Кондрашова Н.В.

Влияние способа разбиения выборки в алгоритмах МГУА на точность прогнозирования

На примере процесса развития инфляции исследовано, в какой мере способ разделения выборки данных в различных алгоритмах МГУА определяет точность предсказания резкого изменения характера процесса. Сделан вывод об эффективности алгоритма адаптивного прогнозирования при квазиоптимальном разбиении данных.

Алгоритмы МГУА давно и эффективно используются для получения моделей однократного и многократного прогнозирования экономических процессов. В работах [1,2] алгоритмы МГУА средствами системы АСТРИД применялись для решения задач экстраполяции экономических процессов.

Традиционно в алгоритмах МГУА для формирования внешних критериев выполняются последовательное (в порядке следования данных) разбиение выборки или разбиение "по дисперсии" — на подобные (ПД) или не подобные (НПД) части выборки. В данной работе на примере данных процесса развития инфляции проводится сравнение квазиоптимального разбиения [3] с перечисленными способами для установления, при каком разбиении выборки получаются модели, дающие наиболее точный результат прогнозирования.

Обычно при введении упреждающих переменных происходит укорачивание таблицы исходных данных вблизи точки прогноза, в результате точность прогноза существенно падает. Для повышения точности предсказания предложено применять адаптивное прогнозирование.

Постановка задачи. Назовем адаптивным прогнозом такое последовательное прогнозирование, при котором в таблице исходных данных и при вычислении прогнозных значений используются значения прогноза на предыдущих шагах.

Рассмотрим его сущность на конкретном примере предсказания резкого изменения характера процесса инфляции.

Исходная таблица данных [4] содержит m=6 аргументов (столбцов) и N=18 строк (точек исходных данных). Аргументами являются экономические показатели [1]: x_1 - личные накопления (\$ млн.); x_2 - число безработных (млн. чел); x_3 - процентные ставки (по Муди); x_4 - личное потребление (\$ млн.); x_5 - личные доходы (\$ млн.); x_6 - валовой национальный продукт (\$ млрд.). Значения индекса инфляции Y (в % прироста к 100%) рассчитывались по формуле приведенной в [4] и учитывали множество экономических факторов x_m, где $m >> 6$. На рис. 1 представлен график изменения во времени индекса инфляции Y_k , $k=\overline{1,N}$.

Линейные по параметрам прогнозирующие модели различались формой шаблона, т.е. разностной схемой

$$Y_{k+L} = F(X_k, X_{k-1},..,X_{k-j}, Y_k,...,Y_{k+L-1}),$$

которая показывает, какие из значений $x_{m,k}$, $m = \overline{1,6}$, Y_k, и в какие моменты времени k-j,..,k+L, связаны между собой, где L= 1,2,3 — число шагов упреждения прогноза, j=1,2 — число запаздываний. Здесь $k = \overline{(j+1), N_1}$, $N_1 = n_A + n_B$; n_A, n_B, n_C — число точек обучающей, проверочной и экзаменационной подвыборок, $(N = N_1 + n_C)$; X_k — исходная матрица состояния входных

