Математические структуры и моделирование
2017. № 2(42). С. 105–114

УДК 004.67

ИДЕНТИФИКАЦИЯ ПАРАМЕТРИЧЕСКОЙ МОДЕЛИ ИЗМЕНЕНИЯ ЧИСЛЕННОСТИ ЗАЧИСЛЕННЫХ АБИТУРИЕНТОВ ПО МАЛОЙ ВЫБОРКЕ

Д.Н. Лавров
к.т.н., доцент, e-mail: lavrov@omsu.ru
А.А. Лаптев
к.ф.-м.н., доцент, e-mail: aalaptev@mail.ru
М.А. Мамонтова
к.и.н., доцент, e-mail: omichka@inbox.ru

Омский государственный университет им. Ф.М. Достоевского

Аннотация. Построена параметрическая модель изменения численности абитуриентов университета по малой выборке. В качестве входного параметра взято количество выпускников школ региона. С помощью информационных критериев выбран порядок авторегрессии-скользящего среднего. Применяя функцию по идентификации систем пакета Scilab, получили прогноз по количеству абитуриентов для приёмной кампании 2017 года. Идентичный результат получен через идентификацию модели в виде передаточной функции.

Ключевые слова: идентификация параметрической модели, модель авторегрессии-скользящего среднего, прогноз количества абитуриентов вуза, информационный критерий Акайке.

Введение

Социальные и демографические процессы, развивающиеся в российском обществе, болезненно сказываются на развитии системы образования. На фоне затяжной демографической ямы происходит значительное сокращение количества выпускников. При общем росте численности населения за 2009–2016 гг. (с 142,7 млн. человек в 2009 г. до 146,5 млн. человек в 2016 г., 146,8 млн. на 20.03.2017 г.)[1] происходит сокращение количества выпускников средних общеобразовательных учреждений (с 898,9 тыс. выпускников в 2009 г. до 716,9 тыс. выпускников в 2017 г.). В 2017 г. система высшего профессионального образования получит наименьшее количество выпускников школ по сравнению с предшествующим периодом (2009–2016 гг.) и прогнозируемым количеством выпускников последующих годов (в 2018–2025 гг. прогнозируется рост с 729,1 тыс. человек в 2018 г. до 915 тыс. в 2025 г.).

[1]По данным Федеральной службы государственной статистики URL: `http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/resources/33a9d8004a4bb6b681e3c9a8ffbe44e3/table.jpg` (дата обращения 01.05.2017 г.)

Сегмент абитуриентов вуза, поступающих на программы бакалавриата и специалитета, складывается из выпускников школ РФ (как записавшихся на ЕГЭ, так и имеющих особые права и сдающих вступительные испытания вуза), выпускников средних (начальных) профессиональных учреждений РФ, иностранных граждан, поступающих в высшие учебные заведения РФ, а также из уже окончивших высшие учебные заведения и получающих второе высшее образование. Последняя категория, как правило, выбирает заочную форму обучения и не имеет права на получение образования за счёт бюджетных средств (п. 3. статьи 5 Федерального закона «Об образовании в Российской Федерации» от 29.12.2012 г. № 273-ФЗ). Выпускники средних профессиональных учреждений тоже большей частью ориентированы на заочную форму обучения, на которую выделяется значительно меньшее количество бюджетных мест. Возможно, при таком выборе сказывается их вовлечённость в выбранную профессиональную сферу и раннее трудоустройство, неуверенность в своих способностях в сравнении с выпускниками-школьниками. Таким образом, основную массу абитуриентов на очную форму обучения, в том числе на бюджетные места, составляют бывшие одиннадцатиклассники, а также выпускники прошлых лет, чьи результаты ЕГЭ действительны на момент поступления. Однако при рассмотрении этой категории необходимо учитывать, что в большинстве своём на перепоступление решаются студенты первого курса и в основном девушки.

