

В.А.ЧИРУХИН, канд. техн. наук, доцент, *tchir12@yandex.ru*
Национальный минерально-сырьевой университет «Горный», Санкт-Петербург

V.A.CHIRUHIN, PhD in eng. sc., associate professor *tchir12@yandex.ru*
National Mineral Resources University (Mining University), Saint Petersburg

ПРИМЕНЕНИЕ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ АППРОКСИМАЦИИ ПЕРИОДИЧЕСКИХ ПРОЦЕССОВ ПРИ ПОСТРОЕНИИ ПРОГНОЗОВ

Проанализированы методы обработки статистических данных о периодическом процессе методом наименьших квадратов с использованием функций Бесселя и рядов Фурье. Предложен несложный метод аппроксимации, в котором выбор значения периода колебания осуществляется, исходя из минимума оценки погрешности.

Ключевые слова: метод наименьших квадратов, аппроксимация, сезонная вариация, оценка погрешности.

THE APPLICATION OF THE LEAST SQUARES METHOD FOR APPROXIMATION OF PERIODIC PROCESSES AT FORECASTS CONSTRUCTION

In the article the methods of processing of statistical data about periodic process by the least squares method with usage of Bessel functions and Fourier numbers are analyzed. The simple method for approximation with a choice of the fluctuation period, proceeding from a minimum of the estimated error is offered.

Key words: least squares method, approximation, season variation, estimated error.

Любое ответственное решение в экономике требует серьезного обоснования, поскольку ошибка в принятии решения грозит большими финансовыми потерями, а иногда и более серьезными издержками социально-го и политического характера. Экспериментальный подход при этом чаще всего невозможен, поскольку постановка эксперимента в экономике весьма дорогостоящее удовольствие как в финансовом, так и в политическом и в социальном планах, ибо, как правило, эксперименты сопровождаются многократными пробами и ошибками. В таких случаях прибегают к моделированию поведения экономической системы.

В подавляющем количестве случаев используются математические модели, которые избавляют исследователя от необхо-

димости проведения дорогостоящих экспериментов. Все возможные варианты можно опробовать на модели, которую, условно говоря, можно резать и перекраивать неоднократно без всяких дополнительных вложений, если, конечно, модель является достаточно полной. Другим достоинством моделирования является формализация задачи, что позволяет воспользоваться для анализа математическим аппаратом, который не зависит от конкретной природы объекта. С помощью математических методов можно провести детальный количественный анализ экономической модели, предсказать, как поведет себя объект в различных условиях, и дать рекомендации для выбора наилучших вариантов решения проблемы. Построение формальных моделей для решения задач

экономики, их анализ и вывод практических рекомендаций – одна из важнейших задач прикладной математики. Современное программное обеспечение персональных компьютеров дает доступный и несложный путь для расчетов с использованием довольно сложных математических моделей экономических процессов.

Среди математических моделей заметное место занимают методы прогнозирования. При построении планов на будущие периоды невозможно обойтись без построения прогноза предполагаемых экономических результатов рассматриваемых процессов. Методы прогнозирования принято делить на качественные и количественные. В свою очередь, среди количественных методов различают каузальные и методы анализа числовых последовательностей. Качественные методы требуют участия экспертов, высказывающих свои мнения по предлагаемым проблемам, и специфической обработки результатов этих мнений. Немаловажно и то, что для построения прогноза в этом случае требуется довольно длительное время. Каузальные методы, хотя и позволяют учитывать большое количество параметров, требуют значительных вычислительных усилий и высокопроизводительной вычислительной техники.