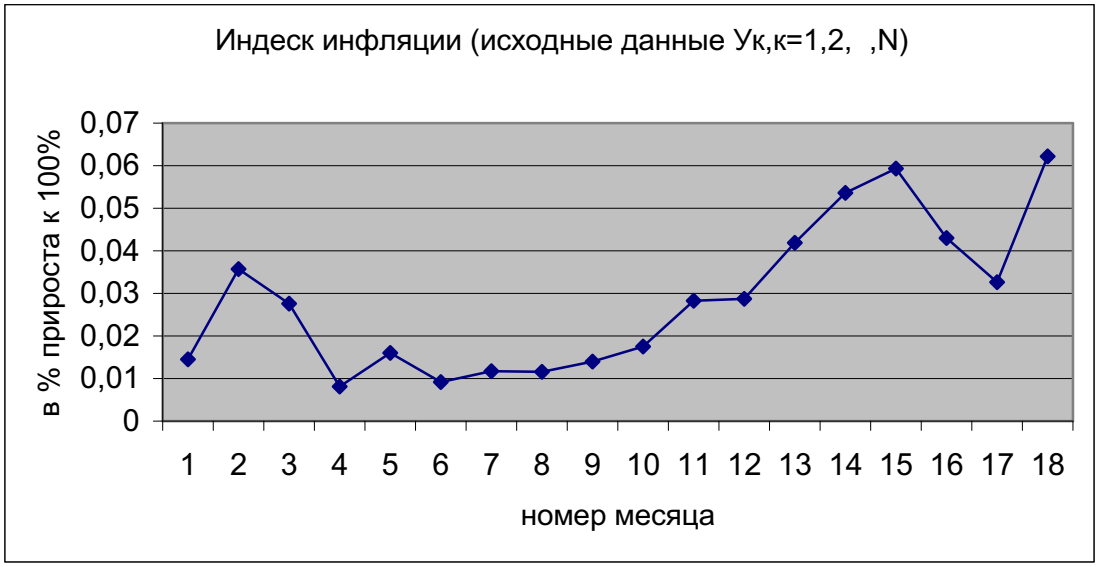

Рис.1

векторов в момент k, $\dim X_{k-j} = N_1 \times 6$, $j = 0,1,2$, N_1- число строк таблицы исходных данных, использующихся для определения структуры, параметров и выбора модели прогноза.

Последние по времени значения индекса инфляции (см. рис. 1) интересны для задачи прогнозирования, т.к. представляют собой ряд резко изменяющихся значений. Для прогноза были выбраны два и три последних по времени значения индекса инфляции. В первом случае считались известными N_1=16 строк таблицы, прогнозировались $n_C = 2$ последних значения, во втором — N_1= 15 строк и прогнозировались $n_C = 3$ последних по времени значения.

Традиционно при введении каждой новой переменной запаздывания по какой-либо одной из входных переменных $x_m, m \in [1,6]$ или упреждения по выходной переменной Y в шаблон таблицу данных расширяется на один столбец. Причем каждое введение запаздывания на 1 шаг по какой-либо переменной матрицы X_k сокращает таблицу на 1

строку сверху. Каждое введение упреждения на 1 шаг по выходной переменной Y_k укорачивает таблицу на 1 строку снизу. Это –1-ый способ формирования “рабочей” таблицы, при котором число ее строк равно:

$$N_2(L, j) = N_1 - j - L\,. \qquad (1)$$

При данном способе для построения каждой из прогнозирующих моделей строится своя “рабочая” таблица данных по формулам, могущим содержать минимально следующий набор аргументов:

$$\hat{Y}_{k+L} = F(X_k, Y_k, ..., Y_{k+L-1})\,, \qquad L = \overline{1, n_C}\,,\ k = N_1\,. \qquad (2)$$

Данный шаблон включает, как частный случай, при полном переборе структур моделей алгоритмом COMBI шаблон вида:

$$\hat{Y}_{k+L} = F(X_k, Y_k)\,, \quad L = \overline{1, n_C}\,, \quad k = N_1\,, \qquad (3)$$

который на этапе прогнозирования не требует знания $Y_{k+1}, ..., Y_{k+L-1}$ неизвестных за пределами “рабочей” таблицы.

Для повышения точности предложено было вводить прогнозные значения в “рабочую” таблицу (2-ой способ формирования таблицы). Прогноз по модели с адаптивным прогнозом на первом шаге находится по аналогичному (2) шаблону число строк равно $N_2 = N_1$-1. При добавлении последующих столбцов “рабочей” таблицы данных $L = \overline{2, n_C}$, для модели с адаптивным прогнозом (на втором и т.д. до n_C-ого шага) — по данным N_1-ой строки таблицы исходных данных— используются прогнозы предыдущих шагов $\hat{Y}_{k+1}, ..., \hat{Y}_{k+L-1}$:

$$\hat{Y}_{k+L} = F(X_k, Y_k, \hat{Y}_{k+1}, ..., \hat{Y}_{k+L-1})\,, \qquad L = \overline{2, n_C}\,, \qquad k = N_1\,. \qquad (4)$$

Такой прием позволяет при составлении условных уравнений Гаусса не уменьшать первоначальное количество строк таблицы исходных данных, т.е. в данном случае число строк таблицы остается тем же – ($N_1 - 1$).

