УДК 519.237.5

С.М. ЛАПАЧ[*], В.М. КУЗЬМІН[**]

ПОБУДОВА БАГАТОЛАНКОВИХ ПОЛІГОНАЛЬНИХ РІВНЯНЬ РЕГРЕСІЇ

[*]Національний технічний університет України «КПІ імені Ігоря Сікорського», Київ, Україна
[**]Національний авіаційний університет, Київ, Україна

Анотація. *Розроблена формалізована процедура визначення абсцис точок перемикання багатоланкової регресії. Базується на попередній побудові найбільш інформативної поліноміальної регресії. З неї аналітичним шляхом визначаються приблизні координати шуканих точок. Уточнення відбувається при пошуку найкращого за описовими характеристиками регресійного полігонального рівняння. Визначені точки дають можливість смислового аналізу процесу і обґрунтованого прогнозу.*
Ключові слова: *полігональна регресія, поліноміальна регресія, інформативність, сплайн-регресія, обчислювальна стійкість, число обумовленості.*

Аннотация. *Разработана формализованная процедура определения абсцисс точек переключения полигональной регрессии. Основывается на предварительном построении наиболее информативной модели полиномиальной регрессии. Из нее аналитически определяются приблизительные координаты искомых точек. Уточнение производится в процессе поиска наилучшего по описательным свойствам регрессионного полигонального уравнения. Определенные таким образом точки дают возможность смыслового анализа процесса и обоснованного прогноза.*
Ключевые слова: *полигональная регрессия, полиномиальная регрессия, информативность, сплайн-регрессия, вычислительная стойкость, число обусловленности.*

Abstract. *A formalized procedure of definition of polygonal absciss of switch points of multi-unit regression was developed. It is based on the preliminary building of the most informative polynomial regression. From the resulting model analytically determined the approximate location of the desired points. Refinements coordinates is in the search for the best descriptive characteristics of polygonal regression equation. Determined points enable semantic analysis process and a reasonable forecast.*
Keywords: *Polygonal regression, polynomial regression, informative, spline regression, computational stability, condition number.*

1. Вступ

При побудові полігональної регресії нерозв`язаною залишається проблема формалізованої і обґрунтованої процедури вибору кількості ланок і абсцис точок перелому при описі процесів, які вимагають смислового аналізу або прогнозу.

Дослідження неперіодичних процесів часто зустрічається з такими випадками, в яких декілька раз відбувається зміна характеру процесу. Опис таких процесів за допомогою поліномів вимагає високої степені полінома, що робить його (а) непридатним для прогнозу; (б) замість тенденцій відстежуються випадкові флуктуації; (в) при використанні звичайних поліномів матриця стає погано обумовленою. Найкраще такі процеси описувати за допомогою полігональної регресії (сплайн-регресії) [1, 2]. Рівняння полігональної регресії має такий загальний вигляд: $\hat{y} = b_0 + \sum_{i=1}^{k} b_i \frac{(x-a_i)+|x-a_i|}{2}$, де a_i – координати точки перелому, причому $a_1 = 0$. Як правило, точки перелому сплайн-регресії в таких процесах є точками зміни ходу процесу, і їх розміщення має значення з точки зору смислового аналізу процесу [3]. Крім того, це має велике значення в задачах прогнозування процесів для вибору фрагмента, за яким буде виконуватись прогноз [4]. Визначенню точок перемикання присвячена велика кількість літератури, зокрема [5–7], в якій використовуються різноманітні математичні методи: генетичні алгоритми, лінійне програмування, кластеризація тощо. При цьому відбувається класична підміна мети: замість мети по визначенню точок зміни

ISSN 1028-9763. Математичні машини і системи, 2017, № 2

тенденції шукають найкращі з точки зору деякого формального критерію, які не мають відношення до вихідної постановки задачі в конкретній галузі дослідження.

Метою роботи є створення формалізованої і обґрунтованої процедури вибору абсцис точок переключення сплайн-регресії (полігональної регресії) при описі процесів, які вимагають смислового аналізу.

