максимумом коэффициента корреляции на проверочной последовательности.

Метод моделирования на основе комбинированного подхода

Непосредственно метод структурной идентификации многофакторной комбинированной модели состоит из нескольких этапов. На первом этапе происходит кросс-корреляционный анализ между всеми переменными в массиве данных, в результате чего происходит редукция первоначального множества факторов и устанавливаются параметры упреждающих зависимостей. По итогам этого строится линейная многофакторная модель с учетом упреждений, если оценка адекватности данной модели не соответствует выдвигаемым требованиям к результирующей модели, то переходим к следующему этапу.

Далее определяются все возможные зависимости вида $X(t) = f(л,(t - _{ij}$ '', при этом оставляются только модели прошедшие проверку адекватности. Зависимости определяются на классе линейных и полиномиальных моделей, с учетом того, что между переменными может существовать упреждающая зависимость. После чего происходит составление аддитивной многофакторной модели из полученных бинарных отношений и параметрическая идентификация полной модели. Последним этапом построения модели является реализация метода группового учета аргументов с квадратичными или кубическими частичными описаниями.

Общий вид моделей получаемые в результате применения этого метода:

- линейные модели, полученные методом группового учета аргументов в виде обобщенных полиномов Колмагорова-Габбора,

$$y^{(t)} = a_0 + \sum_{i=1}^{m} \quad {}^{(t)} + \sum_{i=1}^{m} \sum_{j=1}^{m} j\,(t)\, x_j^{(t)} + \dots,$$

с частичными описаниями вида:

$$f_{,k}(t) = a_0 + a_{,}x_{,}(\,) + a\,A(\,)$$

или

$$f_{,k}(t) = a_0 + a_{,}X_{,}(t) + a_k X_k(t) + a_{,,k} X_{,}(t) X_k(t) + a_{,,,} X_f(t) + a_{k,k} x_l(t)$$

- линейная модель с учетом кросс-корреляционного анализа, полученная методом регрессионного анализа [1, 2],

$$j)(t) = a_0 + a_1 x_1(t - l_1) + a_2 X_2(t - l_2) + \dots + a_m x_m(t - l_m),$$

где: $l_1, l_2, \dots, l_m$ - лаги, максимальной кросс-корреляции между соответствующей независимой переменной и зависимой,

- полиномиальная модель с учетом дисперсионного анализа, полученная методом регрессионного анализа,

$$y(t) = a_0 + \sum_{i=1}^{m} \sum_{k=1}^{n_i} a_i X(t) \bullet$$

Структурная и параметрическая идентификация моделей проходит согласно следующему векторному критерию:

$$\pounds \wedge \min\min, \tag{2}$$

$$\pounds_{00}(a, s) = \min \left(\frac{\pounds_1(a, s)}{\max \pounds_1(a, s)} \quad \frac{\pounds_2(a, s)}{\max \pounds_2(a, s)} \quad \frac{\pounds_3(a, s)}{\max \pounds_3(a, s)} \right) \tag{3}$$

$$\pounds_1(a, s) = 1 \sum (\,) - j \wedge (\,(\,)\,)^2, \quad \pounds_2(a, s) = \ \sum \wedge f \wedge$$

$$£3(a>s) = T f_{t=1} ((() - y_{()}(())),$$

где: $У(a,s)^{(t)}$- модельное значение зависимой переменной (l) при определенном наборе параметров a и структуре модели s . Таким образом, считаем лучшей ту модель, у которой нормированное значение ошибки одного из рассматриваемых типов минимально. Нормирование происходит по всем рассматриваемым структурам моделей.