Эти два основных способа подготовки данных были использованы при рассмотрении задач прогнозирования и прогноза с экстраполяцией. При этом сравнивались модели, полученные по данным различных способов разбиения выборки.

Методика исследований. Остановимся ниже на процессе построения моделей при различных способах разбиений. В примере с шаблонами (2), (4) матрица Ω_L является расширенной столбцами $Y_k, ..., Y_{k+L}$, $L = 1,2$ по отношению к матрице *X*. В случае адаптивного прогноза в матрице Ω_L имеются $(n_C - 1)$ строки, содержащие $(n_C - 1)$ прогнозные переменные $\hat{Y}_{k+1}, .., \hat{Y}_{k+L-1}$ для $L = \overline{2, n_C}$. При разбиении Ω_L на две матрицы Ω_{A_L}, Ω_{B_L}, где $dim[\Omega_{A_L} \vdots \Omega_{B_L}] = N_2(L) \times M_L$, ($M_L$=7+*L*), находится так называемое квазиоптимальное или ρ^2 – пропорциональное разбиение матрицы $\Omega^*_{A_L}, \Omega^*_{B_L}$ с числом

строк соответственно $n^*_{A_L}$, $n^*_{B_L}$ [3]. Затем для матрицы Ω_L применим разбиение "по дисперсии точек наблюдений". При формировании "ранжированных по значению дисперсии точек" подвыборок A и B различными способами - через одну (две, три…) точку (что означает получение "подобных" частей выборки — ПД разбиение), либо разбиение подряд (получение "непохожих" частей — НПД разбиение), будем относить в подвыборку A, найденное выше, число $n^*_{A_L}$ точек и в подвыборку B — $n^*_{B_L}$ точек. Иными словами, при квазиоптимальном, ПД и НПД разбиениях строки меняются местами в каждой ее “рабочей” таблице в соответствии с упомянутыми разбиениями точек подвыборок. При последовательном разбиении точки, расположенные во времени подряд, также разбиваются в отношении $n^*_{A_L}$: $n^*_{B_L}$, но переформирования таблицы не происходит.

После составления таблиц данных и применения алгоритма COMBI анализировались три лучшие модели, отобранные по критерию регулярности. В качестве лучшей модели по результатам ”на экзамене” как на первом шаге $\hat{Y}_{k+1}$, т.е. в (N_1+1)-ой точке таблицы исходных данных, так и на втором шаге $\hat{Y}_{k+2}$ по шаблону вида (4), выбирается та модель, у которой получено наименьшее отклонение $(\hat{Y}_{N_1}-Y_{N_1})$ от табличного значения в 16-ой точке. Это правило сохраняется для всех шагов прогноза $(L=\overline{1,3})$ и в других примерах, кроме одного, с шаблоном (4) при n_c=2. Заметим, что при квазиоптимальном, ПД и НПД разбиениях N_1-ая строка исходной таблицы, содержащая значения Y_{N_1}, может стоять в произвольном месте “рабочей” таблицы данных. Следует отметить также, что минимум ошибки прогнозирования в N_1-ой точке, не всегда совпадает с минимумом критерия регулярности.

Результаты испытаний. На рис 2 представлены результаты прогнозирования и прогнозирования с экстраполяцией алгоритмом из пакета программ АСТРИД полного перебора вариантов структур в заданном базисе переменных (COMBI) для случая $n_C = 2$. Число строк “рабочей” таблицы в данном примере равно $N_2 = 15$ и при $L = 2$ укорачивания такой таблицы не происходит. Здесь прогнозируемые значения ”на экзамене” вычисляются по значениям, соответствующим 16-ой строчке.