Наиболее доступными в повседневной работе являются методы анализа числовых последовательностей. Эти методы могут быть полезны при решении широкого круга экономических, в частности логистических задач. Наиболее универсальным методом анализа числовых последовательностей является метод наименьших квадратов часто именуемый в литературе методом проецирования тренда. Этот метод позволяет аппроксимировать как линейные, так и степенные зависимости; он может быть использован при построении прогнозов, когда некоторая величина периодически изменяется (например, сезонное изменение объемов продаж каких-либо товаров). Известны аддитивный и мультипликативный методы прогнозирования, учитывающие сезонные вариации [2]. Напомним, что под сезонной

вариацией понимается повторение наблюдаемых данных через небольшой промежуток времени. Под термином сезон подразумевают день, неделю, месяц и квартал. Если повторение результатов наблюдений происходит за более длительный промежуток времени, то говорят уже о циклической вариации. В аддитивной модели сезонность выражается через определенное количество, добавляемое или вычитаемое из трендового значения с целью учета сезонных факторов, в мультипликативной модели – как процент от трендового значения. В обоих этих методах в качестве исходных данных фигурируют либо некоторые интегральные, либо некоторые средние за сезон показатели. Если наблюдения производились в течение всего периода изменения некоторой величины, например объема продаж, через равные промежутки времени (значения аргумента равноотстоящие), то хотелось бы использовать по возможности все результаты наблюдений без дополнительной обработки.

В работе [3] приводится формула для аппроксимации периодического процесса по методу наименьших квадратов в интервале $(0; 2\pi)$ тригонометрическим многочленом

$$y = \lambda_0 + \sum_{k=1}^m (\lambda_k \cos kx + \mu_k \sin kx),$$

где при равноотстоящих значениях аргумента x_i оценки коэффициентов λ_k и μ_k определяются формулами Бесселя:

$$\begin{aligned}\tilde{\lambda}_0 &= \frac{1}{n+1} \sum_{i=1}^n y_i; \\ \tilde{\lambda}_k &= \frac{2}{n+1} \sum_{i=1}^n y_i \cos kx_i; \\ \tilde{\mu}_k &= \frac{2}{n+1} \sum_{i=1}^n y_i \sin kx_i \\ (k &= 1, 2, \dots, m).\end{aligned}$$

Объем вычислений по этим формулам весьма велик.

Согласно методу аппроксимации периодического процесса с помощью ряда Фурье[1],

$$y = a_0 + \sum_{k=1}^m (a_k \cos kt + b_k \sin kt),$$

где a_0, a_k, b_k – коэффициенты

$$a_0 = \frac{1}{n} \sum_{k=1}^n y_k; \quad a_k = \frac{2}{n} \sum_{k=1}^n y_k \cos kt;$$

$$b_k = \frac{2}{n} \sum_{k=1}^n y_k \sin kt;$$

Считается, что хорошее приближение можно получить с использованием одной или двух гармоник. Исследуя сезонность, часто принимают n равным числу месяцев в году, т.е. 12. При этом первая гармоника примет вид

$$y = a_0 + a_1 \cos + b_1 \sin t.$$

Здесь

$$a_0 = \frac{1}{12} \sum_{k=1}^{12} y_k; \quad a_k = \frac{1}{6} \sum_{k=1}^{12} y_k \cos kt;$$

$$b_k = \frac{1}{6} \sum_{k=1}^{12} y_k \sin kt.$$

Следует отметить, что тренд процесса при такой аппроксимации остается постоянным и равным a_0 .

В литературе описан также способ аппроксимации периодического процесса функцией

$$y = a + b \cos \frac{2\pi}{T} t.$$

В этом случае тренд процесса тоже остается постоянным, что упрощает вычисления коэффициентов a и b , но снижает точность аппроксимации, если тренд на самом деле растущий или убывающий.

Существует также так называемый периодограмм-анализ. Этот метод был впервые предложен А. Шустером. Он развивает применение рядов Фурье при анализе статистических данных. Основу метода составляет нахождение периода колебательного процесса, т. е. $T = 2\pi / \alpha$ в ряду Фурье,

$$f(x) = A_0 + A_1 \sin(\alpha t + \beta) +$$

$$+ A_2 \sin(2\alpha t + \beta) + A_3 \sin(3\alpha t + \beta) + \dots$$

Для определения T в качестве первого приближения к эмпирическому берут два первых члена ряда Фурье:

$$f(x) = A_0 + A_1 \sin(\alpha t + \beta)$$

и подставляют в него значения t . Для каждого из них вычисляют полуамплитуду A_1 и начальную фазу β_1 . После этого строится периодографик, где на оси абсцисс отмечают испытываемые периоды, а на оси ординат – A_1 или интенсивность соответствующих этим периодам колебаний. Чем выше интенсивность, тем выше вероятность того, что соответствующий ей период колебания действительно существует [1].