Метод моделирования на основе ARIMAX

Проведем сравнительный анализ указанных выше классов моделей с моделью, представленной в виде модели сезонной авторегрессии - проинтегрированного скользящего среднего с экзогенными переменными (САРПССЭ) [8]:

$$o_t = f \; 8_{Ij}(B) \; t\text{-}bj \; + \frac{eq(B)}{Ф{+}{+}(B)} a_t ,$$

где: Q_t, $t = 1,n$ - исходный или преобразованный (нормированный или прологарифмированный) центрированный временной ряд зависимой величины; n - объём выборки; F_t, $t = 1,n$, $1 = 1,N$ - исходные или преобразованные (нормированные или прологарифмированные) центрированные временные ряды изменения значений внешних (экзогенных) факторов; N - количество экзогенных факторов; B - оператор сдвига по времени на одну единицу назад, такой что $B X_k = X_{k-1}$; b_i - величина задержки i-го экзогенного временного ряда по времени, относительно прогнозируемого временного ряда $°t$; $8r(B)$ и $юc(B)$ - полиномы передаточных функций от B степеней r_i и c_i соответственно; $8(B) = 1 - 8_j B - 8_2 B - \dots - 8_r B^r$; $®Cj(B) = - ®_1 B - ®_2 B - \dots - < B^Ч$; $Ф{+}{+}(B)$ - обобщенный оператор авторегрессии порядка $p{+}= p + \sum_{i=1}^{n_s} D_i s_i , p = \sum_{i=1}^{n_s} p_i s_i$; $ф_{p+}(B) = ф_p(B) \dots$;

D_i, $1 = 1,n$, - порядок взятия разности S_i; S_i, $1 = 1,n$, - период периодической компоненты, причем $S_1 = l$; n, - количество периодических компонент; и B^{S_1} - упрощающие операторы такие, что $= (1-B^{1}) x_t = x_t - x_{t-S_1}$.

Анализ результатов

Были рассмотрены данные в период с января 2005 года по декабрь 2008 года ежемесячно, и в результате получилось 48 значения. Стоит отметить, что выборка достаточно короткая, однако при попытке прогнозировать экономические процессы данная ситуация не редкость. Для проверки эффективности предлагаемого подхода и оценки качества структурной идентификации была выбрана величина - Средняя цена жилья на первичном рынке (у.е), а в качестве независимых факторов: ВВП (x_1), инфляция (x_2), учетные ставки НБУ (x_3), курс валют НБУ сред. (x_4), минимальная заработная плата (x_5), М0 (x_6), М1 (X_7), М2 (x_8), М3 (X_9), средневзвешенные процентные ставки по кредитам Хар. Области ($x_{ю}$), средневзвешенные процентные ставки по депозитам Украины ($x_{ц}$), динамика индекса ПФТС (X_{12}).

Все рассмотренные ряды независимых факторов имеют тренды, а также являются персистентными, так как показатель Херста больше 0.5.

Безусловно, перечень факторов влияющих на значение цены на недвижимость может быть существенно расширен. Однако, в данной работе ставилась задача провести оценку эффективности метода структурной идентификации независимо от реальной связи, существующей в данных, а получение некоторого статистического результата связи между факторами.

Линейная модель на основе МГУА без учета упреждений

В результате построения линейной модели МГУА было получено 3 слоя моделей, в лучшей модели 3-го слоя присутствуют такие переменные: x_1, x_2, x_3, x_5, x_6