Значение Y_k при прогнозе на один шаг (L=1 берется из “рабочей” таблицы. Следует отметить, что в данном примере Y_{k+1}, т.е. Y_{17}, известно в исходной (первоначальной) таблице, но выходит за нижнюю границу “рабочей” таблицы данных, т.к. N_1=16. Поэтому такой способ прогнозирования имеет ограниченное применение (на один шаг). Поскольку значение Y_{k+1} при прогнозе на два шага вперед (L=2) берется из первоначальной таблицы

исходных данных при $N_2 = 15$, то на втором шаге по шаблону (3) имеем экстраполяцию $\hat{Y}_{18}$ при известном значении Y_{17}. На рис. 2 можно сравнить с табличными данными $Y_{k+L}, k = 16, L = \overline{1,2},$ значения прогноза с экстраполяцией $\hat{Y}_{k+L}$ ($k = 16;\ L = \overline{1,2}$), полученные для различных способов разбиения исходных данных: квазиоптимальном, последовательном, ПД и НПД разбиениях. Результат по модели прогноза вида (3) квазиоптимального разбиения для таблицы из 15-ти строк также представлен графиком на рис. 2 с пометкой ‘прогноз’. Такая модель уступает по точности модели с экстраполяцией (2) квазиоптимального разбиения, а также модели НПД разбиения вида (3).

Как показало исследование прогнозирования по шаблону (2) на два шага вперед ($n_C = 2,\ L = 2$) при укорачивании “рабочей” таблицы по формуле (1), точность прогноза существенно падает, (иллюстрирующие графики здесь не приводятся).

Рассмотрим способ адаптивного прогнозирования процесса. На рис.3 представлены результаты, полученные по модели с адаптивным прогнозом. На первом шаге прогноз осуществлялся по аналогичному (2) шаблону. Здесь при составлении последней строки второй “рабочей” таблицы данных (k=16, L=2 и вычислении прогноза (4) на втором шаге ”экзамена” по данным N_1-ой строки таблицы исходных данных используется прогноз первого шага $\hat{Y}_{k+1}$ по шаблону (2).