За период с 2011 г. по 2016 г. наблюдается значительное сокращение количества выпускников (с 11368 выпускников в 2011 г. до 8972 выпускника в 2016 г.), а соответственно, и количества абитуриентов, зачисленных в ОмГУ им. Ф.М. Достоевского. Например, наибольшее количество зачисленных абитуриентов выпадает на 2012 г. — 2568 человек, когда в Омской области было наибольшее количество выпускников за обозначенный период — 11916 человек. Доля выпускников школ из общего количества поступивших в 2012 г. тоже была высока — 1167 человек по сравнению с предыдущим 2011 г. (1522 человека) и последующим — 2013 г. (1583 человека). С 2013 г. наблюдается значительный спад как количества выпускников, так и количества поступивших на основе среднего общего образования. Прогнозируя дальнейшее сокращение выпускников по Омской области в 2017 г., авторы предприняли попытку просчитать количество выпускников средних образовательных учреждений, которые будут зачислены в ОмГУ им. Ф.М. Достоевского с 1 сентября 2017 г.

Количество выпускников в этот период уменьшится ещё примерно на 40 тысяч и составит по прогнозам около 710 тысяч. Этот уровень будет самым низким не только за последние годы, но и в будущем такого малого числа выпускников не ожидается. Начиная с 2018 года начнётся увеличение количества 11-классников. Вот прогноз на 2025 год — 915 тысяч, то есть ожидается увеличение более чем на 200 тысяч выпускников.

1. Исходные данные для идентификации

В качестве количественной оценки выберем число поступивших на программы бакалавриата и специалитета. В качестве параметра (входной сигнал)

возьмём количество выпускников школ Омской области или города Омска.

Часть абитуриентов поступает на программы бакалавриата или специалитета не сразу после получения среднего образования, а после обучения на программах среднего или высшего профессионального образования. Поэтому их исключим из рассмотрения. На основе отчётов приёмной комиссии ОмГУ им. Ф.М. Достоевского за 2011–2016 годы получили следующие данные по зачисленным абитуриентам, поступившим на базе среднего образования.

Таблица 1. Число абитуриентов, поступивших в ОмГУ

Год	2011	2012	2013	2014	2015	2016
Количество	1522	1667	1583	987	1019	947

В модели пренебрегаем абитуриентами из других регионов России и из Казахстана, предполагая схожесть демографической картины в этих регионах и учитывая, что основная часть абитуриентов, как правило, обучается в Омской области (в среднем 79,7% за 2014-2016 гг.).

Количество выпускников школ Омской области и города Омска по данным, ежегодно озвучиваемым в прессе перед приёмной кампанией:

Таблица 2. Общее число выпускников школ

Год	2011	2012	2013	2014	2015	2016	2017
Омская область	11368	11916	11410	10502	9357	8972	8872
г. Омск	5712	6260	6090	5755	5372	4887	4833

2. Выбор порядка модели

Для выбора порядка модели ARMA(p, q) (авторегрессии-скользящего среднего) наиболее популярен информационный критерий Акайке (AIC) [1, p. 141]. Для модели линейной регрессии он записывается как

$$AIC = 2K + N[\ln(\widehat{\sigma}^2)]$$

$$SSE = ||f(x_i) - y_i||_2^2 = \sum_{i=1}^{N}(y_i - f(\theta, x_i))^2$$

$$\widehat{\sigma}^2 = \frac{SSE}{N-2},$$

где $K = p + q$ — порядок модели, p – порядок регрессионой части, q – порядок скользящего среднего, n — размер выборки, SSE – квадрат среднеквадратической ошибки предсказания, $\widehat{\sigma}^2$ — оценка дисперсии предсказания, $f(x_i)$ — значение выходного сигнала модели на отсчёте x_i; y_i — наблюдаемое значение выходного сигнала, θ — вектор параметров модели.

Для коротких выборок добавляют штраф за количество параметров модели

$$AIC_c = AIC + \frac{2K \cdot (K+1)}{N-K-1}.$$

Ещё одним критерием является критерий «финальной ошибки предсказания» (Final Prediction Error) [1, p. 141]

