

Интернет-журнал «Наукоедение» ISSN 2223-5167 <http://naukovedenie.ru/>

Том 8, №4 (2016) <http://naukovedenie.ru/index.php?p=vol8-4>

URL статьи: <http://naukovedenie.ru/PDF/81TVN416.pdf>

Статья опубликована 06.09.2016.

Ссылка для цитирования этой статьи:

Новиков Е.И. Оценивание параметров линейных регрессионных моделей с учетом запаздывания влияния факторов на зависимую переменную // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 8, №4 (2016) <http://naukovedenie.ru/PDF/81TVN416.pdf> (доступ свободный). Загл. с экрана. Яз. рус., англ.

УДК 519.237.5 + 517.977.58

Новиков Евгений Иванович

ФГКВОУ ВО «Академия Федеральной службы охраны Российской Федерации», Россия, Орел¹

Доцент кафедры «Автоматизированные информационные системы»

Кандидат технических наук

E-mail: nei05@rambler.ru

Оценивание параметров линейных регрессионных моделей с учетом запаздывания влияния факторов на зависимую переменную

Аннотация. В статье рассматривается проблема параметрической идентификации многофакторных линейных регрессионных моделей, описывающих зависимости между социально-экономическими показателями, представленными временными рядами. Указывается на основную причину сложности идентификации подобных моделей – наличие времени запаздывания влияния контролируемых факторов на зависимый показатель. Актуальность задачи оценивания времени запаздывания обусловлена отсутствием в настоящее время единых формализованных подходов к ее решению. Представлена обобщенная структура регрессионной модели, учитывающей время запаздывания влияния контролируемых факторов на зависимую переменную. Для использования методов оценивания параметров линейных моделей предлагается определять время запаздывания как независимый параметр модели регрессии. Проанализированы различные способы решения этой задачи. Описан предлагаемый подход к оцениванию времени запаздывания, основанный на использовании генетического алгоритма. Отмечены выявленные в ходе проведенных исследований особенности применения алгоритма, влияющие на точность получаемых на его основе оценок времени запаздывания. Дополнительно к внутренним критериям качества подбора неизвестных значений времени запаздывания предложено применить внешние критерии, основанные на сравнении наблюдаемых и модельных значений зависимой переменной и характеризующих качество спецификации регрессионной модели. Выбор конкретного критерия зависит от прикладных целей исследования. Автором предложена методика оценивания параметров регрессионных моделей с учетом времени запаздывания факторов влияния, применение которой позволяет повысить точность регрессионных моделей анализа и прогнозирования социально-экономических показателей.

Ключевые слова: статистические данные; временной ряд; многофакторная линейная регрессионная модель; параметры модели; лаговая переменная; время запаздывания; генетический алгоритм; интерполяция; статистическая значимость; точность модели

¹ 302034, г. Орёл, ул. Приборостроительная, д. 35

Одним из основных способов исследования социально-экономических процессов является их математическое моделирование. Математическая модель позволяет проводить анализ причинно-следственных связей между изучаемыми явлениями, а также получать прогнозы их развития.

Модель может быть построена либо на основании известных физических законов, либо по имеющимся ретроспективным данным, либо, в случае невозможности применения первых двух способов, путем имитационного моделирования. Исходя из того, что в настоящее время собирается и хранится информация по большому количеству различных социально-экономических показателей (например, базы данных Федеральной службы государственной статистики) наибольшее распространение получили экспериментальные модели, формируемые на основе обработки и анализа массивов статистических данных.

Для построения таких математических моделей используется корреляционно-регрессионный анализ, позволяющий количественно описать стохастические зависимости между анализируемыми переменными. Так как в большинстве случаев изменение значений исследуемых социально-экономических показателей определяется множеством причин, то из всего многообразия регрессионных моделей на практике наиболее часто применяются линейные (в том числе квазилинейные) многофакторные модели, поскольку для их оценивания разработаны специальные методы, а также они позволяют интерпретировать полученные зависимости между исследуемыми показателями.