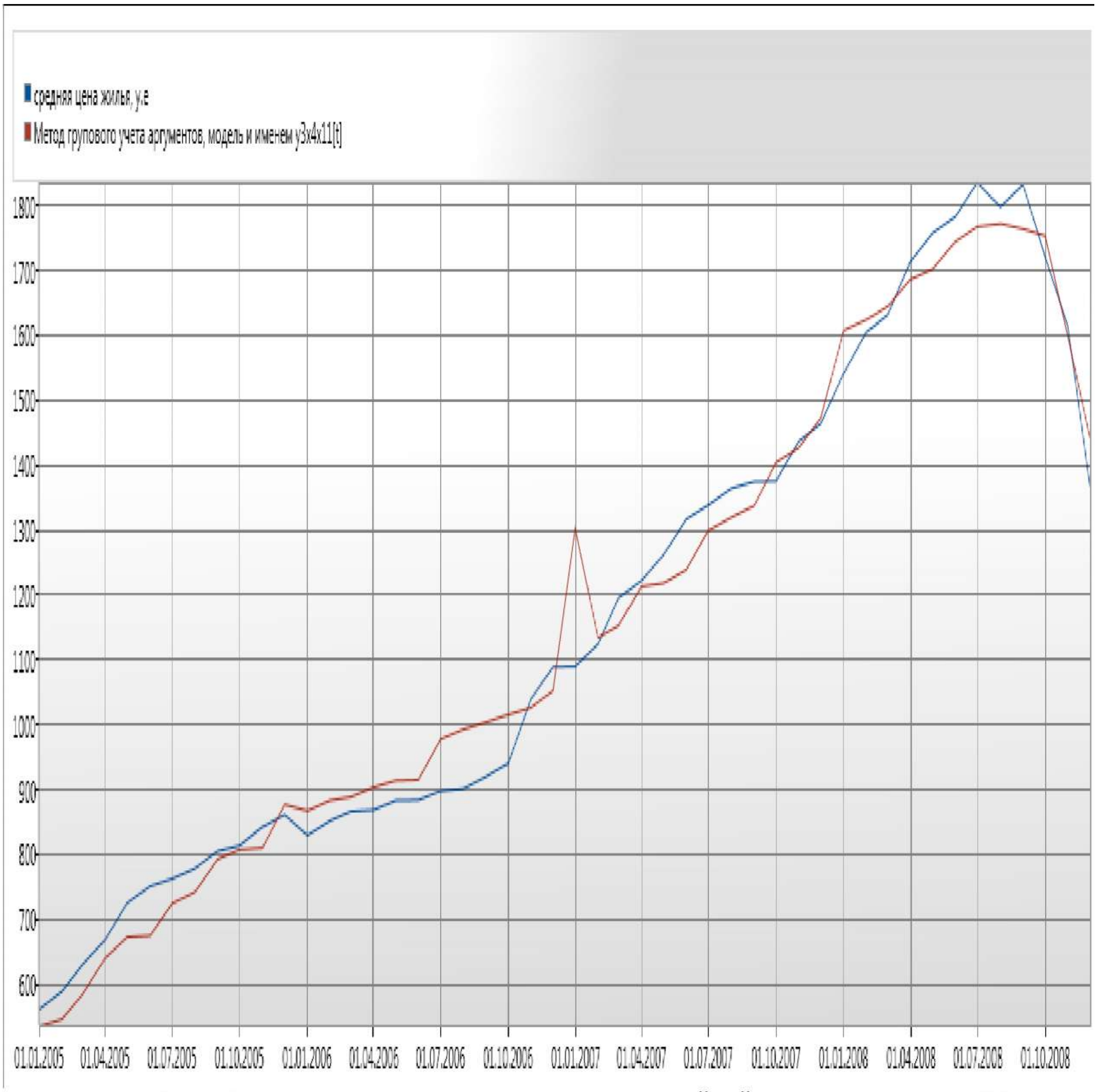

Рис. 1. Результат моделирования на основе линейной модели 3-го слоя МГУА

Результаты проверки адекватности данной модели показали, что зависимость между рассматриваемыми переменными в виде линейной функции неадекватно отражает связь факторов.

Модель ARIMAX

В построении этой модели было сделано предположение о том, что все выбранные независимые переменные влияют на «Средняя цена жилья на первичном рынке».

x_1 - лаг отсутствует, x_2 - лаг отсутствует, x_3 - лаг отсутствует, x_4 - с лагом в 4 значения, x_5 - лаг отсутствует, x_6 - лаг отсутствует, x_7 - лаг отсутствует, x_8 - лаг отсутствует, x_9 - лаг отсутствует, x_{10} - лаг в 2 значения, x_{11} - лаг в 1 значение, x_{12} - лаг отсутствует.

