УДК 004.942+519.816 DOI: 10.20535/1810-0546.2016.1.64895

М.З. Згуровський, С.В. Трухан, П.І. Бідюк
Національний технічний університет України "КПІ", Київ, Україна

МЕТОДИКА ЗАСТОСУВАННЯ ТЕОРІЇ ЕКСТРЕМАЛЬНИХ ЗНАЧЕНЬ ДЛЯ АНАЛІЗУ ДАНИХ

Background. To solve the problems of modeling and forecasting on the basis of large datasets (including singular ones) in conditions of uncertainty it is necessary to develop integrated information decision support systems (DSS). A methodology is proposed for application of extreme value theory for statistical models development and DSS on their basis.
Objective. The goal of the study is in application of the extreme value theory for analysis and estimation of model parameters on the basis of random samples. It is necessary to develop an effective methodology for analysis of pseudorandom sequences and estimation of unknown model parameters; to present examples of analysis using extreme value theory and software developed.
Methods. To solve the problems stated the following approaches were used: pseudorandom sequences generating procedures; probabilistic distributions of the extreme value theory, and methods for estimating unknown model parameters. A multistep methodology is proposed for extreme values processing and DSS is developed for analysis and modeling of pseudorandom sequences.
Results. Using the DSS developed and generated statistical data as well as proposed methodology the procedure was developed for extreme values analysis. The procedure is to be used for estimating of forecasting models for the process of various origin. A comparative analysis of parameter characteristics for GEV-distributions was performed.
Conclusions. Using the instrumentation developed it was shown that the proposed methodology for processing extreme values is convenient for analysis of singular datasets. This is substantiated with the high quality approximation of theoretical probability density by empirical curve. A comparison of model parameters estimation results showed that the estimates converge faster when parameters of form and scale are defined more exactly.
Keywords: extreme value theory; maximum likelihood estimator; simulation and modeling; decision support system.

Вступ

Потужним інструментом дослідження складних систем і процесів, управління якими пов'язане з прийняттям рішень в умовах невизначеності, є імітаційне моделювання. Порівняно з іншими методами імітаційне моделювання дає можливість розглядати велику кількість альтернатив, покращуючи при цьому якість оцінок прогнозів та управлінських рішень. Метою імітаційного моделювання є побудова імітаційної моделі об'єкта і проведення імітаційного експерименту з її використанням для вивчення законів і закономірностей функціонування досліджуваних об'єктів (наприклад, законів розподілу випадкових величин), поведінки об'єктів із врахуванням заданих обмежень, цільових функцій в умовах імітації та взаємодії із зовнішнім середовищем.

Проте використання цього методу при розв'язанні задач практичного спрямування залишається недостатньо поширеним унаслідок складності відповідного математичного апарату, необхідності обробки значних масивів даних та високих обчислювальних витрат. Досить часто трапляються випадки, коли відсутні повні (необхідні за об'ємом) масиви даних, і саме тому виникає необхідність у генеруванні модельних даних із використанням математичних методів для наближеного відтворення цілісної картини відповідного процесу

Тому, в зв'язку з необхідністю розв'язання нових задач моделювання і прогнозування на основі великих обсягів (у тому числі вироджених) вхідних даних, які не можна розв'язати з використанням існуючих методів, виникає потреба у розробці нових інтегрованих інформаційних систем для підтримки прийняття рішень в умовах невизначеності, а також комбінування існуючих методів і підходів до обробки таких даних. Одним із використовуваних підходів є теорія екстремальних значень (ТЕЗ). Вона широко застосовується при розв'язанні таких задач, як регулювання структури портфелю активів у страхуванні, аналіз виникнення ризикових ситуацій у сфері фінансів та кредитування, прогнозування трафіку в галузі телекомунікацій, аналіз екологічних та кліматичних процесів тощо [1–3].

Задачею теорії екстремальних значень є оцінювання ймовірності появи випадкових величин, пов'язаних з екстремальними, тобто рідкісними, подіями. Екстремальні значення не є фіксованими величинами, це нові випадкові величини, які залежать від типу вихідного розподілу та потужності вибірки. Наприклад, у сфері страхування будь-якого майна рідкісною, але ймовірною подією є настання страхових

випадків, які повинні супроводжуватись виплатами значних страхових премій або банкрутством страхових компаній.

Саме тому для розв'язання задачі аналізу псевдовипадкових даних у роботі пропонується ймовірнісна модель, яка будується із застосуванням теорії екстремальних значень. Одним із ключових моментів побудови адекватної моделі досліджуваного процесу є коректний вибір методу оцінювання параметрів математичних моделей за експериментальними даними. Для розв'язання задачі оцінювання невідомих параметрів моделі часто застосовують метод максимальної правдоподібності (ММП), модифікації методу найменших квадратів (МНК) та байєсівський підхід. Останній дає можливість точніше оцінювати моделі в умовах невизначеності, а саме коли статистичні дані мають різні типи розподілів ймовірностей, а також вибрати кращу модель із множини оцінених кандидатів за множиною статистичних критеріїв. Перевагою цього підходу є можливість його застосування до обробки статистичних вибірок відносно малих і великих розмірів, а також за наявності пропусків даних [4, 5]. Популярним і відносно універсальним методом оцінювання параметрів математичних моделей різних класів на сьогодні є метод Монте-Карло для марковських ланцюгів (МКМЛ), який застосовують для оцінювання параметрів лінійних і нелінійних моделей [6–8] в умовах обробки даних з різними типами розподілів.