2. Ідея і алгоритм побудови

Алгоритм побудови формалізованої обґрунтованої процедури базується на двох положеннях:

1) нулі першої та другої похідних полінома відповідають точкам зміни тенденції процесу: зміною тенденції процесу може бути як екстремум, так і перелом функції;

2) найкращим поліномом для опису процесу є рівняння регресії з максимальною інформативністю (в ньому мінімальна частка регресорів, які описують випадкові відхилення, а не тенденції).

Побудову полігональної багатоланкової лінії регресії пропонується виконувати за таким алгоритмом.

1. На першому етапі виконується визначення гіпотетичних точок зміни тенденції.

1.1. Виконується побудова найбільш інформативного полінома степені k для опису набору даних.

1.2. Визначення нулів першої і другої похідних побудованого найбільш інформативного полінома степені k.

1.3. Відбір точок зміни тенденції з результатів 1.2 і аналізу спостережень.

2. На другому етапі відбувається уточнення розміщення точок зміни тенденції.

2.1. Побудова найкращого рівняння полігональної регресії, уточнюючи гіпотетичні точки зміни тенденції.

2.2. Побудова варіантів рівняння полігональної регресії, виключаючи точки, де зміна тенденції сумнівна (якщо це необхідно).

3. Аналіз отриманих моделей і вибір моделі для використання.

П.п. 1.1 і 1.2 першого етапу можуть виконуватись повністю автоматично, без участі людини. В п. 1.1 для забезпечення стійкості розв’язку необхідно використовувати ортогональні поліноми Чебишева, оскільки матриця, сформована із звичайних поліномів, при збільшенні степені швидко стає сильно закорельованою або навіть виродженою. Максимальна інформативність визначається як максимальне значення розрахункового критерію Фішера для значимості множинного коефіцієнта кореляції. Необхідність п. 1.3 викликана потребою зменшити кількість варіантів у п. 2.1 і 3. Але цей пункт може бути і пропущений.

На другому етапі п. 2.1 теж може бути виконаний автоматично. Виконання ж п. 2.2 потребує участі спеціаліста.

Необхідність п. 3 викликана можливістю отримання кількох моделей, які рівнозначні з точки зору статистичних показників, що їх описують, але різні з точки зору смислового аналізу.

3. Приклад розв’язання задачі

Вихідні дані

Детальне пояснення алгоритму виконаємо на реальних даних, взятих з [8], які приведені в табл. 1.

Таблиця 1. Вихідні дані (імпорт нафти в США)

№ пп	Рік	Імпорт нафти (тис. барелів за добу)	№ пп	Рік	Імпорт нафти (тис. барелів за добу)
1	1973	6556,145	17	1989	8060,545
2	1974	6112,184	18	1990	8017,521
3	1975	6955,712	19	1991	7626,748
4	1976	7312,598	20	1992	7887,697
5	1977	8807,249	21	1993	8620,422
6	1978	8363,411	22	1994	8996,222
7	1979	8356,129	23	1995	8834,94
8	1980	6909,025	24	1996	9478,492
9	1981	5995,673	25	1997	10161,56
10	1982	5113,311	26	1998	10708,07
11	1983	5051,353	27	1999	10852,26
12	1984	5436,982	28	2000	11459,25
13	1985	5067,144	29	2001	11871,34
14	1986	6223,512	30	2002	11530,24
15	1987	6677,696	31	2003	12264,39
16	1988	7402,021	32	2004	13145,09

Побудова поліному степені k для опису набору даних

Визначити ланки і приблизні точки перелому можна було б і візуально, як зазвичай і робиться. Але при цьому або залишається сумнів (у складних ситуаціях), або необхідно перебирати декілька варіантів ланок. Крім того, такий вибір не є обґрунтованим, а спирається тільки на думку людини. Зауважимо, що у складних ситуаціях різні експерти пропонують різні варіанти розбиття, виходячи з одних і тих же даних.

У зв'язку з цим пропонується початкову кількість ланок і попередню оцінку розміщення точок перелому отримувати за допомогою апроксимації вибірки поліномом високої степені. Проблемою тут є визначення оптимальної степені поліному, так як його степінь можливо підвищувати до $N-1$, де N – розмір вибірки. При цьому частина статистичних характеристик відповідного рівняння регресії буде асимтотично монотонно зростати. Це приводить до того, що з деякого моменту рівняння починає відслідковувати не тенденції, а випадкові флуктуації.