$$FPE = \frac{N+K+1}{N-K-1}\widehat{\sigma}^2.$$

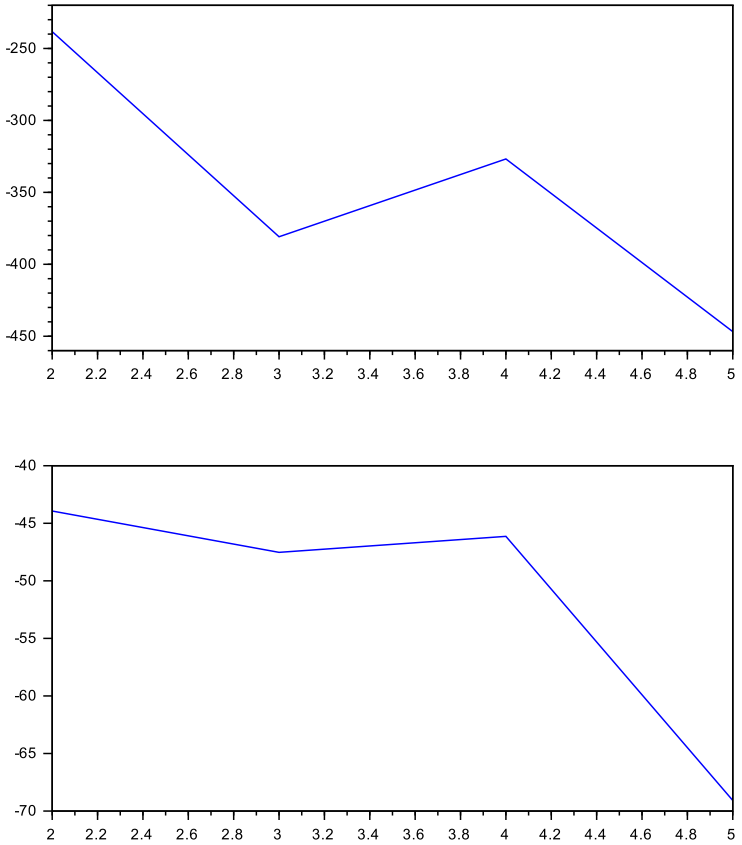

Рис. 1. Порядок модели p по оси абсцисс. Сверху: AIC — по оси ординат, снизу: $\ln$ FPE – по оси ординат

Минимум каждого критерия и даёт нам подходящий порядок модели.

В результате проведения компьютерного моделирования было установлено, что наиболее адекватные (при моделировании согласующиеся с наблюдаемыми данными) параметрические модели ARMA(p, q) получаются при $p = q$ и p, $q > 2$.

С другой стороны выборка крайне мала ($n = 6$), поэтому выбор модели может быть осуществлён в коротком интервале от $p = 2$ до 5. Значение $p = 2$

взято для контроля. Значение $p = 5$ будет давать максимальное соответствие (минимальную ошибку предсказания), но в критерии Акайке сумма квадратов ошибки предсказания при этом будет заведомо очень близка к нулю. Параметры модели при ($p = q = 5$) вычисляются по наименьшему числу уравнений, а значит, оценка более чувствительна к точности данных. В этом случае более предпочтительной моделью будет модель меньшей размерности. Однако расчёты были произведены и для случая $p = q = 5$. На рисунке 1 показаны результаты расчёта AIC и FPE.

Из графиков видно, что при $p = 3$ достигается локальный минимум, а случай $p = 5$ хоть и даёт лучший результат по сравнению с $p = 3$, но, как упомянуто выше, является предельным результатом, а потому его следует избегать для построения прогноза.

Так как отношение поступивших по области к поступившим по городу из года в год примерно одинаковое, то косвенным критерием качества модели будет минимальное расхождение предсказания по области и по городу. Этот критерий был использован для первоначального отсева моделей ($p \neq q$) и будет использован далее для сравнения с основным предсказанием.

На рисунке 2 представлен код на языке Scilab [2], использующий стандартную функцию armax для идентификации систем ARMA и моделирования поведения временных рядов.

На рисунке 3 представлены результаты моделирования.

3. Идентификация и прогнозирование

Будем использовать параметрический метод идентификации модели ARMA(3,3) исследуемой системы, который производится на основе априорной информации. Построим модель в форме передаточной функции. Эта модель определяется коэффициентами числителя и знаменателя.