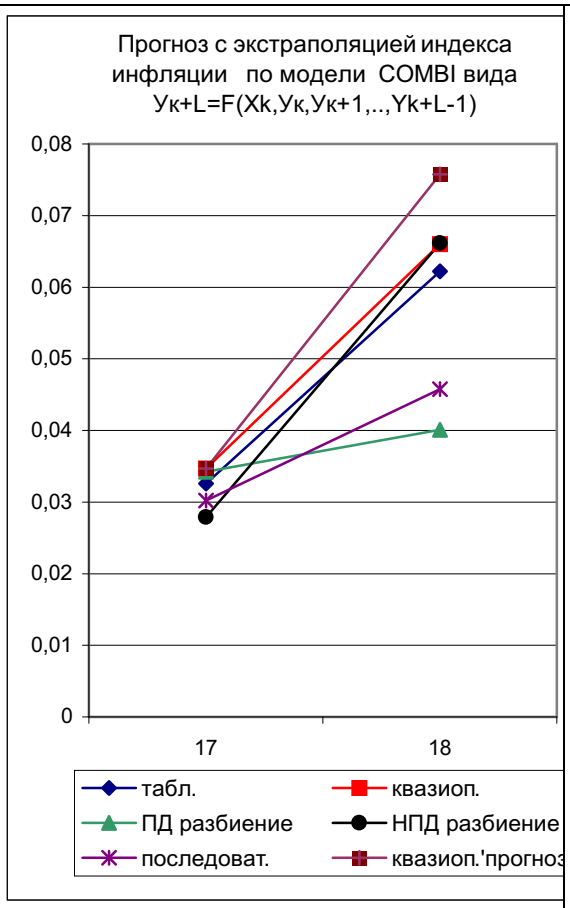

Рис. 2

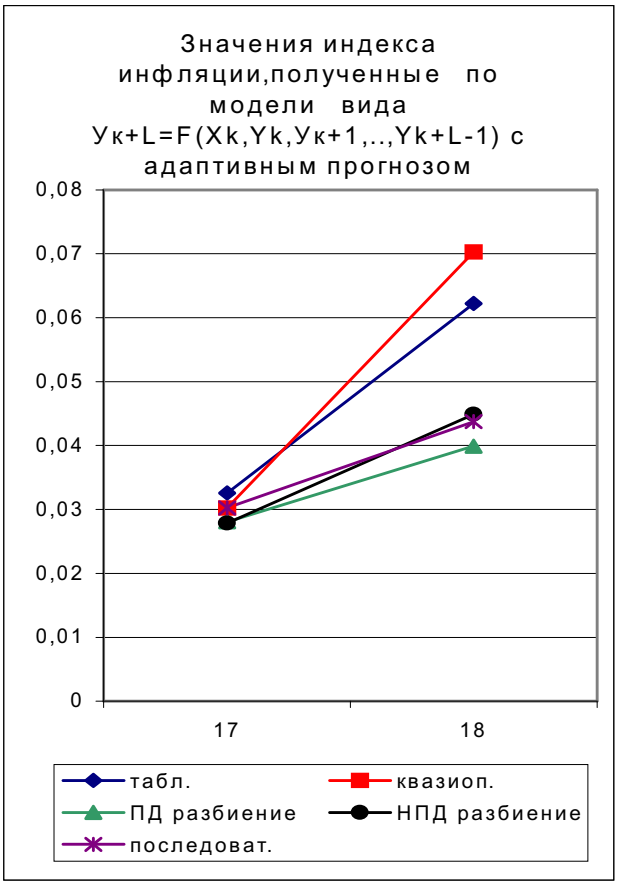

Рис. 3

При прогнозе на 2 шага по двум лучшим отобранным моделям $\hat{Y}_{18} = 0{,}03$ и $\hat{Y}_{18} = 0{,}07$ ошибка отклонения $(\hat{Y}_{16} - Y_{16})$ была соответственно 3,1E-4 и 7,4E-4, но выбор был сделан в пользу той модели, прогноз по которой в дальнейшем подтвердился по другим моделям.

Учитывая не вполне убедительный результат прогнозирования на $n_C = 2$ шага вперед, при увеличении интервала прогноза до $n_C = 3$ трудно было ожидать хорошего результата по моделям вида (2)-(4), включающим только переменные упреждения. Моделирование подтвердило эти ожидания, иллюстрирующие графики здесь не приводятся.

Для того, чтобы точнее прогнозировать, необходимо расширять исходный базис запаздывающими аргументами. В работе [1] введением лаговых переменных повышалась точность экстраполяции.

На рис.4 и 5 представлены результаты прогнозирования на три шага ($n_C = 3$) по 1-ому и 2-ому способам построения “рабочей” таблицы данных. По первому способу при L=1, j=0 “рабочая” таблица имеет N_2=15 строк и 8 столбцов, при L=2, j=0 — 14 строк и 9 столбцов. Для алгоритма COMBI число столбцов таблицы при L=3 не должно превышать $M_{\max}$=14. Дополнительные запаздывающие аргументы $x_{1,k-2}, x_{3,k-1}, x_{5,k-2}$, таких же, что и в работе [1] входных векторов, расширяют базис переменных до девяти (m+3+L<$M_{\max}$). Введение

указанных запаздывающих входных переменных сокращает таблицу исходных данных сверху на j=2 строки, т.е. число строк (1) “рабочей” таблицы равно $N_2(L)=12$ припрогнозе на 1 шаг

.На рис. 4 имеем результаты моделирования индекса инфляции с использованием адаптивного прогноза. Прогноз ”на экзамене” на 1 шаг осуществляется по шаблону:

$$\hat{Y}_{k+1} = F(X_k, x_{1,k-2}, x_{3,k-1}, x_{5,k-2}, Y_k), \qquad k=N_1. \tag{5}$$

При прогнозе на 2 и 3 шага в “рабочую” таблицу последовательно добавляются одна и две строки, содержащие значения $\hat{Y}_{k+1}$ и $\hat{Y}_{k+1}, \hat{Y}_{k+2}$, так что число строк ($N_2=12$) остается неизменным, не зависящим от L. Значения прогноза при этом вычисляются по разностной схеме из следующего набора аргументов:

$$\hat{Y}_{k+L} = F(X_k, x_{1,k-2}, x_{3,k-1}, x_{5,k-2}, Y_k, \hat{Y}_{k+1}..., \hat{Y}_{k+L-1}), \qquad L=2,3;\ k=N_1. \tag{6}$$

На рис. 5 представлены результаты прогнозирования для “рабочих” таблиц, составленных по 1-ому способу, т.е. число строк $N_2(L)$ уменьшается для различных $L=\overline{1,3}$ в соответствии с формулой (1). Таблицы содержат столбцы $\Omega_L=[Y_{k+L}, X_k, x_{1,k-2}, x_{3,k-1}, x_{5,k-2}, Y_k, ..., Y_{k+L-1}]$, $k=\overline{1,(N_2)}$.