Розв'язання проблеми можливе за допомогою розрахункового значення критерію Фішера для множинного коефіцієнта кореляції: оптимальна степінь відповідає його максимальному значенню. На відміну від більшості статистичних характеристик розрахункове значення критерію Фішера для множинного коефіцієнта кореляції (F_R) не зростає монотонно, воно має максимум при максимальній інформативності моделі і після цього монотонно спадає при збільшенні степені полінома.

У табл. 2 приведено тенденції зміни показників для прикладу, який розглядається. Ми бачимо, що зі зростанням степені полінома регресійного рівняння множинний коефіцієнт кореляції асимтотично зростає, а середньоквадратична помилка зменшується. Тобто, формально якість апроксимації з ростом степені покращується. Але розрахункове значення критерію Фішера для R (F_R) зменшується. Це пов'язано з тим, що найкраща апроксимація набору рівнянь і найкраще рівняння регресії – це зовсім не одно і теж. Найкраща апроксимація буде відслідковувати, крім закономірностей, і випадкові відхилення. Якщо ж враховувати чутливість методу найменших квадратів до «викидів», а саме «перетягування» рівняння до аномальних спостережень, то найкраща апроксимація може бути незадовільною з точки зору опису поведінки досліджуваного процесу.

Таблиця 2. Деякі статистичні характеристики поліноміальних моделей

Статистичні показники регресії	Степінь полінома			
	3	4	5	6
Множинний коефіцієнт кореляції (R)	0,918773	0,935402	0,958123	0,960134
Розрахункове значення критерію Фішера для R (F_R)	50,55107	47,23966	58,21476	49,15506
Середньоквадратична помилка (σ)	940,2844	857,6131	707,7772	704,6103

Як видно з таблиці, найбільшу інформативність має рівняння поліному п'ятої степені, яке має такий вигляд:

$$\hat{y} = 3923{,}831 + 2334{,}999x - 457{,}514x^2 + 33{,}95877x^3 - 1{,}05523x^4 + 0{,}011827x^5.$$

На рис. 1 приведено графік, побудований за вказаною вище моделлю, з відображенням відповідних даних спостережень. Моделі отримані за допомогою стандартних засобів надбудови «Аналіз даних» Excel. Побудова моделей стандартними поширеними програмними засобами має два недоліки:

1) необхідність самому будувати кожну матрицю гаданого поліному як вихідні дані відповідної моделі;

2) сильна закорельованість матриць і, відповідно, відсутність обчислювальної стійкості коефіцієнтів регресії при високій степені полінома (вище третьої).

Використовуючи програмний засіб (ПЗ) ПРІАМ (планування, регресія і аналіз моделі) [9], немає необхідності в побудові кількох поліноміальних моделей з наступним вибором найкращої: цей програмний засіб забезпечує такі дії автоматично, включаючи побудову матриць поліномів Чебишева до вихідних факторів.

Результати роботи (графік побудованої моделі і результати спостережень) приведені на рис. 2.