Постановка задачі

Мета роботи полягає у застосуванні теорії екстремальних значень для аналізу та оцінювання параметрів статистичних моделей на основі даних стосовно розвитку випадкових процесів. Експериментальні дані модельного характеру (псевдовипадкові послідовності) генеруються за машинними алгоритмами.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

1) дослідити властивості розподілів екстремальних значень та методи оцінювання невідомих параметрів моделей екстремальних значень, зокрема оцінити можливості використання методу максимальної правдоподібності;

2) розробити і реалізувати програмно архітектуру системи підтримки прийняття рішень (СППР) для аналізу та моделювання випадкових процесів з екстремальними значеннями;

3) дослідити деякі алгоритми машинного генерування псевдовипадкових послідовностей;

4) розробити методику аналізу випадкових значень та оцінювання невідомих параметрів статистичних моделей, побудованих на основі таких вхідних даних;

5) навести приклади аналізу псевдовипадкових значень за допомогою теорії екстремальних значень та інструментарію розробленої бібліотеки мовою програмування R.

Архітектура систем підтримки прийняття рішень

СППР – інтерактивна комп'ютерна автоматизована система (програмний комплекс), призначена для надання допомоги та підтримки різних видів діяльності особи, що приймає рішення (ОПР) стосовно розв'язання слабко структурованих і неструктурованих задач [9]. СППР надає максимально можливу допомогу ОПР, якщо вона побудована на тих же принципах, які використовуються ОПР при прийнятті рішень у повсякденному житті.

У загальному випадку проектування СППР складається з 3-х етапів: 1 – декомпозиція про-

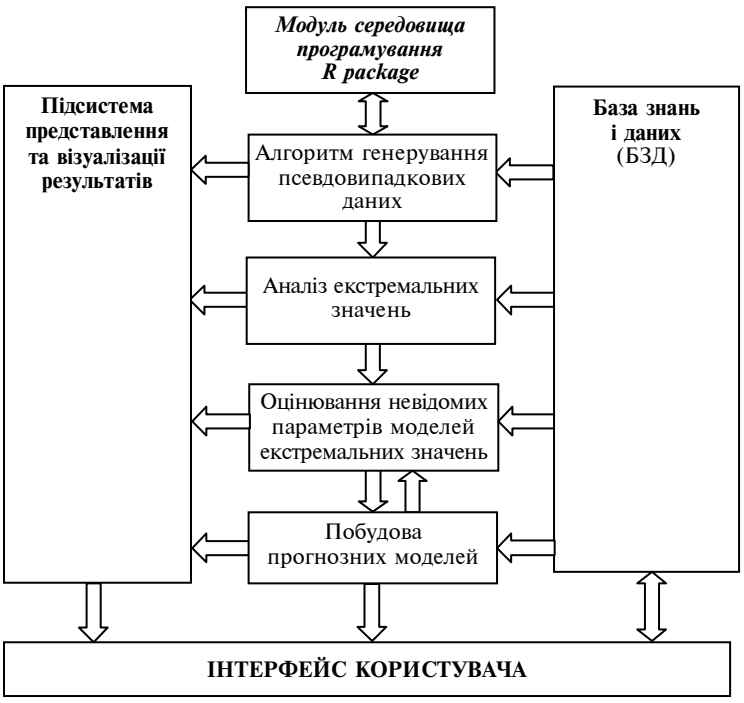

Рис. 1. Архітектура СППР для аналізу екстремальних даних

цесу ухвалення рішення на елементарні операції, докладний опис виконання цього процесу особою, яка приймає рішення; 2 – аналіз конкретної задачі стосовно ухвалення рішення, проектування СППР на функціональному рівні; 3 – докладна специфікація функцій системи, її реалізація, верифікація (тестування) і супроводження [10]. Архітектура СППР для аналізу та моделювання псевдовипадкових значень показана рис. 1.

Очевидно, що наведена архітектура може бути використана СППР багатьох типів, які призначені для розв'язування задач побудови математичних моделей за статистичними даними та їх застосування для оцінювання прогнозів і генерування альтернативних рішень. Конкретні застосування подібних систем визначаються множинами обчислювальних алгоритмів, що містяться в системі, критеріальною базою, можливими множинами правил для генерування альтернатив і вибору кращих результатів.