Экспериментальное моделирование социально-экономических процессов может быть основано на данных двух видов. Первый вид – пространственные данные, которые описывают исследуемые показатели, значения которых измерены на различных объектах (например, субъектах) в один момент времени. Второй вид – временные ряды, представляющие исследуемые показатели, полученные на одном объекте в различные, как правило, равноотстоящие друг от друга дискретные моменты времени (месяц, квартал, год).

Применение корреляционно-регрессионного анализа для построения пространственных моделей представляет собой несложную задачу, методы и средства решения которой достаточно хорошо отработаны и подробно описаны в различных учебных изданиях и научных публикациях [1, 2]. Общий вид такой модели можно описать следующим выражением:

$$y(t) = b_0 + b_1 \cdot x_1(t) + b_2 \cdot x_2(t) + \dots + b_k \cdot x_k(t), \quad (1)$$

где: $y(t)$ – значение исследуемого социально-экономического показателя, измеренное в момент времени t ; x_i – значения контролируемых факторов, предположительно обуславливающих изменение исследуемого показателя, измеренные в момент времени t ($i = 1, k$, где k – количество факторов влияния); b_i – оценки параметров модели, определяющие степень влияния факторов на зависимый показатель.

В случае построения математических моделей, описывающих зависимости социально-экономических показателей, значения которых представлены в виде временных рядов, возникает необходимость реализации дополнительных процедур, которые учитывают особенности применения регрессионного анализа в части методов оценивания параметров моделей.

Основной проблемой при построении модели регрессии временных рядов является наличие запаздывания факторов влияния на зависимую переменную. Такая ситуация обусловлена тем, что при исследовании социально-экономических процессов факторные показатели в большинстве случаев определяют изменение зависимого показателя не

одномоментно, а по истечении некоторого промежутка времени, называемого лагом. Например, если рассмотреть зависимость между уровнем преступности и уровнем миграции, то можно увидеть, что для большинства субъектов с увеличением числа мигрантов возрастает уровень преступности. Однако такая тенденция проявляется не сразу, а через определенные временные периоды.

Для учета этого обстоятельства регрессионная модель должна быть дополнена переменной, характеризующей время запаздывания влияния факторов на изменение зависимого показателя [3]:

$$y(t) = b_0 + b_1 \cdot x_1(t - \tau_1) + b_2 \cdot x_2(t - \tau_2) + \dots + b_k \cdot x_k(t - \tau_k), \quad (2)$$

где τ_i – время запаздывания влияния контролируемых факторов на значение зависимого показателя ($i = 1, k$, где k – количество факторов влияния).

Модель вида (2) является нелинейной. Нелинейные модели значительно усложняют расчеты и в большинстве случаев методы, применяемые для оценивания параметров таких моделей, отсутствуют, либо трудно формализуемы. Поэтому целесообразно свести модель (2) к линейной форме, для чего предлагается рассматривать время запаздывания τ_i как независимый параметр и находить его оценку непосредственно перед нахождением оценок параметров b_i .

В настоящее время предложено несколько подходов к решению задачи оценивания времени запаздывания, среди которых наиболее разработанными являются подходы, применяемые в эконометрике для моделей с распределенными лагами типа [4, 5]:

$$y(t) = b_0 + b_1 \cdot x(t) + b_2 \cdot x(t - 1) + \dots + b_k \cdot x(t - k) \quad (3)$$

Для идентификации таких моделей существуют известные методы, среди которых наибольшее распространение получили методы геометрической прогрессии, Койка и Алмон [4, 5, 6].

Однако идентификация регрессионных моделей вида (2) с применением этих методов практически невозможна, поскольку они позволяют получить адекватные результаты только в случае рассмотрения одного фактора влияния в различные моменты времени (модель (3)). В связи с этим актуальной является задача разработки альтернативных подходов к оцениванию времени запаздывания влияния факторов на зависимые показатели.