Будем предполагать, что структура модели соответствует передаточной функции линейной дискретной системы в виде:

$$G(z) = \frac{B(z)}{A(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3}}{1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3}}.$$

Или в операторной форме:

$$A(z)Y(z) = B(z)U(z);$$

$$[1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3}]Y(z) = [b_0 + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3}]U(z),$$

где $Y(z)$ и $U(z)$ — z-преобразования выходного и входного сигнала соответственно. Рассматривая z^{-1} как оператор запаздывания, получим разностное уравнение:

$$y_t + a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} = b_0 u_t + b_1 u_{t-1} + b_2 u_{t-2} + b_3 u_{t-3}.$$

Тогда на выходе получим:

$$y_t = b_0 u_t + b_1 u_{t-1} + b_2 u_{t-2} + b_3 u_{t-3} - a_1 y_{t-1} - a_2 y_{t-2} - a_3 y_{t-3}.$$

```
// поступившие
y = [1522 1667 1583 987 1019 947]
// абитуриентов по области
u = [11368 11916 11410 10502 9357 8972]
uu = [u 8872] // абитуриентов по области с данными по 2017.
n=length(u); nn=length(u); m=length(y);
t = 2011:2018;
label=["black -- ^","black -- +","black -- *","black -- o"]
T=7; eps=1e-30
SSE=[]; AIC=[]; FPE=[]; pr=[]
scf(0)
subplot(211)
r=[2011,800,2017.5,2000]
plot2d(t(1:length(y)),y,rect=r)
subplot(212)
plot2d(t(1:length(y)),y,rect=r)
subplot(211)
inter=2:5
for p=inter
 q=p
 ArmaEst=armax(p,q,y,u); // идентифицируем модель
 yy=arsimul(ArmaEst,uu((q+1):nn),uu(q:-1:1),y(p:-1:1),[]);
 yy=[y(1:p) yy]
 plot(t(1:T),yy(1:T),label(p-inter(1)+1))
 sse=sum((y-yy(1:6)).^2)
 SSE=[SSE sse]
 K=p+q
 AIC=[AIC 2*K+n*log((sse+eps)/(n-2))+2*(K+1)*(K)/(n-K-1)]
 FPE=[FPE log((n+K+1)/(n-K-1)*sse)]
 up=u(n:-1:n-q+1)
 yp=y(m:-1:m-p+1)
 yy = arsimul(ArmaEst, [8872], up, yp, []);
 pr=[pr yy] // Предсказание на 2017 год.
end
```

Рис. 2. Листинг кода моделирования идентифицированной системы

Идентификация модели сводится к поиску коэффициентов a_1, a_2, a_3, b_0, b_1, b_2, b_3, дающих результат с минимальной ошибкой предсказания.

Ошибка предсказания (рассогласование выходного сигнала модели $\widehat{y}_t$ и системы y_t): $\varepsilon_t = y_t - \widehat{y}_t$. Предполагается, что имеется последовательность дискретных отсчётов входного (u_t) и выходного (y_t) сигналов системы.

Для каждого момента времени составим уравнение:

$$y_t - [b_0 u_t + b_1 u_{t-1} + b_2 u_{t-2} + b_3 u_{t-3} - a_1 y_{t-1} - a_2 y_{t-2} - a_3 y_{t-3}] = 0.$$