Генерування псевдовипадкових чисел

На практиці досить часто потрібна послідовність випадкових чисел з гауссовим розподілом. Найчастіше для їх генерування використовують центральну граничну теорему, згідно з якою розподіл суми N однаково розподілених випадкових величин підпорядковується нормальному закону розподілу при $N \to \infty$. Тому послідовність незалежних рівномірно розподілених чисел $\{x_n\}$ перетворюється на послідовність чисел із гауссовим розподілом $\{y_n\}$ згідно з таким правилом:

$$y_n = \frac{1}{N}\sum_{i=0}^{N-1} x(n\cdot N - i). \qquad (1)$$

При цьому параметр N повинен бути достатньо великим [11].

Різновидом цього методу є пропозиція Рейдера, яка полягає в тому, що послідовність із L незалежних рівномірно розподілених випадкових величин перетворюється за допомогою матриці Адамара на нову послідовність із L некорельованих випадкових величин із гауссовим розподілом. Кожна із L гауссових величин отримується додаванням (відніманням) L чисел з рівномірним розподілом. Тому при $L > 16$ спостерігається максимальне наближення до нормального розподілу. Такий підхід досить ефективний, оскільки з L рівномірно розподілених величин отримуємо L нормально розподілених випадкових величин, а не L/N, як при використанні формули (1)

Крім того, існує прямий метод перетворення пари рівномірно розподілених випадкових величин на пару псевдовипадкових величин. Якщо представити послідовність рівномірно розподілених на інтервалі (0, 1) випадкових величин через $\{x_n\}$ та визначити $\{y_n\}$ таким чином:

$$y_n = \sqrt{2\sigma^2 \ln\left[\frac{1}{x_n}\right]}, \qquad (2)$$

то $\{y_n\}$ буде мати розподіл Релея, тобто:

$$P_y(y_0) = \frac{y_0}{\sigma^2}\exp\left(\frac{-y_0^2}{2\sigma^2}\right). \qquad (3)$$

Щільність і функцію розподілу Релея зображено на рис. 2 і 3.

Якщо сформувати дві нові випадкові величини $\{w_n\}$ та $\{w_{n+1}\}$ згідно з формулами (4), (5):

$$w_n = y_n \cos[2\pi x_{n+1}], \qquad (4)$$

$$w_{n+1} = y_n \sin[2\pi x_{n+1}], \qquad (5)$$

то нові величини будуть мати нормальний розподіл з нульовим середнім та дисперсією, рівною σ^2. Також слід зауважити, що в результаті величини $\{w_n\}$ та $\{w_{n+1}\}$ не корелюють між собою, що еквівалентно незалежності для нормально розподілених випадкових величин [11].

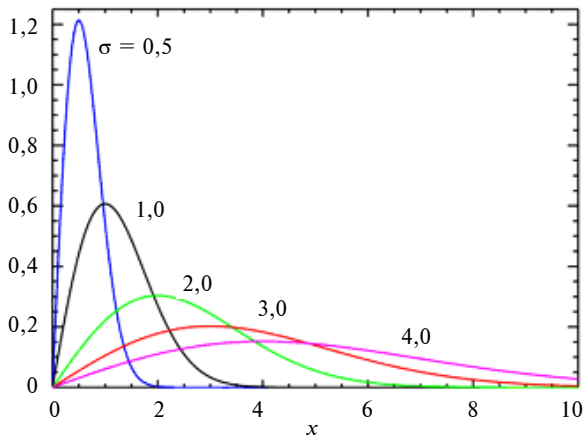

Рис. 2. Щільність розподілу Релея

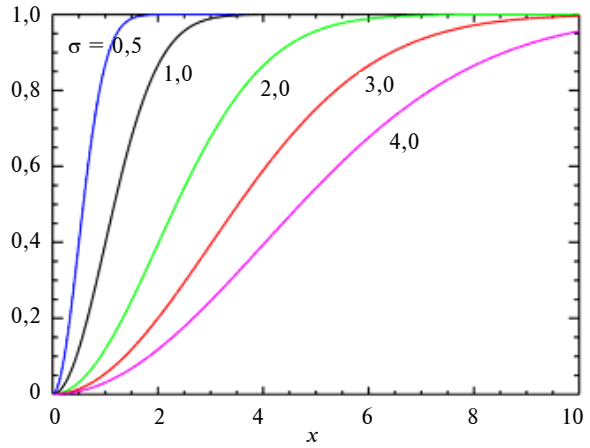

Рис. 3. Функція розподілу Релея

Отже, описаний вище метод дає на практиці корисні результати, але потребує додаткових обчислень логарифмів, синусів, косинусів переважно за рахунок великих часових витрат та достатньої процесорної потужності. Тому рекомендована потужність вибірки для оптимальних часових втрат не повинен перевищувати 300 точок [11].