По своей постановке эта задача относится к задачам поиска оптимального значения параметра – времени запаздывания – в соответствии с заданным критерием. Для ее решения можно предложить несколько способов:

- визуальный анализ графиков динамики зависимого показателя и факторов влияния;
- анализ корреляционной функции;
- полный перебор возможных значений времен запаздывания.

Несложный анализ этих способов позволяет сделать следующие выводы. Визуальный анализ слабо формализован с точки зрения критерия оптимизации, поэтому его можно использовать только для оценивания интервала изменения значений времен запаздывания. Метод полного перебора является самым точным, однако имеет большую вычислительную сложность. Например, вычислительная сложность алгоритма, реализующего этот метод, при

числе факторов влияния равном 4 и интервале запаздывания 90 дней составляет 4^{90} операций перебора. Это свидетельствует о невозможности использования этого подхода для определения время запаздывания. При использовании корреляционной функции, во-первых, сложно определить нижнюю границу значений коэффициента корреляции, начиная с которой можно говорить о наличии зависимости между фактором влияния и зависимым показателем, во-вторых, при достаточно широком интервале возможных значений времен запаздывания теряется большой объем исходных, что приводит к значительному уменьшению статистической значимости получаемых значений коэффициентов корреляции.

Исходя из вышеизложенного, для оценивания времен запаздывания τ_i предлагается использовать подход, основанный на генетическом алгоритме. Предпосылками такого решения являются результаты успешного применения генетических алгоритмов для идентификации различных математических моделей, содержащих параметры, аналогичные времени запаздывания в модели (2) [7, 8].

Схема функционирования генетического алгоритма для оценивания времени запаздывания является стандартной и основана на использовании операторов скрещивания (кроссовер), мутации и инверсии. Механизмы реализации этих операторов подробно описаны в различных учебных изданиях [9, 10]. Поэтому рассмотрим только выявленные в ходе исследования особенности применения генетического алгоритма для идентификации моделей вида (2).

1. Перед реализацией генетического алгоритма необходимо провести интерполяцию исходных значений показателей. Это обуславливается тем, что измерение социально-экономических показателей производится с периодичностью месяц, квартал, полугодие, год, хотя характеризующие ими социально-экономические процессы развиваются непрерывно. Интерполяция позволяет восстановить «пропущенные» значения показателей, а механизмы ее реализации обеспечивают возможность выбора исследователем необходимо периода дискретизации (неделя, день и т.д.). Как следствие в результате интерполяции исходных значений показателей увеличивается точность оценивания времени запаздывания. Для решения этой задачи предлагается использовать интерполяцию кубическими сплайнами, основанную на разбиении интервала изменения анализируемых социально-экономических показателей на небольшие отрезки, на каждом из которых функция задается полиномом третьей степени.
2. Число генов в хромосоме определяется числом контролируемых факторов.
3. Для отбора наиболее приспособленных особей целесообразно использовать метод рулетки.
4. Ключевую роль в успешном функционировании генетического алгоритма играет корректное определение целевой функции, называемой фитнес-функцией, либо функцией пригодности. Значение этой функции показывает степень пригодности каждой особи или точность получаемого с использованием найденной особи решения.

Проведенные исследования показали, что для задачи определения времени запаздывания факторов влияния вид целевой функции целесообразно определять в зависимости от прикладного характера использования модели (2). При решении задачи выявления факторов, статистически значимо влияющих на зависимый показатель, в качестве фитнес-функции целесообразно применить стандартную ошибку оценки. Тогда критерий

пригодности полученного с помощью генетического алгоритма решения будет иметь следующий вид:

$$s_{\varepsilon}^2 = \frac{\sum_{i=1}^n \varepsilon_i^2}{n - (k + 1)} \rightarrow \min, \quad (4)$$

где: s_{ε}^2 – оценка дисперсии возмущающей переменной ε , n – число наблюдений зависимого показателя и факторов влияния, k – число факторов влияния.