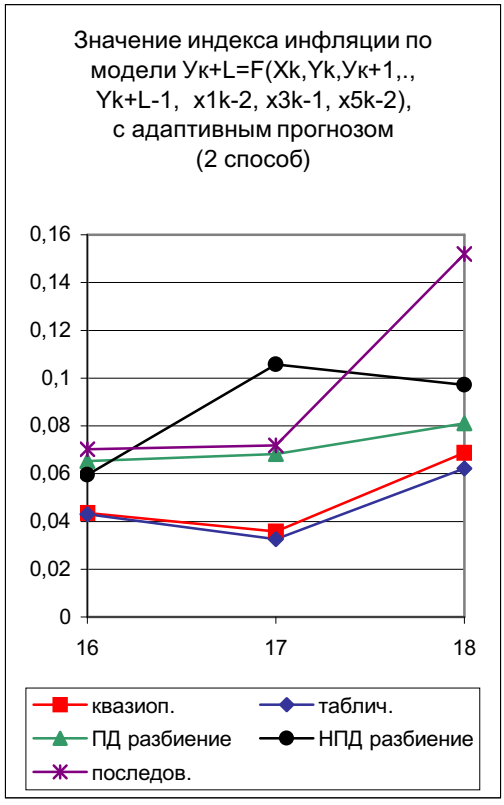

Рис. 4

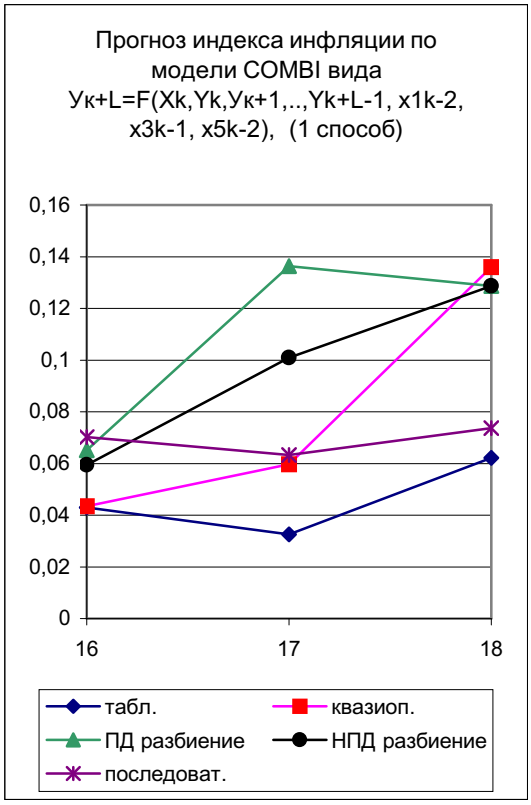

Рис. 5

При вычислении выходной переменной модели прогноза:

$$\hat{Y}_{k+L} = F(X_k, x_{1,k-2}, x_{3,k-1}, x_{5,k-2}, Y_k, ..., Y_{k+L-1}), \quad L = 1,2,3, \quad k=\overline{1,(N_1 - L)} \tag{7}$$

на этапе отбора лучших моделей по критерию регулярности использовались только табличные значения. При вычислении прогноза ”на экзамене” в 17-ой (L=2 и 18-ой (L=3 точках использовались последовательно соответственно прогнозные значения $\hat{Y}_{16}$ и $\hat{Y}_{16}, \hat{Y}_{17}$, если аргументы Y_{k+1} и Y_{k+1}, Y_{k+2} входили в структуру отобранных лучших моделей.