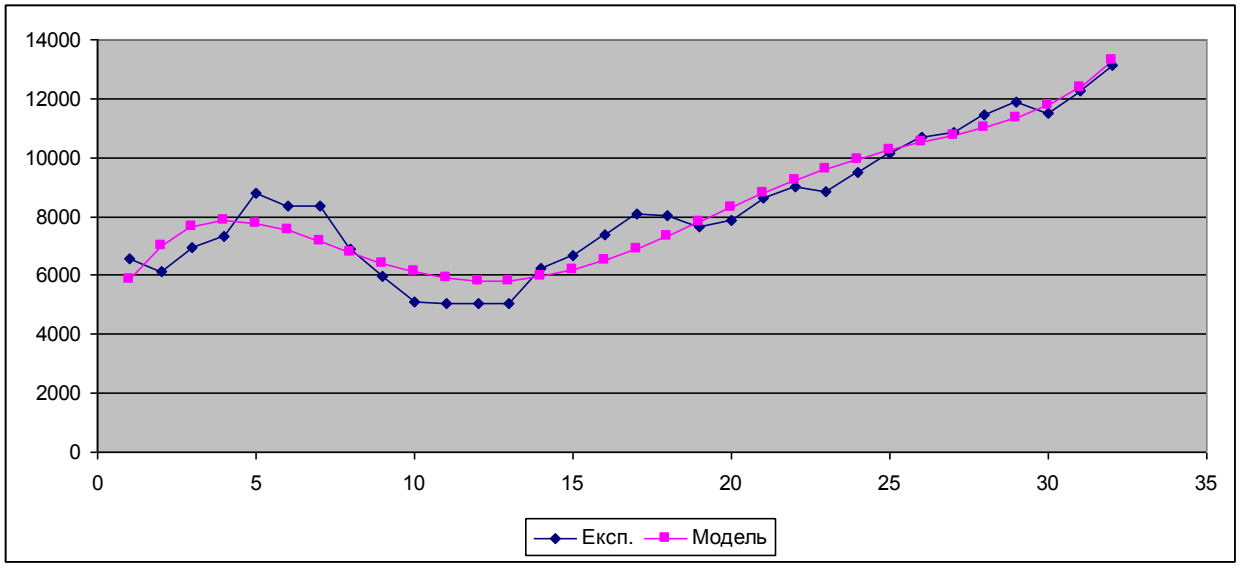

Рис. 1. Вихідні дані і опис їх поліномом 5-ї степені

Хоча отримані статистичні характеристики моделі (табл. 3) незначно відрізняються від приведених вище в табл. 2, перевагою ПРІАМ є не тільки автоматизація побудови найкращої моделі, але й забезпечення її структурної і обчислювальної стійкості. Число обумовленості моделі, побудованої ПРІАМ, дорівнює 1 (теоретично ідеальне значення), а в мо-

делі п’ятої степені, побудованою Excel, – Cond=$2{,}7\times10^{15}$, тобто матриця погано обумовлена. Стійкість забезпечена побудовою моделі в поліномах Чебишева.

Модель, побудована ПЗ ПРІАМ, має такий вигляд:

$$\hat{y} = 8295{,}47 + 2957{,}67\, f^{(1)} + 2113{,}56\, f^{(2)} + 956{,}914\, f^{(5)} - 857{,}751 f^{(4)},$$

де

$$f^{(1)} = 0{,}0645161(X - 16{,}5);\ f^{(2)} = 1{,}55((f^{(1)})^2 - 0{,}354839),$$
$$f^{(4)} = 5{,}3504((f^{(1)})^4 - 0{,}909469(f^{(1)})^2 + 0{,}0963714,$$
$$f^{(5)} = 11{,}0575((f^{(1)})^5 - 1{,}17586(f^{(1)})^3 + 0{,}266295.$$

Більш висока інформативність цієї моделі, порівняно з моделлю, побудованою Excel, пояснюється меншою кількістю членів.

Таким чином, застосування ПЗ ПРІАМ дозволяє автоматично виконувати п. 1.1 описаного вище алгоритму.