Выбор этого критерия обусловлен тем, что стандартная ошибка оценки характеризует ошибку спецификации регрессионной модели, что позволяет адекватно подобрать время запаздывания влияния контролируемых факторов на зависимую переменную.

При решении задачи прогнозирования динамики развития социально-экономических показателей в качестве целевой функции предлагается использовать средне абсолютную процентную ошибку (MAPE – Mean Absolute Percentage Error). Тогда критерий пригодности полученного с помощью генетического алгоритма решения будет иметь следующий вид [11]:

$$\text{MAPE} = \frac{1}{n} \cdot \sum_{i=1}^n \left(\frac{|y_i^{\text{мод}} - y_i^{\text{набл}}|}{y_i^{\text{набл}}} \right) \cdot 100 \% \rightarrow \min, \quad (5)$$

где: $y_i^{\text{набл}}$ – наблюдаемое значение социально-экономического показателя, полученное в ходе проведения пассивного эксперимента, $y_i^{\text{мод}}$ – модельное значение социально-экономического показателя, рассчитанное по уравнению вида (2).

Показатель MAPE в сравнении с другими показателями точности прогноза имеет критерии, позволяющие интерпретировать его возможные значения.

На основании вышеизложенных соображений можно описать методику оценивания параметров линейных регрессионных моделей с учетом запаздывания влияния факторов на зависимую переменную, включающую следующие основные этапы.

1. Интерполяция наблюдаемых значений исследуемых социально-экономических показателей с применением кубических сплайнов.
2. Построение графиков динамики изменения социально-экономических показателей на основе имеющихся в распоряжении исследователя ретроспективных данных.
3. Определение интервала, в пределах которого будет осуществляться оценивание времени запаздывания. Для этого можно использовать результаты визуального анализа графиков динамики показателей. Также рекомендуется учитывать экспертные оценки, отражающие особенности развития анализируемых социально-экономических процессов.
4. Определение начального значения времени запаздывания, с которого начнется работа генетического алгоритма.
5. Определение прикладных целей исследования – анализ влияния факторов либо прогнозирование значений показателей. Исходя из этого, выбирается вид

критерия пригодности – выражение (4) для первой задачи, выражение (5) – для второй.

6. Оценивание времени запаздывания влияния контролируемых факторов τ_i с применением внутренних критериев останова генетического алгоритма. Для этого можно использовать любую доступную прикладную программу. В качестве внутреннего критерия останова целесообразно применить критерий неизменности хромосом.
7. Подстановка найденных значений времен запаздывания в модель регрессии (2) и расчет оценок ее параметров b_i с помощью метода наименьших квадратов в любом статистическом пакете, например Statistica.
8. Расчет значения функции пригодности по выражениям (4) или (5), в зависимости от выбранного критерия с сохранением полученного результата.
9. Увеличение значения времени запаздывания на выбранную величину дискретизации (день, неделя и т.д.).
10. Повтор этапов 6-8 до тех пор, пока не будет выполнен внешний критерий останова, описываемый выражениями (4) или (5).

Для верификации описанной методики была проведена ее экспериментальная проверка. С этой целью были сформированы временные ряды, описывающие динамику изменения зависимой переменной и четырех факторов влияния. После этого рассчитаны значения времен запаздывания на основании внутреннего критерия неизменности хромосом и внешнего критерия, описываемого выражениями (4) и (5). При этом величина интервала изменения значений времен запаздывания была выбрана равной 120 дням.

В таблице 1 приведены рассчитанные значения времен запаздывания τ_i для каждого фактора, полученные по результатам проведения одной реплики эксперимента.