Методика обробки екстремальних значень

Математичну модель екстремальних даних можна подати у вигляді [12]:

$$M_n = \max\{X_1,...X_n\}, \qquad (6)$$

де $X_1,...,X_n$ – послідовність незалежних випадкових величин з функцією розподілу F. У виразі (6) величина M_n позначає максимум досліджуваного процесу на інтервалі часу n і має розподіл [12]

$$\Pr\{M_n \le z\} = \Pr\{X_1 \le z,..., X_n \le z\} =$$

$$= \Pr\{X_1 \le z\} \times ... \times \Pr\{X_n \le z\} = \{F(z)\}^n. \qquad (7)$$

Функція F невідома, а тому розглядається наближена оцінка для F^n. Якщо послідовність констант $\{a_n > 0\}$ і $\{b_n > 0\}$ таких, що

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to F(a_n x + b_n x)^n \to G(z),$$

при $n \to \infty$, то G – невироджена функція розподілу, яка належить до одного з розподілів екстремальних значень, наприклад до узагальненого розподілу екстремальних значень (Generalized Extreme Value – GEV):

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad (8)$$

де μ – параметр розподілу; σ – параметр масштабованості; ξ – параметр форми розподілу.

Відповідно до теореми про типи розподілів екстремальних значень виділяють три типи таких розподілів:

1) розподіл Гумбела:

$$G(z) = \exp\left\{-\exp\left(-\left(\frac{z-b}{a}\right)\right)\right\}, \; -\infty < z < \infty;$$

2) розподіл Фреше:

$$G(z) = \begin{cases} 0, \; z \le b; \\ \exp\left(-\left(\dfrac{z-b}{a}\right)^{-\alpha}\right), \; z > b; \end{cases}$$

3) розподіл Вейбулла:

$$G(z) = \begin{cases} \exp\left(-\left(-\left(\dfrac{z-b}{a}\right)\right)^{\alpha}\right), \; z < b; \\ 1, \; z \ge b. \end{cases}$$

Для всіх трьох випадків $a > 0$, b – дійсне число. Для другої та третьої функцій $\alpha > 0$. Ці три класи розподілів називають розподілами екстремальних значень, вони зображені на рис. 4.

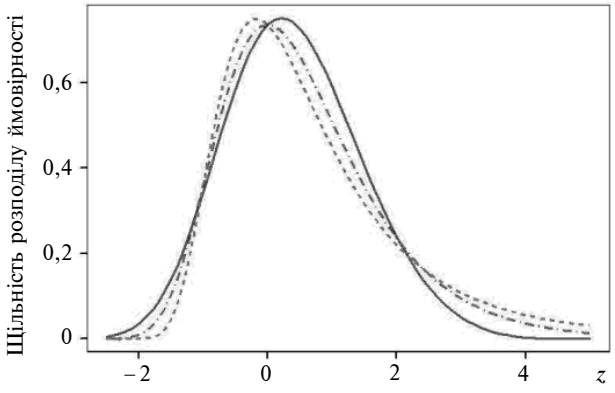

Рис. 4. Функції щільності розподілу для трьох типів розподілів: —— – Вейбула; – – – – Фреше; — · · – Гумбела

З рис. 4 видно, що кожен із розподілів має свою форму поведінки хвоста. Наприклад, для розподілу Вейбула хвіст має кінцеву точку $z_{\sup} = \dfrac{\mu - \sigma}{\xi}$, а для розподілів Фреше та Гумбе-

ла $z_{\sup} = \infty$. Крім того, щільність розподілу Гумбела експоненційно затухає, тоді як щільність розподілу Фреше затухає поліноміально. Розподіл Гумбела є наближеним до класу таких відомих розподілів: нормального, лог-нормального та гамма. Розподіл Фреше має тяжкий хвіст, для якого за означенням $E(X^r) = \infty$ при $r \geq \frac{1}{\xi}$ (що означає нескінченність дисперсії при $\xi \geq 1/2$). В окремий клас виділяють узагальнений розподіл Парето (Generalized Pareto Distribution – GPD), який отримуємо за умови: X – це розподіл, що умовно перевищує деякий поріг u:

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}, \qquad (9)$$

де $u \to w_F = \sup\{x : F(x) < 1\}$, що найчастіше зводиться до пошуку границі:

$$F_u(y) \approx G(y, \sigma_u, \xi),$$

де G – узагальнений розподіл Парето, еквівалентний виразу [2]

$$G(y, \sigma, \xi) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)_+^{-1/\xi}.$$

Якщо $\xi > 0$, то маємо довгий хвіст $x^{-1/\xi}$, що еквівалентно розподілу Парето; якщо $\xi = 0$, при $\xi \to 0$ отримаємо $G(y, \sigma, 0) = 1 - \exp\left(-\frac{y}{\sigma}\right)$, тобто експоненціальний розподіл із середнім σ; і якщо $\xi < 0$, то кінцева верхня точка знаходиться на рівні $-\sigma/\xi$. Також однією із переваг GEV-розподілів є інваріантність кожного з розподілів, які належать до цього класу.