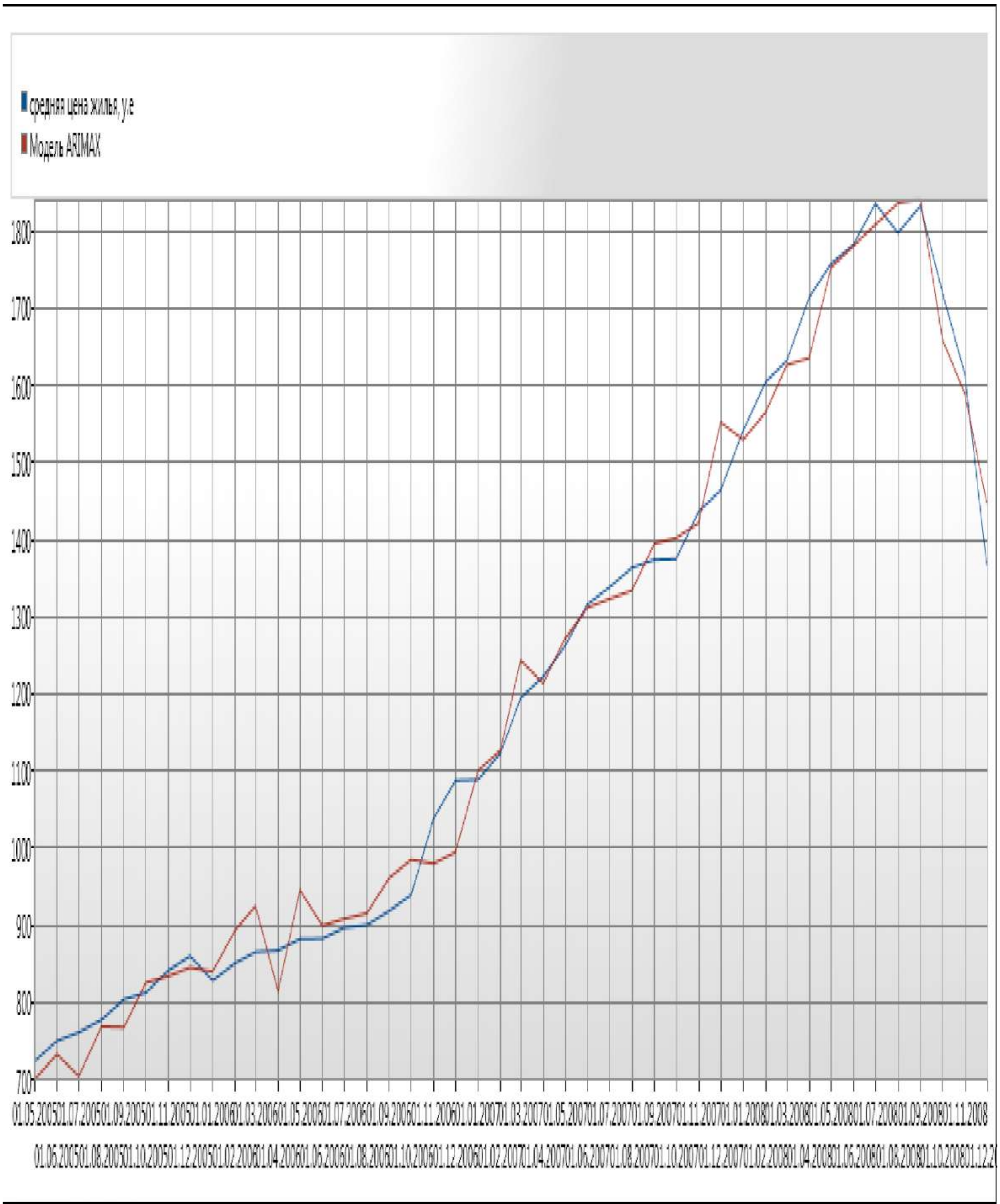

Рис. 2. Результат моделирования на основе модели ARIMAX

Нелинейная модель на основе МГУА без учета упреждений

В результате построения нелинейной модели МГУА было получено 10 слоев моделей, в лучшей модели 4-го слоя присутствуют такие переменные: x_1, x_2, x_3, x_4, x_5