Таблица 1

Значения времен запаздывания τ_i факторов влияния (составлено автором)

Время запаздывания τ_i , (день)	Фактор 1, $x_1(t)$	Фактор 2, $x_1(t)$	Фактор 3, $x_1(t)$	Фактор 4, $x_1(t)$
	93	108	118	104

Как было отмечено выше, значения времен запаздывания τ_i , а также интервал варьирования времени запаздывания можно оценить с помощью визуального анализа графиков динамики анализируемого показателя и контролируемых факторов, выполнив интерполяцию исходных данных. Следовательно, если построить график динамики зависимого показателя по его наблюдаемым значениям и графики динамики контролируемых факторов по значениям, сдвинутым относительно значений зависимого показателя на найденные величины времен запаздывания τ_i , то в случае прямой зависимости соответствующие периоды увеличения значений показателя и факторов должны совпадать (в случае обратной зависимости между показателем и фактором влияния должны совпадать периоды увеличения и уменьшения их значений). Как видно из рисунка 1 времена запаздывания τ_i , найденные с помощью генетического алгоритма, совпадают с их соответствующими значениями, полученными в результате визуального анализа графиков временных рядов, построенных с применением пакета символьной математики Mathcad 14.

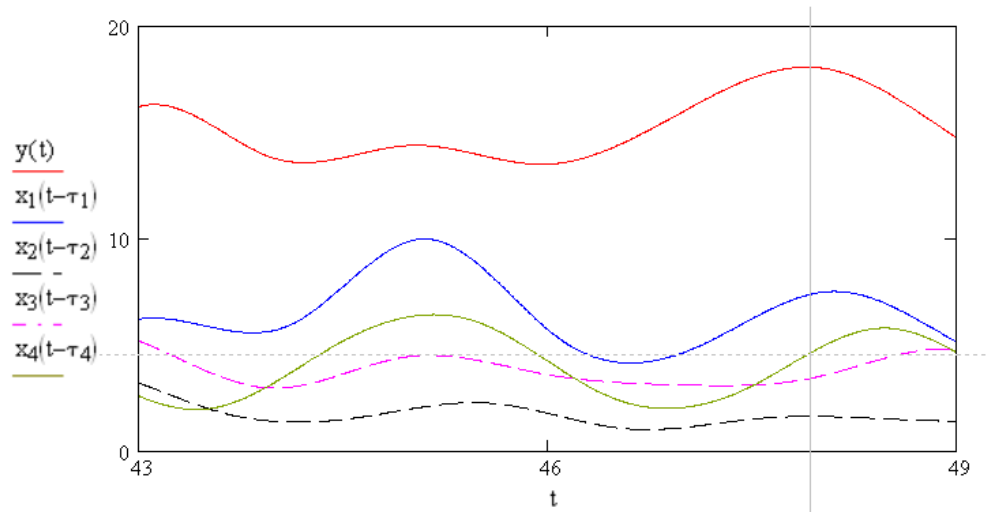


Рисунок 1. Графики динамики зависимого показателя $y(t)$ и контролируемых факторов влияния с учетом времен запаздывания τ_i (разработано автором)

С учетом найденных значений времен запаздывания были рассчитаны параметры соответствующей модели вида (2). Проведенный регрессионный анализ показал, что все коэффициенты уравнения регрессии b_i значимы на уровне $\alpha = 0,05$, величины их стандартных ошибок минимум на порядок меньше значений самих параметров, скорректированный коэффициент детерминации имеет значение 0,71, что свидетельствует о достаточно высокой точности модели регрессии.

Предложенная методика прошла апробацию для решения задач анализа и прогнозирования социально-экономических процессов в регионе, результаты которой свидетельствуют о достаточной точности определения времени запаздывания влияния контролируемых факторов на значение зависимой переменной, что позволяет получать более достоверные оценки степени влияния факторов, а также более точные прогнозные оценки развития социально-экономических показателей.