Розглянемо методику обробки екстремальних значень. Для обробки статистичного ряду з n незалежних, однаково розподілених змінних $X_1, ..., X_n$ застосовується послідовність таких дій: групування вибірок даних з n спостережень (такі вибірки повинні містити від 50 до 100 значень); визначення максимуму Z_i для кожного блоку i; встановлення наближення кожного блоку максимумів до GEV-розподілу. Зазвичай за довжину блоку беруть величину першого року, але для зручності часто використовують дані річного максимуму Z_i для i-го року.

Після апроксимації GEV-розподілом для кожного з річних максимумів розраховується функція квантиля [3, 4]:

$$z_p = \begin{cases} \mu - (\sigma/\xi)(1 - (-\log(1-p))^{-\xi}), & \xi \neq 0; \\ \mu - \sigma \log(-\log(1-p)), & \xi = 0. \end{cases}$$

Припустимо, що $y_p = -\log(1-p)$, тоді квантиль-функція матиме вигляд

$$z_p = \begin{cases} \mu - (\sigma/\xi)(1 - (y_p)^{-\xi}), & \xi \neq 0; \\ \mu - \sigma \log(y_p), & \xi = 0. \end{cases}$$

Якщо зобразити z_p залежно від $\log(y_p)$, то графік буде мати лінійний характер при $\xi = 0$. Якщо $\xi < 0$, отримаємо випуклу криву з асимптотичною границею $(\mu - \sigma)/\xi$ при $p \to 0$, а при $\xi > 0$ отримаємо увігнутий графік без кінцевої границі. Такий графік називається графіком повернення рівня (return level plot), він вважається інструментом або способом представлення згладженої моделі [3]. Загалом методику обробки екстремальних значень можна подати у вигляді п'яти кроків.

1. Оцінювання параметрів моделі та пошук оптимальної довжини блоку.

Остання задача зводиться до пошуку співвідношення між величинами відхилення та дисперсії. Наприклад, коли довжина блоків незначна, то наближення розподілів до границь є поганим і призводить до відхилень у оцінюванні та екстраполяції. З іншого боку, великі блоки породжують значення з великими оцінками дисперсії.

Для оцінювання параметрів моделей часто використовується метод максимальної правдоподібності (ММП). Однак умова регулярності оцінювання не задовольняється при застосуванні ММП до GEV-розподілів, тому що кінцева точка розподілів залежить від значення параметра. Це означає, що стандартні асимптотичні результати аналізу за методом максимальної правдоподібності недоречно застосовувати до GEV-розподілів. Цю проблему дослідив Сміт у 1985 р. з такими результатами [3]: якщо $\xi > -0{,}5$, то оцінювання за ММП має стандартний асимптотичний характер; якщо $-1 < \xi < 0{,}5$, то оцінки ММП можуть бути отримані, але не із заданими асимптотичними властивостями; якщо $\xi < -1$, то оцінки ММП вважаються неправдоподібними.

Окремий випадок: якщо $\xi < -0{,}5$, то це еквівалентно розподілу з дуже коротким обмеженим верхнім хвостом, який є рідкісним явищем для теорії екстремальних значень. Логарифмічна

функція правдоподібності для GEV-розподілів, коли $\xi \neq 0$, має вигляд

$$l(\mu,\sigma,\xi) = -m\log\sigma -$$
$$-(1+1/\xi)\sum_{i=1}^{m}\log\left(1+\xi\frac{z_i-\mu}{\sigma}\right)-\sum_{i=1}^{m}\left(1+\xi\frac{z_i-\mu}{\sigma}\right)^{-1/\xi}$$

за умови, що $\left(1+\xi\frac{z_i-\mu}{\sigma}\right)>0$ для $i=1,...,m.$ Коли остання умова перестає виконуватись, функція правдоподібності дорівнює нулю і логарифмічна функція правдоподібності набуває значення нескінченності.

Для розподілу Гумбела при $\xi = 0$ логарифмічна функція правдоподібності має вигляд

$$l(\mu,\sigma) = -m\log\sigma -$$
$$-\sum_{i=1}^{m}\log\left(\frac{z_i-\mu}{\sigma}\right)-\sum_{i=1}^{m}\left(-\frac{z_i-\mu}{\sigma}\right). \qquad (10)$$

Після використання методів числової оптимізації та максимізації виразу (10) отримуємо оцінку параметрів за методом максимальної правдоподібності у вигляді $(\hat{\mu},\hat{\sigma},\hat{\xi})$ [3].

2. Графічна перевірка наближення GEV-моделей.

Для обґрунтування екстраполяції GEV-моделей можна скористатись способами графічного аналізу даних.