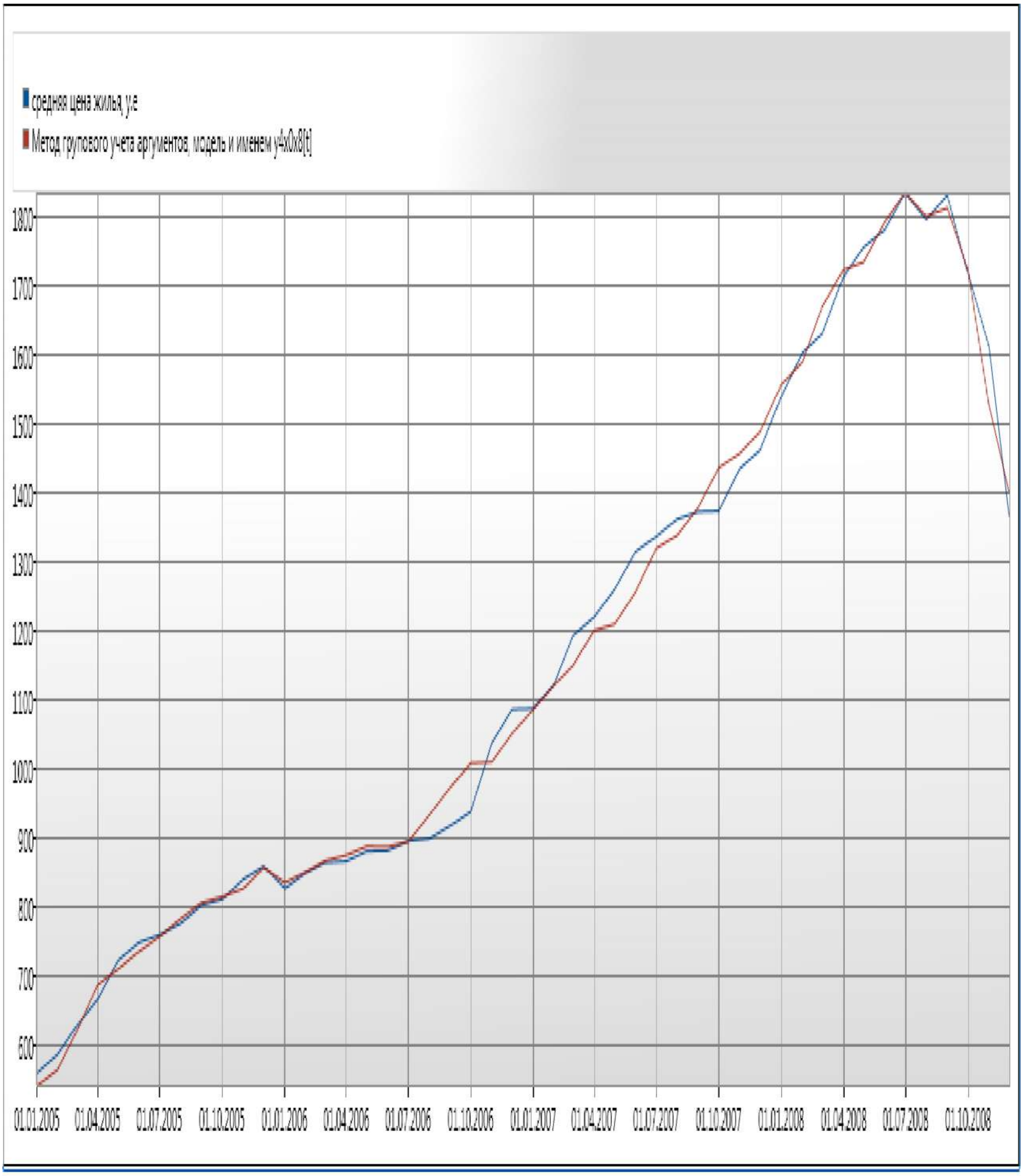

Рис. 3. Результат моделирования на основе модели 4-го слоя МГУА

Нелинейная модель, полученная методом группового учета аргументов на основе квадратичных частичных описаний значительно более точно позволяет на заданном множестве независимых факторов получить модель зависимой переменной.

Нелинейная модель с учетом дисперсионного анализа

Для построения этой модели необходимы результаты расчета коэффициентов детерминации. В результате построения этой модели некоторые факторы могут быть отброшены, так как не оказывают существенного влияния на зависимую переменную.

x_1 - отсутствует в модели, x_2 - отсутствует в модели, x_3 - в виде квадратичной функции, x_4 - отсутствует в модели, x_5 - линейная функция, x_6 - в виде квадратичной функции, x_7 - в виде квадратичной функции, x_8 - в виде квадратичной функции, x_9 - в виде квадратичной функции, x_{10} - отсутствует в модели, x_{11} - отсутствует в модели, x_{12} - в виде квадратичной функции.