Процедура вычислений, как видно из описания, довольно трудоемкая, и не учитывает возможного изменения тренда.

В настоящей работе предлагается аппроксимировать периодический процесс функцией

$$y = at + b + c \sin \frac{2\pi}{T} t$$

или

$$y = at + b + c \cos \frac{2\pi}{T} t.$$

В этом случае часть функции $at + b$ дает тренд процесса, растущий или убывающий, а синусоида моделирует периодический процесс, развивающийся вокруг тренда. Коэффициент c в формулах задает амплитуду синусоиды, а величина T – период процесса. Возьмем для примера первую функцию

$$y = at + b + c \sin \frac{2\pi}{T} t.$$

Для выбранной функции можно записать нормальные уравнения для определения коэффициентов по методу наименьших квадратов. Это нетрудно сделать обычным способом, составив вспомогательную функцию

$$\varphi(a, b, c, T) = \sum_{i=1}^n \left(at_i + b + c \sin \frac{2\pi}{T} t_i - y_i \right)^2$$

и, приравняв нулю ее частные производные по a, b, c и T . Так получим систему нормальных уравнений, решение которых определит значения a, b, c и T .

$$\begin{cases} a \sum_{i=1}^n t_i + nb + c \sum_{i=1}^n \alpha = \sum y_i; \\ a \sum_{i=1}^n t_i^2 + b \sum_{i=1}^n t_i + c \sum_{i=1}^n t_i \alpha = \sum y_i t_i; \\ a \sum_{i=1}^n t_i \alpha + b \sum_{i=1}^n \alpha + c \sum_{i=1}^n \alpha^2 = \sum_{i=1}^n y_i \alpha; \\ a \sum_{i=1}^n t_i^2 \beta + b \sum_{i=1}^n t_i \beta + c \sum_{i=1}^n t_i \alpha \beta = \sum_{i=1}^n y_i t_i \beta, \end{cases}$$

где $\alpha = \sin \frac{2\pi}{T} t_i$, $\beta = \cos \frac{2\pi}{T} t_i$.

Решение этой системы можно существенно упростить, исключив последнее уравнение. По имеющимся данным можно построить график процесса и определить период T . Такой подход известен. Однако при определении периода T по графику возможна (если не неизбежна) ошибка. Эта ошибка может быть исправлена позднее путем подбора периода по минимуму оценки погрешности прогноза.

Предположим, что величина T нам известна. Тогда система примет вид

$$\begin{cases} a \sum_{i=1}^n t_i + nb + c \sum_{i=1}^n \alpha = \sum y_i; \\ a \sum_{i=1}^n t_i^2 + b \sum_{i=1}^n t_i + c \sum_{i=1}^n t_i \alpha = \sum y_i t_i; \\ a \sum_{i=1}^n t_i \alpha + b \sum_{i=1}^n \alpha + c \sum_{i=1}^n \alpha^2 = \sum y_i \alpha. \end{cases}$$

Система легко решается по теореме Крамера, т. е. вычислением определителей. Можно решить систему в общем виде, а можно определить суммы, входящие в систему, и решить ее, подставив численные значения сумм. Для примера попробуем вычислить параметры периодического процесса. Возьмем, например, функцию

$$y = 0,2t + 20 + 20 \sin \frac{2\pi}{10} t.$$

Здесь $a = 0,2$; $b = 20$ и $c = 20$. Рассчитаем ее значения на промежутке значений t от 1 до 22 (можно было бы взять любой другой промежуток t) и попробуем по числовому ряду восстановить значения коэффициентов функции. Ряд до третьего знака после запятой представлен в табл.1.