Графік щільності розподілу. В основі такого графіка лежить порівняння емпіричної та апроксимуючої функцій щільності розподілу. Абсциса точки на графіку щільності розподілів є емпіричною функцією розподілу, в яку замість аргументу підставляють дані з вибірки, а ордината – це теоретична функція розподілу, куди аналогічно замість аргументу підставляють дані зі статистичної вибірки. Функція емпіричного розподілу оцінюється в i-му впорядкованому блоці максимумів Z_i і має вигляд

$$\tilde{G}_i(Z_i) = i/(m+1).$$

Апроксимуюча функція щільності розподілу в тій самій точці має такий вигляд:

$$G(Z_i) = \exp\left\{-\left(1+\hat{\xi}\left(\frac{z_{(i)}-\hat{\mu}}{\hat{\sigma}}\right)\right)^{-1/\hat{\xi}}\right\}.$$

Для того щоб отримати найкраще наближення моделі, необхідно задовольнити рівність $\tilde{G}(Z_i) = \hat{G}(Z_i)$. За допомогою цього графіка на практиці часто вдається запобігти ефекту "виродженості". Тобто коли множина точок $\{\tilde{G}(Z_i), \hat{G}(Z_i)\}, i=1,...,m,$ лежить близько до діагоналі, обидві функції є обмеженими в околі одиниці та значення абсциси z збільшуються.

Графік квантилів (*Q-Q plot*). Недоліком класичної методології оцінювання фінансових ризиків VaR є припущення про нормальність розподілу та наявність симетрії в розподілі. На практиці більшість економічних процесів асиметричні, а фінансові ряди мають вироджений хвіст. Саме графік квантилів дає можливість оцінити ступінь довіри для низки параметричних моделей. Графік квантилів визначається як множина точок:

$$\left\{X_{k,n},\ F^{-1}\left(\frac{n-k+1}{n}\right),\ k=1,...,n\right\}.$$

Якщо параметрична модель надає прийнятне згладжування, то графік має лінійну форму. Тому графік дає можливість порівняти оцінені моделі та вибрати найкращу, а також оцінити, як вибрана модель апроксимує хвіст емпіричного розподілу. Тобто якщо ряд апроксимується нормальним розподілом і емпіричні дані мають вироджений хвіст, то графік квантилів буде характеризувати криву на вершині правого кінця або на дні лівого кінця розподілу. Крім розглянутих вище видів графічного аналізу, існують графік рівня процесу (return level plot) та середня функція ексцесу (mean excess function) [3, 4].

3. Визначення порога екстремального значення.

Для забезпечення ефективнішого результату наближення екстремальних даних до одного з GEV-розподілів застосовують так звані порогові моделі. Нехай множина статистичних даних перевищує деякий поріг u, а $X_1,...,X_n$ – послідовність незалежних однаково розподілених змінних з функцією розподілу F. Тоді умовна ймовірність визначається так:

$$F_u(y) = P(X \le u+y \,|\, X > u),$$

або

$$F_u(y) = \frac{F(u+y)-F(u)}{1-F(u)}.$$

Цей вираз дає можливість визначити ступінь наближення значень ймовірності для великих значень порога u.

Задача вибору оптимального порога ідентична задачі визначення розміру блоку. Обидві задачі спрямовані на визначення балансу між відхиленням та дисперсією. Низький рівень призводить до порушень асимптотичної апроксимації, а високий рівень забезпечує велику дисперсію. Метод вибору порога базується на основі середнього GPD-розподілу. Якщо Y – випадкова змінна у GPD-розподілі з параметрами σ і ξ, коли $\xi < 1$, то математичне сподівання $E(Y) = \sigma/(1-\xi)$. В інших випадках середнє є нескінченністю.

Якщо модель є істинною відносно порога u_0, то вона також істинна для всіх інших порогів u, більших за u_0. Тобто для забезпечення високого рівня адекватності побудованої моделі достатньо знайти одне значення порога, а всі інші припустити проміжними при оцінюванні невідомих параметрів моделі. Середнє для обох випадків визначається так [5]:

$$e(u_0) = E(X - u_0 / X > u_0) = \tilde{\sigma}_{u_0}/(1-\xi),$$

$$e(u) = E(X - u/X > u) = \tilde{\sigma}_u/(1-\xi) = \quad (6)$$

$$= (\tilde{\sigma}_{u_0} + \xi(u - u_0))/(1-\xi).$$

Оскільки $e(u) = E(X - u/X > u)$ – це лінійна функція від u, то, враховуючи вираз (6), оцінювання величини порогу можна виконати так [3, 13]:

1) побудувати графік кривої залишків, що відображають множину точок:

$$\left(u, \sum_{i=1}^{n_u} (x_i - u)/n_u \right), u < x_{\max},$$

де n_u – число значень, які перевищують u; $x_{\max}$ – верхня границя досліджуваного значення;

2) вибрати порогове значення, над яким графік має наближено лінійний характер стосовно u.

Також для визначення порога екстремального значення використовують метод умовно прийнятного вибору, який базується на такому правилі: поріг встановлюється в тій області, де хвіст становить 5–10 % від усієї вибірки. Головне припущення: він не повинен бути більшим ніж 10–15 %. На практиці 10 %-на границя використовувалась досить часто [12, 13].

4. Оцінювання невідомих параметрів моделі.