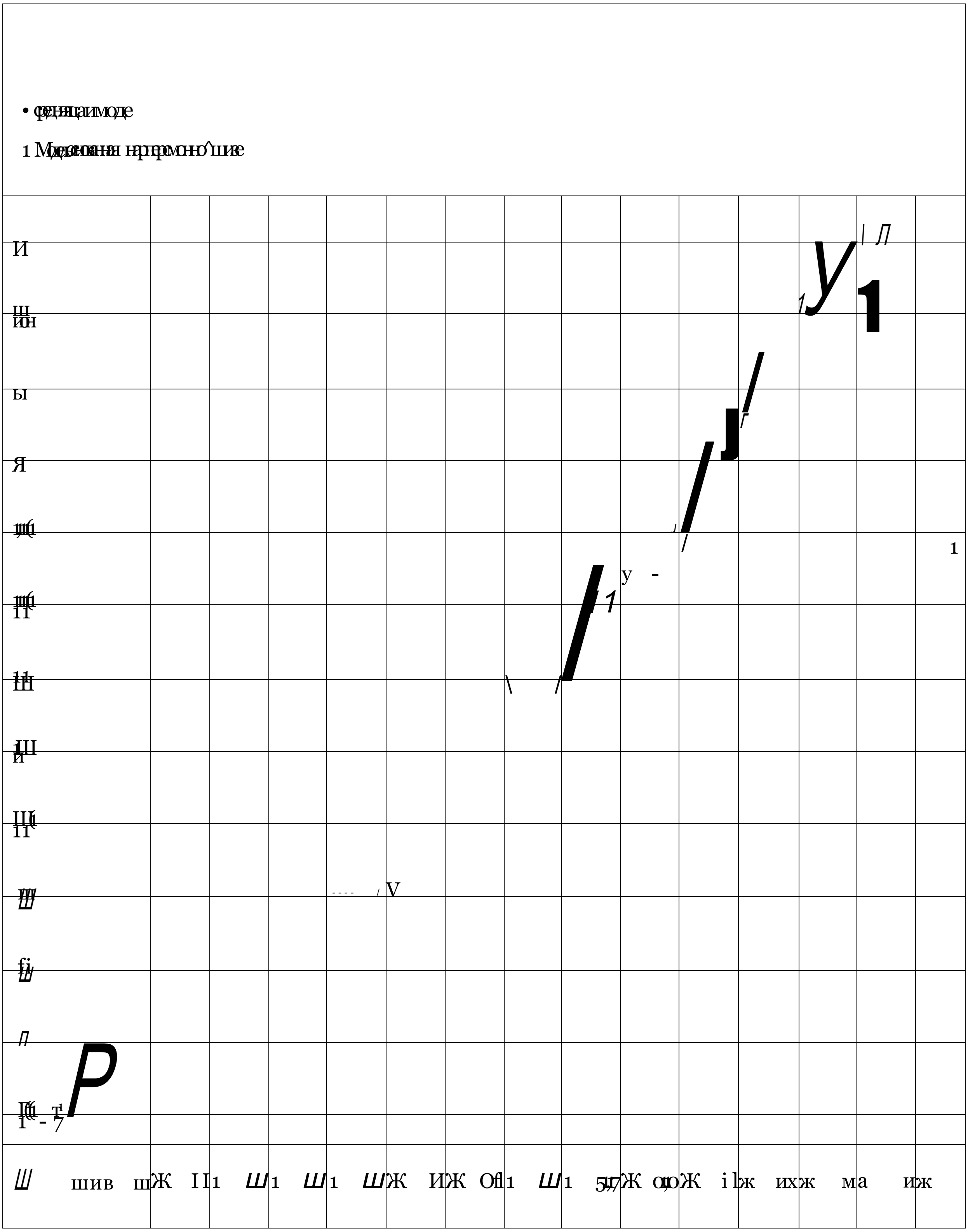

Рис. 4. Результат моделирования на основе нелинейной модели с учетом дисперсионного анализа

Комбинированная модель

В лучшей модели присутствуют такие переменные:

x_1 - линейная функция, x_2 - отсутствует в модели, x_3 - в виде квадратичной функции, x_4 - отсутствует в модели, x_5 - линейная функция, x_6 - в виде квадратичной функции, x_7 - в виде квадратичной функции, x_8 - в виде квадратичной функции, x_9 - в виде квадратичной функции, x_{10} - отсутствует в модели, x_{11} - отсутствует в модели, x_{12} - в виде квадратичной функции.

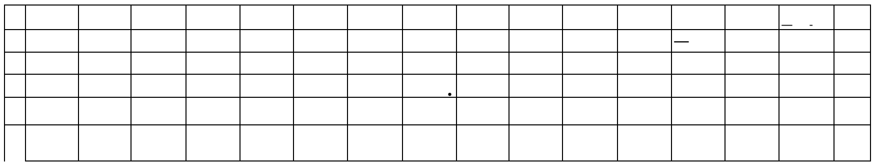

Рис. 5. Результат моделирования на основе комбинированной модели

Таблица

Сводная таблица характеристик моделей

Вид модели	Максимальная ошибка	Минимальная ошибка	Процентная ошибка (от 0 до 1)
Линейная модель с учетом кросс-корреляции, лучшая модель 3-го слоя	214,2679	4,0967	0,0412
МГУА, лучшая модель 4-го слоя	85,8721	0,177	0,0182
Модель основанная на дисперсионном анализе	163,6977	0,6744	0,0394
ARIMAX	91,2909	1,9081	0,028
Обобщенная модель вида: МГУА + Модель основанная на дисперсионном анализе, МГУА с частичными описаниями вида: $y = a_0 + a_1 x_i x_j$.	154,3229	1,5743	0,03695