ЛИТЕРАТУРА

1. Кремер Н. Теория вероятности и математическая статистика [Текст]: учеб. пособие для вузов. – 2-е изд., – М.: Наука, 2006. – 573 с.
2. Айвазян С.А. Теория вероятностей и прикладная статистика [Текст]: учебник / С.А. Айвазян, В.С. Мхитарян – М.: ЮНИТИ-ДАНА, 2001. – 656 с.
3. Тараканов О.В., Новиков Е.И. Модель краткосрочного прогнозирования этноконфессиональной напряженности в субъектах РФ [Текст] // «Информационные системы и технологии» ИЗВЕСТИЯ ОрелГТУ № 4/54(567) июль-август 2009, с. 60-69.
4. Андерсон Т. Статистический анализ временных рядов [Текст] / пер. с англ. – М.: «Мир», 1976. – 757 с.
5. Джонстон Дж. Эконометрические методы [Текст] / пер. с англ. – М.: Статистика, 1980. – 444 с.
6. Ханк Д.Э., Уичерн Д.У., Райтс А. Дж. Бизнес-прогнозирование, 7-е издание [Текст] / пер. с англ. – М.: Издательский дом «Вильямс», 2003. – 656 с.
7. Сокольчик Павел Юрьевич, Малимон М.В. Оценочное определение времени запаздывания подачи исходного сырья при производстве сухих тампонажных смесей [Электронный ресурс] // Вестник ПНИПУ. Химическая технология и биотехнология. 2013. №1. – Режим доступа: <http://cyberleninka.ru/article/n/otsenochnoe-opredelenie-vremeni-zapazdyvaniya-podachi-ishodnogo-syrya-pri-proizvodstve-suhih-tamponazhnyh-smesey>, свободный. – Загл. с экрана. – Яз. рус., англ.
8. Энгельгардт Владислав Викторович Генетический алгоритм структурно-параметрической идентификации линейных динамических систем с помехами на входе и выходе [Электронный ресурс] // Известия ВУЗов. Поволжский регион. Технические науки. 2013. №4 (28). – Режим доступа: <http://cyberleninka.ru/article/n/geneticheskiy-algoritm-strukturno-parametricheskoy-identifikatsii-lineynyh-dinamicheskikh-sistem-s-romehami-na-vhode-i-vyhode>, свободный. – Загл. с экрана. – Яз. рус., англ.
9. Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы [Текст] / Под ред. В.М. Курейчика. – 2-е изд., испр. и доп. – М.: ФИЗМАТЛИТ, 2006. – 320 с.
10. Емельянов, В.В. Теория и практика эволюционного моделирования [Текст] / В.В. Емельянов, В.В. Курейчик, В.М. Курейчик. – М.: Физматлит, 2003. – 432 с.
11. Льюис К.Д. Методы прогнозирования экономических показателей [Текст] / пер. с англ. и предисл. Е.З. Демиденко. – М.: Финансы и статистика, 1986. – 133 с.

Novikov Evgeniy Ivanovich

The Academy of Federal security guard service of the Russian Federation, Russia, Orel
E-mail: nei05@rambler.ru

Parameter estimation of the linear regression models taking into account delay of influence of factors on an effect variable

Abstract. In article the problem of parametrical identification of the multifactorial linear regression models describing dependences between the socio-economic indexes presented by time series is considered. It is specified the main reason for complexity of identification of similar models – existence of a delay time of influence of controlled factors on a dependent indicator. Relevance of a problem of estimation of a delay time is caused by lack of the uniform formalized approaches to its decision now. The generalized structure of the regression model considering a delay time of influence of controlled factors on an effect variable is presented. For use of methods of a parameter estimation of the linear models it is offered to define a delay time as the independent parameter of model of regression. Various ways of the solution of this task are analysed. The offered approach to estimation of a delay time based on use of genetic algorithm is described. Are noted revealed during the conducted researches of feature of application of algorithm, influencing accuracy of the delay time estimates received on its basis. In addition to internal criteria of quality of selection of unknown values of a delay time it is offered to apply the external criteria based on comparison of observed and model values of an effect variable and characterizing quality of the specification of regression model. The choice of concrete criterion depends on applied research objectives. The author offered a technique of a parameter estimation of regression models taking into account a delay time of factors of influence which application allows to increase accuracy of regression models of the analysis and prediction of socio-economic indexes.