Таблица 1

Значения функции y при заданных значениях t

t	y	t	y	t	y
1	31,956	9	10,044	17	4,379
2	39,421	10	22,000	18	4,579
3	39,621	11	33,956	19	12,044
4	32,556	12	41,421	20	24,000
5	21,000	13	41,621	21	35,956
6	9,444	14	34,556	22	43,421
7	2,379	15	23,000		
8	2,579	16	11,444		

Требуется рассчитать суммы, входящие в три нормальных уравнения, и, решив систему уравнений, найти значения параметров аппроксимирующей функции.

Для начала воспользуемся тем, что известно значение периода: 10. При работе с числовым рядом, описывающим реальный процесс, как уже говорилось, надо было бы построить график и подсчитать период этого процесса по графику. Расчеты удобно вести в Microsoft Excel, они дают точные значения: $a = 0,2$; $b = 20$ и $c = 20$. С высокой точностью получается исходный ряд. Оценка средней относительной погрешности при этом очень мала: $7,8 \cdot 10^{-16}$ (практически ноль). Таким образом, используя в качестве наблюдаемых величин точные значения выбранной функции, мы получили её точные параметры.

Применив этот алгоритм вычислений для $T = 9$ и $T = 11$, несложно получить значения a , b , и c для этих случаев. Значения коэффициентов, конечно, другие, но главное то, что оценку погрешности получим соответственно 45 и 40 %, т.е. значительно больше, чем в первом случае. Это дает возможность уточнения периода T по минимуму погрешности. Вариант периода T , при котором оценка погрешности будет наименьшей, будет указывать правильное значение этого параметра.

Теперь внесем отклонения в числовой ряд в пределах ± 5 %, моделирующие влияние случайных воздействий на результаты наблюдений (табл.2). То обстоятельство, что внесена заранее известная погрешность, позволяет сравнить оценку погрешности по результатам вычислений с внесенной величиной. Пользуясь все той же схемой расче-

тов в Microsoft Excel, при $T = 10$ получим $a = 0,184 \pm 0,009$; $b = 20 \pm 1$ и $c = 20 \pm 1$. Погрешность 5 %. Рассчитанные параметры хорошо согласуются с исходными данными в пределах внесенных искажений.

Таблица 2

Значения функции y с внесенными искажениями при заданных значениях t

t	y	t	y	t	y
1	34	9	11	17	4,5
2	37	10	23	18	5
3	38	11	35	19	13
4	35	12	40	20	23
5	20	13	43	21	35
6	10	14	33	22	44
7	2	15	22		
8	3	16	12		

Для $T = 9$ найдем $a = -0,16 \pm 0,07$; $b = 24 \pm 10$ и $c = 13 \pm 6$, для $T = 11$ – $a = -0,5 \pm 0,2$; $b = 18 \pm 7$ и $c = 15 \pm 6$. Соответственно погрешности 43 и 40 %. Оценки погрешностей для истинного и ошибочных вариантов периода разнятся в 8 раз. Такого отличия вполне достаточно, чтобы правильно определить величину периода T .

Можно смоделировать разброс наблюдений с помощью генератора случайных чисел. Например, можно выбрать подряд из полученного ряда случайных чисел те числа, которые не превышают 0,5. Разделив их на 10 для того, чтобы внесенная таким спо-

собом погрешность не превышала 5 %, нетрудно из первоначального ряда (см. табл. 1) получить ряд с погрешностью (табл. 3). В табл. 3 значения $y_{\text{пог.}}$ получены из исходного ряда попеременно путем сложения со случайными числами, деленными на 10, и вычитанием из них. Это и будет ряд с внесенными искажениями, моделирующими погрешность наблюдений.