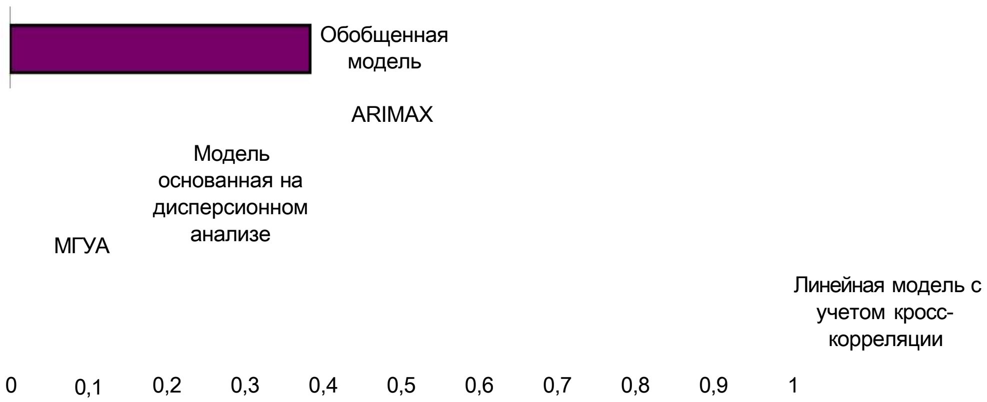

Рис. 6. Нормализированные ошибки моделей

Результаты проведенных исследований показали, что на данном наборе данных МГУА обладает меньшими значениями ошибок.

Вывод

Научная новизна данной работы состоит в разработке многофакторной математической модели цены за один квадратный метр на жилую недвижимость города Харькова, с учетом экономического состояния города и страны. Данная модель позволит моделировать значение ценовых показателей, исходя из различных сценариев, которые возможны в экономике.

К настоящему времени разработано множество моделей для решения задачи прогнозирования временного ряда, среди которых наибольшую применимость имеют авторегрессионные и результаты данной статьи показали, что наиболее точный результат на данном соотношении объемов выборки и количества факторов у МГУА. Но наиболее перспективным направлением развития моделей прогнозирования с целью повышения точности является создание комбинированных моделей.

Список литературы

1. Ванюкевич О.Н., Попов А.А. Критерии выбора модели при построении размытой регрессионной зависимости. / О.Н. Ванюкевич, А.А Попов // Сборник научных трудов НГТУ. - 2004. - №4(38). - с. 15-30.
2. Дронов С.В. Многомерный статистический анализ: учебное пособие. Барнаул: Изд-во Алт. гос. ун-та, 2003. - 213 с.
3. Цветков В. В, Сумин В. И. Об алгоритмах и моделях, данных в решении задач принятия решения/ В. В. Цветков, В. И. Сумин // Научные ведомости БелГУ. 2010. № 13 (84). Выпуск 15/1. - С. 138-141.
4. Ханк Д.Э., Уичерн Д.У., Райтс А.Дж. Бизнес прогнозирование, 7-е издание. - М.: Издательский дом «Вильямс», 2003. - 193 с.
5. Ferster E., Rents B. Methods of correlation and regressionanalysis: Transl. fromgerman - Moskow: Finance and statistics, 1983. (In russian). - 302 p.
6. Madala H.R. and Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modeling, Boca Raton: CRC Inc., 1994. - 384 p.
7. Self-organizing methods in modeling: GMDH type algorithms / Ed. S.J.Farlow — New York, Basel: Marcel Decker Inc., 1984. — 350 p.
8. Viviana Fernandez, Forecasting commodity prices by classification methods: The cases of crude oil and natural gas spot prices. [Электронный ресурс] / Viviana Fernandez. - Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.2375&rep=rep1&type=pdf.

method of structural identification multifactorial model of residential real estate prices

D. V. BULAIENKO
O. I. SINELNIKOVA

Kharkov National University of Radioelectronics

e-mail:
dbulaenko@gmail.com
ol.sinelnikova @gmail.com

The work is devoted to the problems of developing mathematical model of multi-factor prices per square meter for residential real estate of the city of Kharkov, in view of the economic situation of the city and the country. This model will allow to model the value of price indicators, based on the different scenarios that are possible in the economy.

Keywords: multi-factor model, GMDH, combined model, methods of structural and parametric identification of multi-factor model.