Keywords: statistical data; time series; multifactorial linear regression model; model parameters; lag variable; delay time; genetic algorithm; interpolation; statistical significance; accuracy of model

REFERENCES

1. Kremer N. Teoriya veroyatnosti i matematicheskaya statistika [Tekst]: ucheb. posobie dlya vuzov. – 2-e izd., – M.: Nauka, 2006. – 573 s.
2. Ayvazyan S.A. Teoriya veroyatnostey i prikladnaya statistika [Tekst]: uchebnik / S.A. Ayvazyan, V.S. Mkhitaryan – M.: YuNITI-DANA, 2001. – 656 s.
3. Tarakanov O.V., Novikov E.I. Model' kratkosrochnogo prognozirovaniya etnokonfessional'noy napryazhennosti v sub"ektakh RF [Tekst] // «Informatsionnye sistemy i tekhnologii» IZVESTIYA OrelGTU № 4/54(567) iyul'-avgust 2009, s. 60-69.
4. Anderson T. Statisticheskii analiz vremennykh ryadov [Tekst] / per. s angl. – M.: «Mir», 1976. – 757 s.
5. Dzhonston Dzh. Ekonometricheskie metody [Tekst] / per. s angl. – M.: Statistika, 1980. – 444 s.
6. Khank D.E., Uichern D.U., Rayts A. Dzh. Biznes-prognozirovaniye, 7-e izdanie [Tekst] / per. s angl. – M.: Izdatel'skiy dom «Vil'yams», 2003. – 656 s.
7. Sokol'chik Pavel Yur'evich, Malimon M.V. Otsenochnoe opredelenie vremeni zapazdyvaniya podachi iskhodnogo syr'ya pri proizvodstve sukhikh tamponazhnykh smesey [Elektronnyy resurs] // Vestnik PNIPU. Khimicheskaya tekhnologiya i biotekhnologiya. 2013. №1. – Rezhim dostupa: <http://cyberleninka.ru/article/n/otsenochnoe-opredelenie-vremeni-zapazdyvaniya-podachi-ishodnogo-syrya-pri-proizvodstve-suhih-tamponazhnyh-smesey>, svobodnyy. – Zagl. s ekrana. – Yaz. rus., angl.
8. Engel'gardt Vladislav Viktorovich Geneticheskiiy algoritm strukturno-parametricheskoy identifikatsii lineynykh dinamicheskikh sistem s pomekhami na vkhode i vykhode [Elektronnyy resurs] // Izvestiya VUZov. Povolzhskiy region. Tekhnicheskie nauki. 2013. №4 (28). – Rezhim dostupa: <http://cyberleninka.ru/article/n/geneticheskiiy-algoritm-strukturno-parametricheskoy-identifikatsii-lineynykh-dinamicheskikh-sistem-s-pomehami-na-vhode-i-vyhode>, svobodnyy. – Zagl. s ekrana. – Yaz. rus., angl.
9. Gladkov L.A., Kureychik V.V., Kureychik V.M. Geneticheskie algoritmy [Tekst] / Pod red. V.M. Kureychika. – 2-e izd., ispr. i dop. – M.: FIZMATLIT, 2006. – 320 s.
10. Emel'yanov, V.V. Teoriya i praktika evolyutsionnogo modelirovaniya [Tekst] / V.V. Emel'yanov, V.V. Kureychik, V.M. Kureychik. – M.: Fizmatlit, 2003. – 432 s.
11. L'yuis K.D. Metody prognozirovaniya ekonomicheskikh pokazateley [Tekst] / per. s angl. i predisl. E.Z. Demidenko. – M.: Finansy i statistika, 1986. – 133 s.