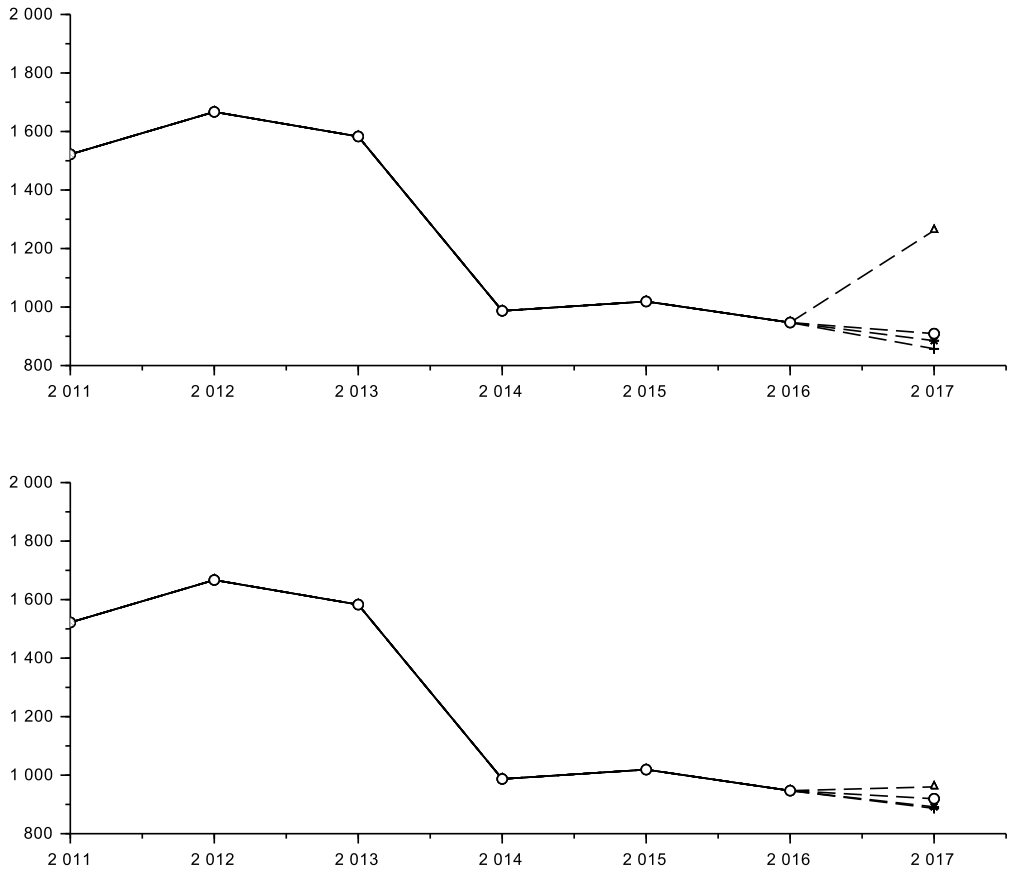

Рис. 3. Моделирование временного ряда абитуриентов по годам. Сверху моделирование ряда и предсказание по количеству абитуриентов в области, снизу — по количеству абитуриентов в городе. Метки моделируемых рядов: Δ – ARMA(2,2), $+$ – ARMA(3,3), $*$ – ARMA(4,4), $\circ$ – ARMA(5,5)

Получим систему уравнений:

$$\begin{aligned} y_t &= b_0u_t + b_1u_{t-1} + b_2u_{t-2} + b_3u_{t-3} - a_1y_{t-1} - a_2y_{t-2} - a_3y_{t-3}, \\ y_{t-1} &= b_0u_{t-1} + b_1u_{t-2} + b_2u_{t-3} + b_3u_{t-4} - a_1y_{t-2} - a_2y_{t-3} - a_3y_{t-4}, \\ y_{t-2} &= b_0u_{t-2} + b_1u_{t-3} + b_2u_{t-4} + b_3u_{t-5} - a_1y_{t-3} - a_2y_{t-4} - a_3y_{t-5}. \end{aligned}$$

В матричном виде система:

$$y = \Phi\theta, \tag{1}$$

здесь $\theta = (b_0, b_1, b_2, b_3, -a_1, -a_2, -a_3)'$ — вектор параметров, Φ — матрица, составленная из отчётов входов и выходов системы, $y = (y_t, y_{t-1}, y_{t-2})'$.

$$\Phi = \begin{pmatrix} u_t & u_{t-1} & u_{t-2} & u_{t-3} & y_{t-1} & y_{t-2} & y_{t-3} \\ u_{t-1} & u_{t-2} & u_{t-3} & u_{t-4} & y_{t-2} & y_{t-3} & y_{t-4} \\ u_{t-2} & u_{t-3} & u_{t-4} & u_{t-5} & y_{t-3} & y_{t-4} & y_{t-5} \end{pmatrix}.$$

Решение системы (1) невозможно найти в общем виде в силу переопределённости и возможной несовместности системы. Будем искать псевдорешение. Псевдорешением системы называется такой вектор, который при подстановке в (1) даёт вектор невязки, минимальный по евклидовой норме (метод наименьших квадратов).

Решение этой задачи хорошо известно и даётся выражением:

$$\widehat{\theta} = \Phi^{\#} y,$$

где $\Phi^{\#} = (\Phi'\Phi)^{-1}\Phi'$ — псевдообратная матрица.