При $T = 10$ расчетные значения $a = 0,2 \pm 0,0006$, $b = 20 \pm 0,06$ и $c = 20 \pm 0,06$. Оценка погрешности 3 %. Для $T = 9$ $a = -0,15 \pm 0,07$, $b = 24 \pm 11$ и $c = 13 \pm 6$; для $T = 11$ $a = 0,5 \pm 0,2$, $b = 18 \pm 7$ и $c = 15 \pm 6$. Соответственно погрешности 43 и 40 %. Для вариантов расчетов, в которые искажения вносились в первоначальный ряд с точностью ± 5 % или в пределах 5 % с помощью генератора случайных чисел, вычисленные значения параметров и оценки погрешностей практически совпали. Это вполне ожидаемый результат, так как выборка оказалась достаточно велика.

В качестве примера рассмотрим аппроксимацию функции изменения цен в пенсах на пшеницу в Англии с 1898 по 1917 г., т.е. за 20 лет. Данные взяты из работ Кондратьева (табл. 4).

Расчеты показывают, что для $T = 3$ года оценка погрешности 25,1 %, для $T = 4$ года 25,3 %, для $T = 5$ лет 24,9 %, для $T = 6$ лет 24,4 %, для $T = 7$ лет 25,8 %. Наименьшая

Таблица 3

Значения функции y с внесенными искажениями с помощью генератора случайных чисел при заданных значениях t

Параметр	t										
	1	2	3	4	5	6	7	8	9	10	11
Случайное число	0,208	0,377	0,289	0,284	0,340	0,456	0,449	0,197	0,358	0,061	0,110
Случайное число, деленное на 10	0,021	0,038	0,029	0,028	0,034	0,046	0,045	0,020	0,036	0,006	0,011
y	31,956	39,421	39,621	32,556	21,000	9,444	2,379	2,579	10,044	22,000	33,956
$y_{\text{пог.}}$	31,977	39,345	39,65	32,528	21,034	9,398	2,424	2,559	10,08	21,994	33,967

Параметр	t										
	12	13	14	15	16	17	18	19	20	21	22
Случайное число	0,074	0,082	0,256	0,498	0,324	0,268	0,066	0,364	0,378	0,043	0,441
Случайное число, деленное на 10	0,007	0,008	0,026	0,050	0,032	0,027	0,007	0,036	0,038	0,004	0,044
y	41,421	41,621	34,556	23,000	11,444	4,379	4,579	12,044	24,000	35,956	43,421
$y_{\text{пог.}}$	41,414	41,629	34,53	23,05	11,412	4,406	4,572	12,08	23,962	35,96	43,377

Таблица 4

Изменения цены на пшеницу по годам

t	y	t	y	t	y
1898	408	1905	356	1912	442
1899	308	1906	339	1913	391
1900	323	1907	367	1914	416
1901	321	1908	384	1915	632
1902	337	1909	443	1916	699
1903	321	1910	380	1917	936
1904	340	1911	390		

погрешность при $T = 6$ лет. Надо отметить, что последние три точки ряда, соответствующие годам войны, должны изменить тренд. Если их отбросить по названным соображениям, то погрешность для $T = 6$ и $T = 7$ снижается до 6 %.

ЛИТЕРАТУРА

1. Венецкий И.Г. Основные математико-статистические понятия и формулы в экономическом анализе: Справочник / И.Г.Венецкий, В.И.Венецкая. М., 1979. 101 с.
2. Просветов Г.И. Управление запасами: задачи и решения. М., 2008. 192 с.
3. Сборник задач по теории вероятностей, математической статистике и теории случайных функций / Под ред. А.А.Свешникова. М., 1965. 622 с.

REFERENCES

1. Venetskij I.G., Venetskaja V.I. Main mathematics-statistical concepts and formulas of the economic analysis: The directory. Moscow, 1979. 101 p.
2. Prosvetov G.I. Storekeeping's Gleams: problems and decisions. Moscow, 2008. 192 p.
3. The Collection of problems in probability theory, mathematical statistics and the theory casual functions / Ed. A.A.Sveshnikov. Moscow, 1965. 622 p.