Після визначення порога потрібно оцінити невідомі параметри узагальненого розподілу Парето. Як відомо, при оцінюванні невідомих параметрів моделі поширеним є метод максимальної правдоподібності. Нехай $y_1, ..., y_k$ – це значення k залишків від порога; тоді логарифмічна функція правдоподібності при $\xi \neq 0$

$$l(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1}^{k} \log(1 + \xi y_i/\sigma),$$

коли $(1 + \xi y_i/\sigma) > 0$, а для будь-яких інших випадків $l(\sigma, \xi) = -\infty$. При $\xi = 0$ логарифмічна функція правдоподібності має вигляд:

$$l(\sigma) = -k(\log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i).$$

Другим поширеним методом оцінювання невідомих параметрів є байєсівський підхід. Перевагою байєсівського аналізу при застосуванні до моделей обробки екстремальних значень є його незалежність від регулярності припущень стосовно характеру початкового розподілу, як цього потребує метод максимальної правдоподібності. Практичне застосування байєсівського підходу для оцінювання невідомих параметрів проілюстровано на прикладі узагальнених лінійних моделей [13].

Обчислювальні експерименти і результати

Експериментальне дослідження ефективності методики аналізу екстремальних даних виконано за допомогою розробленої СППР для аналізу псевдовипадкових чисел та машинного генерування статистичних вибірок згідно з розглянутою вище процедурою та принципами проектування СППР. Потужність статистичної вибірки становила 250 вимірів, які включають такі вхідні змінні: параметр розподілу, параметр масштабованості, параметр форми розподілу, розмір статистичної вибірки. Розглянуто дві статистичні вибірки однакової потужності та з різними значеннями вхідних параметрів. Обґрунтований підбір вхідних параметрів забезпечується за допомогою раніше проведених досліджень з реальними статистичними даними [13].

Для реалізації алгоритму генерування псевдовипадкових чисел та виконання попереднього аналізу даних, а також для реалізації окремих кроків алгоритму обробки екстремальних значень використовувалось інструментальне середовище програмування R2.9.2 для статистичної обробки даних та роботи з графікою. В середовищі програмування R2.9.2 виконано ін-

тегрaцію модулів Rcmdr, extRemes, evdbayes та mcmcPack, функцію 'SimulateData' модуля extRemes.

На рис. 5 подано результат генерування псевдовипадкових даних із заданими вхідними параметрами. Візуальний аналіз отриманої вибірки свідчить про те, що масив даних є досить виродженим і розсіяним по всій числовій осі, не спостерігається збіжності до дельта-околу чіткого інтервалу. Аналіз описових статистик дає можливість припустити про наближення даних до GEV- або GPD-розподілу.

Описові статистики для масиву згенерованих даних такі: $N = 250$; $\mu = 125,0$; median = $= 125,5$; min $= 1$; max $= 250$.

На рис. 6 подано графічні результати оцінювання GEV-моделі, а саме: графік розподілу ймовірностей, графік квантилів, графік щільності розподілу і графік "рівнів повернення". Найкращим наближенням експериментальних даних до побудованої моделі є наявність ланцюжка на діагоналі ймовірності та графіка квантилів без значних відхилень від візуальної прямої. Графік квантилів порівнює квантилі моделі з даними емпіричних квантилів. Графік квантилів, який значно відхиляється від прямої лінії, свідчить про хибність отриманої оцінки та відповідних припущень щодо належності експериментальних даних до вибраного класу розподілів. Графік рівня повернення ілюструє наявність взаємозв'язку між періодом часового

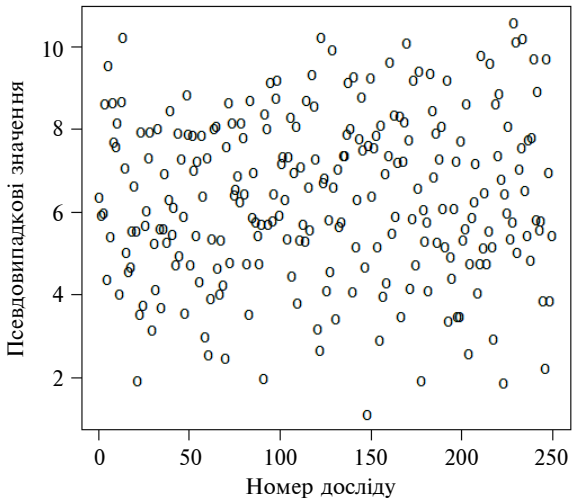

Рис. 5. Результати машинного генерування вибірки об'ємом 250 точок і з такими значеннями вхідних параметрів: $\mu = 5,86; \sigma = 1,97; \xi = -0,36$