Одношаговый прогноз вычисляется по формуле:

$$y_{t+1} = b_0 u_{t+1} + b_1 u_t + b_2 u_{t-1} + b_3 u_{t-2} - a_1 y_t - a_2 y_{t-1} - a_3 y_{t-2}.$$

Ниже приведён код, реализующий данный подход.

```
// Идентифицируем систему
Phi=[8972  9357 10502 11410    1019  987 1583;
     9357 10502 11410 11916     987 1583 1667;
    10502 11410 11916 11368    1583 1667 1522]
y=[947; 1019; 987]
theta =  pinv(Phi)*y
disp(theta')
// Строим прогноз
uy=[8872 8972 9357 10502   947 1019 987]'
y_2017=sum(theta.*uy)
disp(y_2017)
```

Результаты расчётов, проведённых по указанным формулам, дают точно такие же результаты, что и встроенная в Scilab функция armax, что подтверждает правильность проведённых вычислений и исключает ошибки программиста.

4. Результаты прогнозирования

Все результаты предсказания для различных моделей собраны в таблице 3. Полужирным шрифтом выделено наше предпочтение, на выбор которого указывает информационный критерий Акайке.

Таблица 3. Результаты одношагового предсказания

Модель	2017 (область)	2017 (город)
ARMA(2,2)	1261,8129	960,01816
ARMA(3,3)	**857,2997**	886.58083
ARMA(4,4)	885,03696	891,59887
ARMA(5,5)	909.10012	919,41979

Итак, итоговый прогноз мы должны выбрать по модели ARMA(3,3) и это — 857 абитуриентов. Это минимальное из всех значений предсказание.

Заключение

С точки зрения управленческих решений наихудший прогноз приёма — лучший. Он стимулирует принятие решений по преодолению влияния демографического кризиса на приёмную кампанию.

Представленная модель далека от идеала, её реальная полезность будет ясна после проведения приёмной кампании 2017 года. По результатам кампании будет принято решение о доработке и уточнении модели.

Таким образом, используемый параметрический метод идентификации модели исследуемой системы, производимый на основе априорной информации, позволил построить модель в форме передаточной функции. Привлечённые количественные данные выпускников школ города Омска и Омской области, абитуриентов Омского государственного университета им. Ф.М. Достоевского за 2011–2016 годы, позволили смоделировать ситуацию приёмной кампании 2017 г. Построенная модель дала возможность вывести количество абитуриентов предстоящей приёмной кампании 2017 года — 857 человек. При устойчивой тенденции снижения количества выпускников и, соответственно, абитуриентов, учитывая сползание в «демографическую яму» в последние несколько лет, авторы сумели просчитать общее количество абитуриентов, что поможет оценить риски приёмной кампании на программы бакалавриата и специалитета в 2017 г.

Благодарности

Благодарим сотрудников Регионального информационно-аналитического центра системы образования Омской области Надежду Изаровну Жолобову и Анастасию Викторовну Коротченко за помощь в сборе статистки выпускников по Омской области и г. Омску.

Литература

1. Topics in Applied Physics. Nonlinear Methods of Spectral Analysis. Haykin, S. (Ed.). V. 34. Springer-Verlag, 1979.
2. Scilab. Open source software for numerical computation. [Заголовок с экрана]. URL: `http://www.scilab.org/`.

IDENTIFICATION OF A PARAMETRIC MODEL OF CHANGES IN THE NUMBER OF STUDENTS ENROLLED ON A SMALL SAMPLE

D.N. Lavrov
Ph.D. (Eng.), Associate Professor, e-mail: lavrov@omsu.ru
A.A. Laptev
Ph.D. (Phys.-Math.), Associate Professor, e-mail: aalaptev@mail.ru
M.A Mamontova
Ph.D. (History), Associate Professor, e-mail: omichka@inbox.ru

Dostoevsky Omsk State University

Abstract. A parametric model of changes in the number of students of the University with a small sample is built. The number of high school graduates in the region is taken as an input parameter. With the help of information criteria the order of autoregressive moving-average is selected. Applying the identification system feature of Scilab, the forecast of the number of applicants for admission campaign in 2017 was made. A similar result was obtained through the identification of the model in the form of the transfer function.

Keywords: parametric identification, autoregressive-moving-average model, forecast of the number of high school students, Akaike information criterion.

Дата поступления в редакцию: 10.05.2017