Из сравнения графиков (рис. 4,5) можно сделать вывод, что прогнозирование при квазиоптимальном разбиении более точное, чем при всех других рассмотренных способах разбиений. Кроме того, моделирование с адаптивным прогнозом более эффективно, т.к. в большей мере учитывает информацию об исследуемом процессе в непосредственной близости к точкам прогноза. Как видно из графиков (рис. 2 – 5), наилучший прогноз получается по моделям квазиоптимального разбиения. Значения критериев регулярности, максимального отклонения и среднеквадратической ошибки (СКО) на данных “рабочей” таблицы приблизительно сопоставимы для моделей различных разбиений.

Для сравнения точности моделей на рис. 6 изображен ”коридор” среднеквадратической ошибки при квазиоптимальном способе разделения данных для всех вышерассмотренных примеров прогнозирования (рис.2—5). График ошибки на “экзамене” (*EP*) в логарифмической шкале значений для тех же примеров имеем на рис. 7, где

$$EP = \frac{\sum_{i=1}^{i=n_C} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{i=n_C} (Y_i - \overline{Y})}, \quad \text{где} \quad \overline{Y} = \frac{1}{n_C} \sum_{i=1}^{n_C} Y_i \quad .$$

Числа-отметки по оси абсцисс (рис. 6,7) соответствуют номерам предыдущих рисунков и связанным с ними особенностями прогнозирования (разное число упреждений, различные шаблоны, модели с адаптацией прогноза и без него)