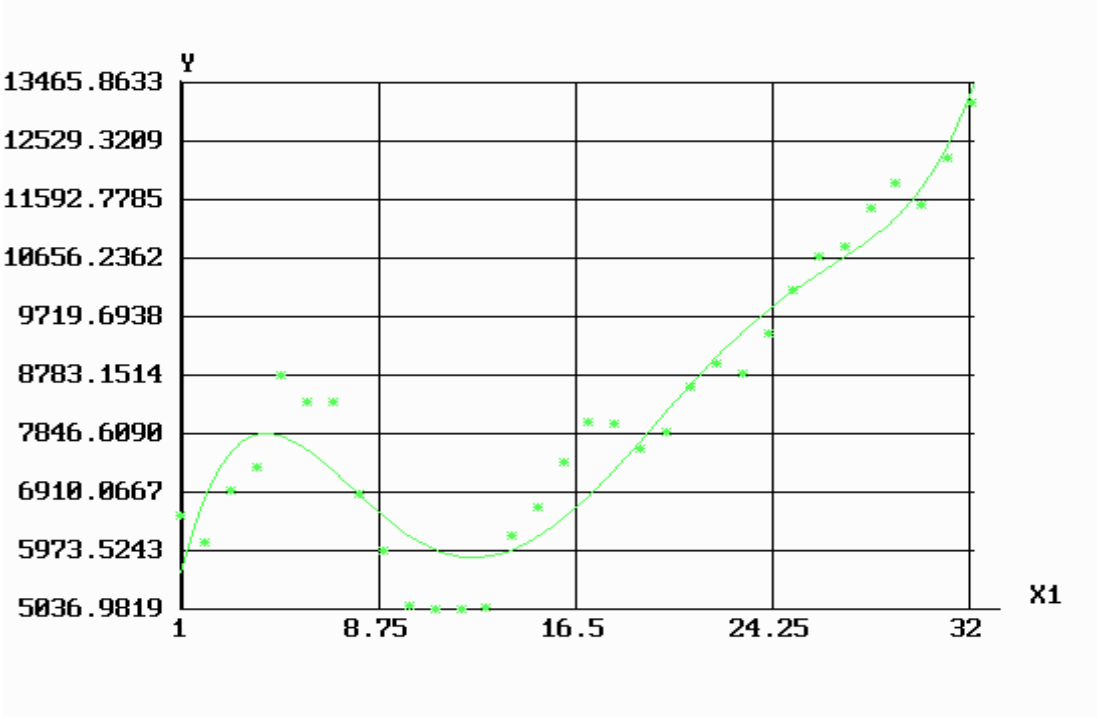

Рис. 2. Графік рівняння регресії, побудованого ПРІАМ

Таблиця 3. Статистичні показники рівняння регресії

Назва	Значення
Множинний коефіцієнт кореляції (R)	0,95727
Розрахункове значення критерію Фішера для R (F_R)	73,9578
Середньоквадратична помилка рівняння регресії (σ)	701,438
Число обумовленості (cond)	1

Визначення гіпотетичних точок зміни тенденції

Точки зміни тенденції відповідають або точкам екстремуму, або точкам перелому функції. Для визначення цих точок необхідно визначити нулі першої і другої похідних. На рис. 3 і 4 приведені графіки рівнянь першої і другої похідних відповідно, а в табл. 4 корені цих рівнянь, визначені чисельними методами.

Побудова рівнянь похідних для поліноміальної функції легко може бути автомати-

зована, як і знаходження коренів відповідних рівнянь чисельними методами.

Таблиця 4. Корені рівнянь першої та другої похідних

Номер точки	Перша похідна		Друга похідна	
	Відгук	Фактор	Відгук	Фактор
1	-0,00036	4,179958	-1,4E-05	7,488339
2	-0,00038	12,36942	-7,3E-06	19,35513
3			-4,6E-06	26,6902

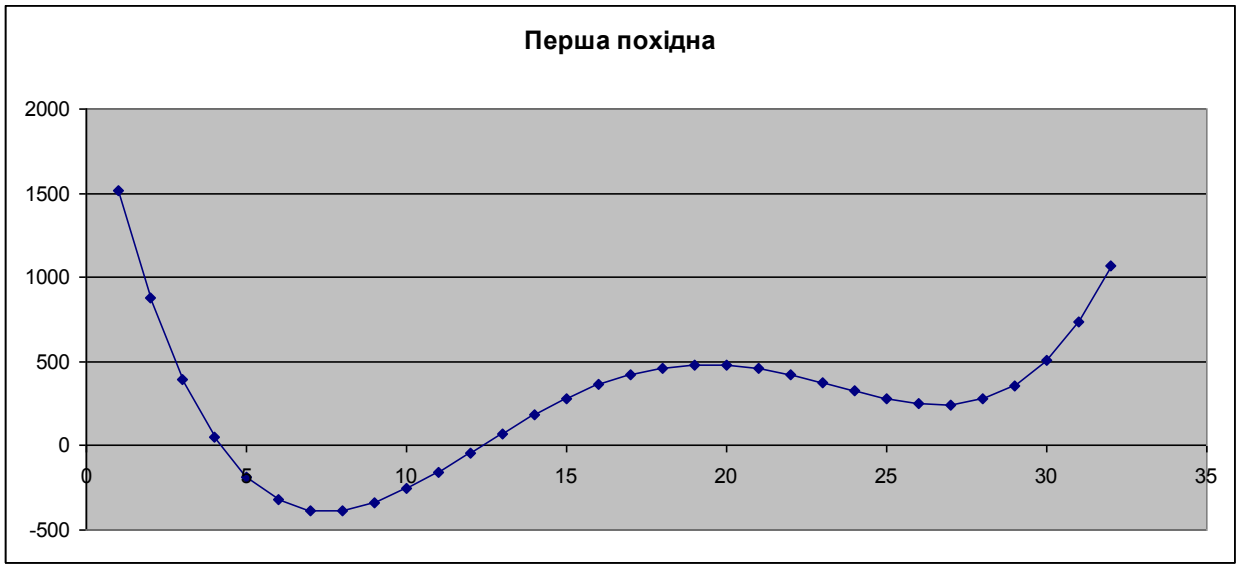

Рис. 3. Графік першої похідної функції опису

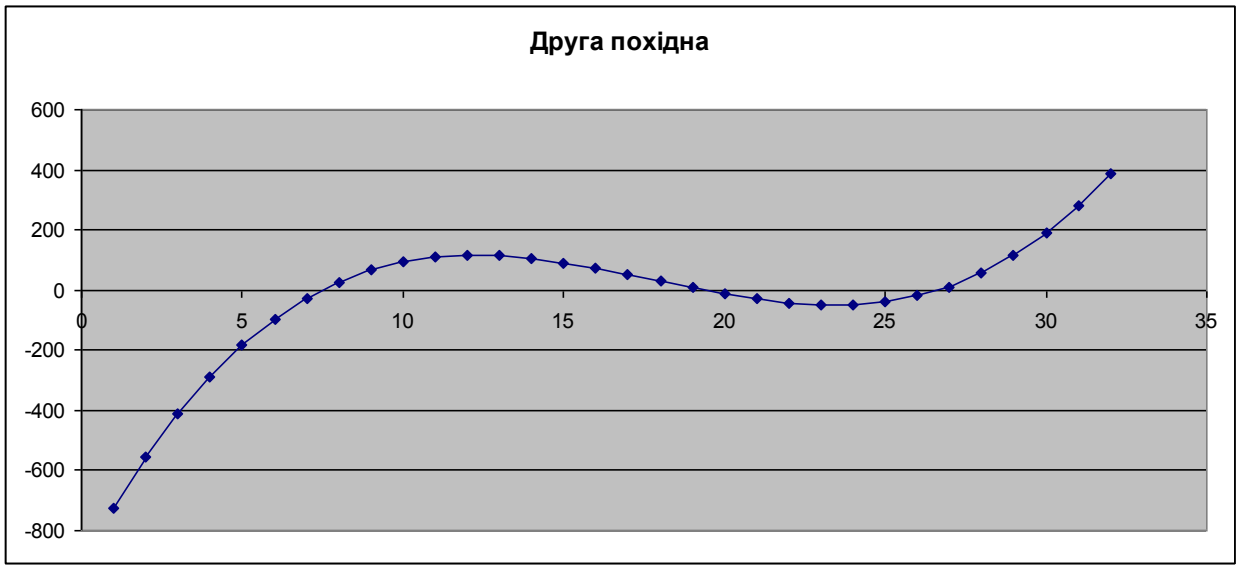

Рис. 4. Графік другої похідної функції опису

Відбір точок зміни тенденції з аналізу спостережень

Виходячи з аналізу графіка вихідних даних до розгляду, як гіпотетичні точки зміни тенденції приймаються обидва корені нулів першої похідної і останній корінь другої. Рішення опирається на думку експертів, які вважають, що перші два кореня другої похідної відображають випадкові викиди, а не тенденції процесу. Зауважимо ще раз, що цей пункт може бути пропущено і для розгляду прийняті всі точки з наступним відхиленням точок випадкових відхилень на етапі 2.2.

Уточнення розміщень точок зміни тенденції. Побудова найкращого полігонального рівняння регресії

Уточнення необхідне у зв'язку з тим, що регресійна модель у представленій постановці задачі не апроксимує похідні, а тільки саму функцію і не може використовуватись як формула для оберненого визначення незалежної змінної [10]. У зв'язку з цим при побудові багатоланкового полінома відбувається підбір найкращих з точки зору мінімальної залишкової дисперсії координат точок зміни тенденції, базуючись на визначених на попередньому кроці приблизних координатах. Підбір виконується перебором варіантів за допомогою спеціально розробленого макросу для електронної таблиці Excel.
Отримана модель

$$Y=5160{,}6531 + 708{,}93113X - 1327{,}971\,(X-5)_+ + 984{,}30302\,(X-11)_+ + 28{,}449927\,(X-26)_+,$$

де $(X - X_П)_+ = \dfrac{(X - X_П) + |X - X_П|}{2}$.