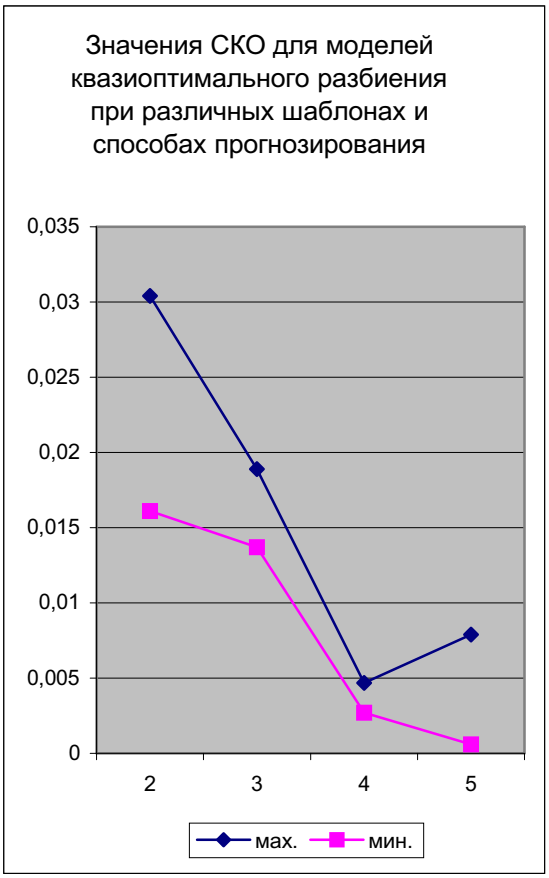

Рис. 6

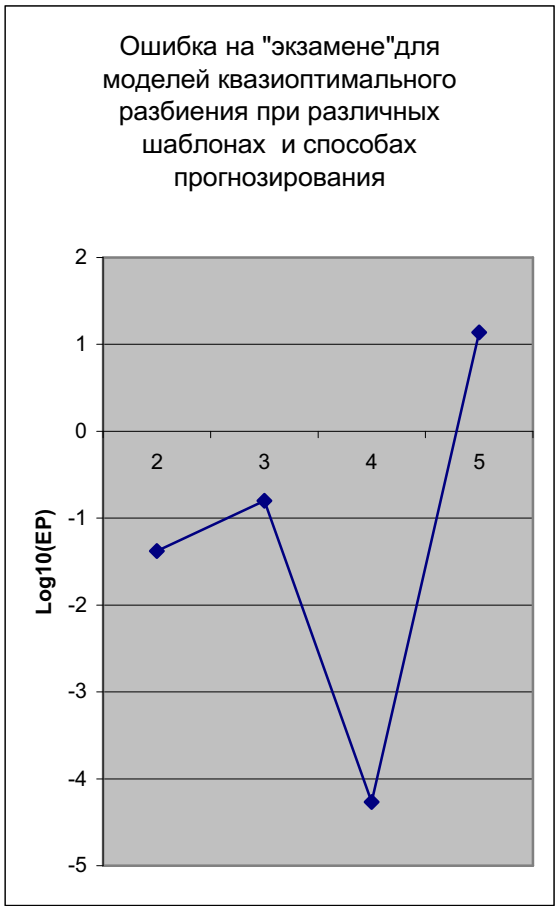

Рис. 7

Кривые максимумов (мах) и минимумов (мин.) (рис. 6) отражают диапазон, в котором изменяется СКО для моделей прогнозирования на число шагов $L = \overline{1, n_C}$. Из графиков (рис. 6) видно, что для моделей вида (5), (6) квазиоптимального разбиения с адаптивным прогнозом (рис. 4), диапазон СКО является минимальным. Шаблоны (5) - (7), содержащие указанный набор запаздывающих аргументов, является оптимальным с точки зрения точности моделирования на всех $N_2(L)$ точках (рис.4, 5).

Из графика рис. 7 видно, что модели вида (6), содержащие лаговые переменные, с адаптивным прогнозом наиболее точны на “экзамене”.

Выводы. Анализ влияния способа разбиения данных на точность прогнозирования на примере моделирования индекса инфляции позволяет сделать следующие выводы:

1. С ростом числа шагов прогноза увеличивается различие способов разбиения данных в алгоритмах МГУА в смысле влияния на точность прогнозирования;

2. Модели с адаптацией прогноза существенно повышают точность прогнозирования рассмотренного процесса;

3. Квазиоптимальное разбиение выборки с адаптивным прогнозом по модели, содержащей запаздывающие аргументы, позволило получить наиболее точный результат прогнозирования.

Список литературы.

1. Степашко В.С., Коппа Ю.В., Опыт применения системы АСТРИД для моделирования экономических процессов по статистическим данным.// Кибернетика и вычислительная техника.– 1998.– вып. 117. – с.24–31.

2. Коппа Ю.В., Степашко В.С. Порівняння ефективності застосування регресійного аналізу та МГВА для прогнозування економичних процесів. – Праці I Міжнародної конференції з індуктивного моделювання, Львів, 20-25 травня 2002: Т. 3. – с. 123-128.

3. Степашко В.С., Кондрашова Н.В. Исследование способов генерации вариантов разбиения выборки в алгоритмах МГУА. – Праці I Міжнародної конференції з індуктивного моделювання, Львів, 20-25 травня 2002: Т. 1.,Ч. 1. – с. 90-94.

4. Мостелер Ф., Тьюки Дж. Анализ данных и регрессия: В 2-х вып. Вып. 2. – М: Финансы и статистика, 1982. – 239с.

Кондрашова Н.В.

Влияние способа разбиения выборки в алгоритмах МГУА на точность прогнозирования

На примере процесса развития инфляции исследовано, в какой мере способ разделения выборки данных в различных алгоритмах МГУА определяет точность предсказания резкого изменения характера процесса. Сделан вывод об эффективности алгоритма адаптивного прогнозирования при квазиоптимальном разбиении данных.

Кондрашова Н.В.

Вплив способу розбиття вибірки в алгоритмах МГУА на точність прогнозування

На прикладі процесу розвитку інфляції досліджено, якою мірою спосіб розбиття вибірки даних у різних алгоритмах МГУА визначає точність прогнозування різкої зміни характеру процесу. Зроблено висновок про ефективність алгоритму адаптивного прогнозування з квазіоптимальним розбиттям даних.

Kondrashova N.V.

Influence of a sample division method in GMDH algorithms on forecasting accuracy

For an example of inflation development process, it is investigated to what extent a method of division of data sample in GMDH algorithms defines prediction accuracy of a sharp changing of a process character. Conclusion about efficiency of an algorithm of adaptive forecasting with a suboptimal data division is drawn.