Статистичні характеристики моделі приведені в табл. 5.

Таблиця 5. Характеристики полігональної моделі

Множинний коефіцієнт кореляції, R	0,9824639		
Частка, пояснювана моделлю, R^2	0,9652352		
Розрахункове F-відношення для R	187,41216		
Критичне значення для F_R	2,7277653	V_1=4	V_2=27
Залишкова дисперсія	204516,75		
Число обумовленості	7760,189		
		Точки перелому	
Номер	1	2	3
Координата	5	11	26
Коефіцієнти регресії			
Номер	Значення		
0	5160,6531		
1	708,93113		
2	-1327,971		
3	984,30302		
4	28,449927		

Уточнення розміщень точок зміни тенденції. Побудова варіантів рівняння полігональної регресії, виключаючи точки, де зміна тенденції сумнівна

Візуальний аналіз графіка викликає гіпотезу, що остання точка перелому є випадковою, а не відображає зміну тенденції. Для її перевірки побудуємо полігональну регресію без останнього члена. Статистичні характеристики такої моделі приведені в табл. 6. Видно, що хоча коефіцієнт множинної кореляції практично не змінився (відмінність у четвертому знаку), розрахункове F-відношення для R збільшилось майже в півтора рази. Це дозволяє нам вважати останній «перелом» реакцією на випадкові зміни, а не зміною тенденції. Графік цього рівняння приведено на рис. 5.

Таблиця 6. Статистичні характеристики вибраної моделі

Множинний коефіцієнт кореляції, R	0,9823765		
Частка, пояснювана моделлю, R^2	0,9650637		
Розрахункове F-відношення для R	257,81918		

Продовж. табл. 6

Критичне значення для F_R	2,9466853	V_1=3	V_2=28
Залишкова дисперсія	198185,99		
Число обумовленості	7864,4879		
	Точки перелому		
Номер	1	2	
Координата	5	11	
Номер	Значення		
0	5154,7104		
1	711,90252		
2	-1337,31		
3	996,20232		

Таким чином, отримана модель

$$Y=5154{,}7104 + 711{,}90252X - 1337{,}31\,(X - 5)_+ + 996{,}20232\,(X-11)_+,$$

де $$(X - X_П)_+ = \frac{(X - X_П) + |X - X_П|}{2}.$$

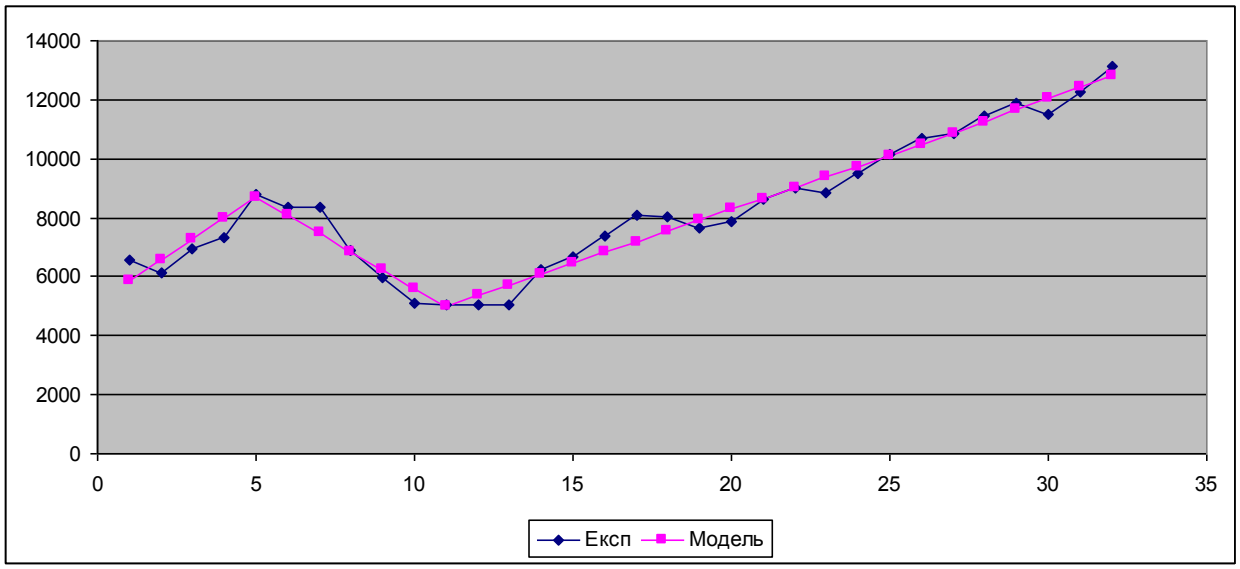

Рис. 5. Графік прийнятого рівняння регресії і дані спостереження

Метод дозволяє знаходити абсциси координат точок зміни тенденції з більш високою точністю, наприклад, до місяця, тижня чи дня.

3. Висновки

Пропонується розроблена процедура побудови сплайн-регресії з формалізованим визначенням точок перелому за рахунок попередньої побудови найбільш інформативного поліноміального рівняння регресії високої степені. Визначення особливих точок цього рівняння дозволяє отримати перше наближення точок зміни тенденції досліджуваного процесу. На другому етапі координати точок уточнюються для побудови багатоланкового полігонального рівняння регресії з найкращими описовими властивостями. Хоча така багатокрокова процедура вимагає значних обчислювальних затрат, вона дозволяє обґрунтовано визначити точки перелому, які у смисловому значенні є точками зміни тенденції. Використовуючи спеціальні програмні засоби (ПРІАМ і розроблений макрос для побудови багатолан-

кових рівнянь регресії), дослідник має змогу автоматизувати формальні обчислення і йому залишається тільки виконувати смисловий аналіз на кожному етапі роботи.

Перевагами запропонованої процедури є високий рівень автоматизації за рахунок наявності спеціальних програмних засобів і обґрунтованість прийняття рішень.

У майбутньому можливе об'єднання програмних засобів в один з подальшим використанням як засобу підтримки прийняття рішень при аналізі тенденцій чи прогнозування подібного виду процесів.

СПИСОК ЛІТЕРАТУРИ

1. Бородич С.А. Эконометрика / Бородич С.А. – Мн.: Новое знание, 2001. – 408 с.
2. Грін В.Г. Економетричний аналіз / Грін В.Г.; пер. з англ. – К.: Основи, 2005. – 1197 с.
3. Кузьмін В.М. Використання полігональної регресії в економічних дослідженнях / В.М. Кузьмін, С.М. Лапач // Економіка і управління. – 2004. – № 3. – С. 79 – 84.
4. Лапач С.Н. Прогнозирование с использованием полигональной регрессии / С.Н. Лапач, А.В. Чубенко, П.Н. Бабич // Провизор. – 2003. – № 16. – С. 11 – 13.
5. Казаченок В.В. Построение сплайновой регрессии по экспериментальным данным / В.В. Казаченок // Вестник Белорусского государственного университета. – (Серия 1 «Физика, математика, информатика»). – 1997. – № 1. – С. 70 – 71.
6. Остропицкий В.М. Методы поиска узлов склеивания сплайн-регрессий / В.М. Остропицкий, А.Ф. Приставка // Вопросы прикладной математики и математического моделирования: сб. науч. тр. – Д.: ДГУ, 1997. – С. 121 – 125.
7. Алгоритмы и программы восстановления зависимостей / В.Н. Вапник, Т.Г. Глазкова, В.А. Кощеев [и др.]; под ред. В.П. Вапника. – М.: Наука, ГРФМЛ, 1984. – 816 с.
8. Douglas M.C. Applied Statistics and Probability for Engineers [Fours Edition] / M.C. Douglas Montgomery, G.C. Runger. – NJ.: John Wiley & Sons, Inc., 2007. – 768 p.
9. Лапач С.Н. Планирование, регрессия и анализ моделей PRIAM (ПРИАМ) / С.Н. Лапач, С.Г. Радченко, П.Н. Бабич // Каталог программные продукты Украины. – К., 1993. – С. 24 – 27.
10. Дрейпер Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. – [3-е изд.]. – М.: Издательский дом «Вильямс», 2007. – 912 с.

Стаття надійшла до редакції 07.11.2016