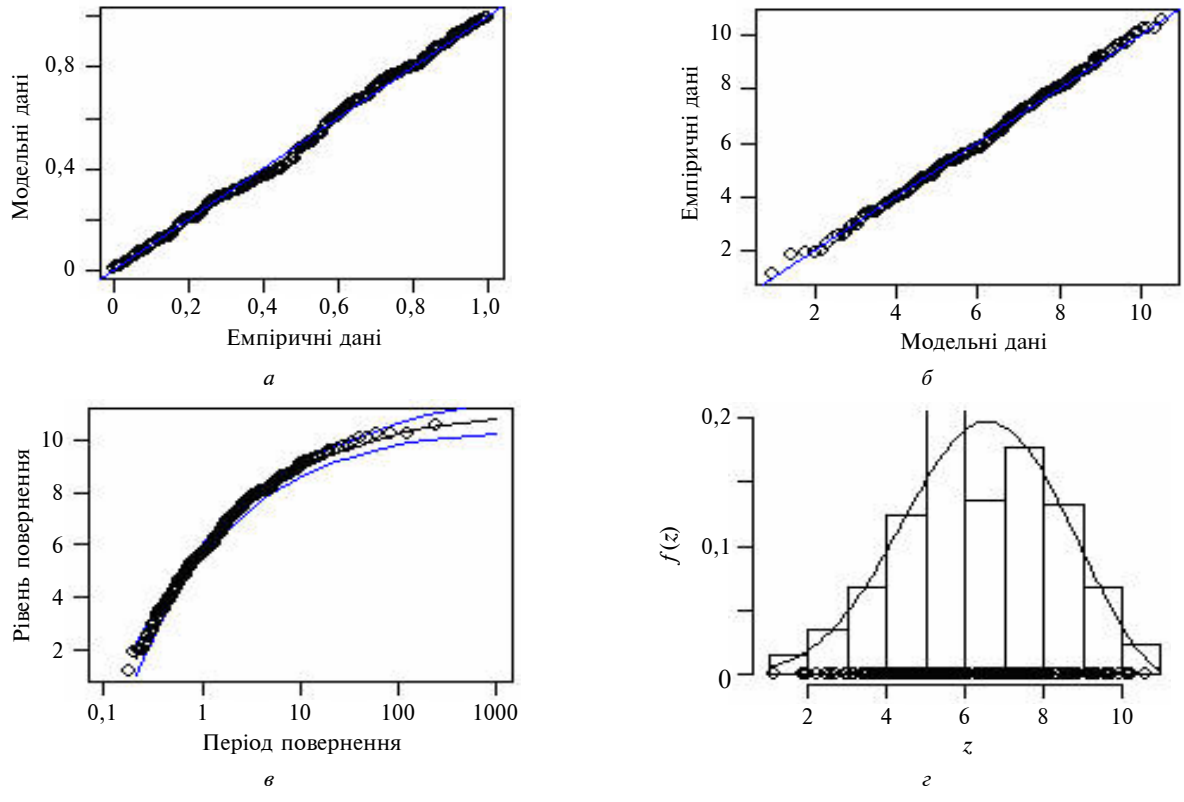

Рис. 6. Графічне зображення оціненої GEV-моделі для 1-го експерименту: *а* – графік розподілу ймовірностей; *б* – графік квантилів; *в* – графік "рівнів повернення"; *г* – графік щільності розподілу

ряду та рівнем повернення і відображає оцінку 95 %-ного довірчого інтервалу. “Рівень повернення” для цього експерименту – це деякий уявний рівень на множині псевдовипадкових значень, на якому спостерігаються значні стрибки значень у середньому за певний період часу (уявний рік, місяць тощо). Період часового ряду – це часова одиниця, за яку відбуваються конкретні перевищення “рівня повернення”. Наприклад, із рис. 6 можна було б очікувати максимальне значення для статистичної вибірки, наближене до 10 у середньому на кожному 100-му експерименті. Числові результати оцінювання побудованої GEV-моделі за допомогою методу максимальної правдоподібності такі: $\mu = 5,738\ (0,141)$; $\sigma = 2,015\ (0,103)$; $\chi = -0,362\ (0,043)$. У дужках наведені значення стандартних похибок (СП) оцінок.

Порівняльна характеристика параметрів GEV-розподілів, виходячи з двох масивів експериментальних даних, відображена в таблиці. Аналізуючи показники за двома експериментами, слід зазначити: чим точніше задати параметр масштабованості та форми і при цьому вибрати нульове значення для параметру розподілу, тим кращим буде наближення практичної кривої до теоретичної (рис. 7).

При порівнянні графіків щільності розподілу для побудованих моделей (рис. 6 і 7) помітні значні покращення моделі GEV-розподілів у термінах вибору початкових параметрів, а саме моделі з нульовим значенням параметра

Таблиця. Порівняльна характеристика параметрів GEV-розподілів

Експеримент	Результати	σ – параметр масштабованості		ξ – параметр форми розподілу		μ – параметр розподілу		Правдоподібність
		ММП	СП	ММП	СП	ММП	СП	
1	Емпіричні	1,953	0,71	−0,650	0,09	5,738	0,140	518,829
	Теоретичні	1,97		−0,36		0,586		–
2	Емпіричні	0,955	0,05	0,209	0,05	0,093	0,068	413,192
	Теоретичні	1		0,2		μ = 0		–

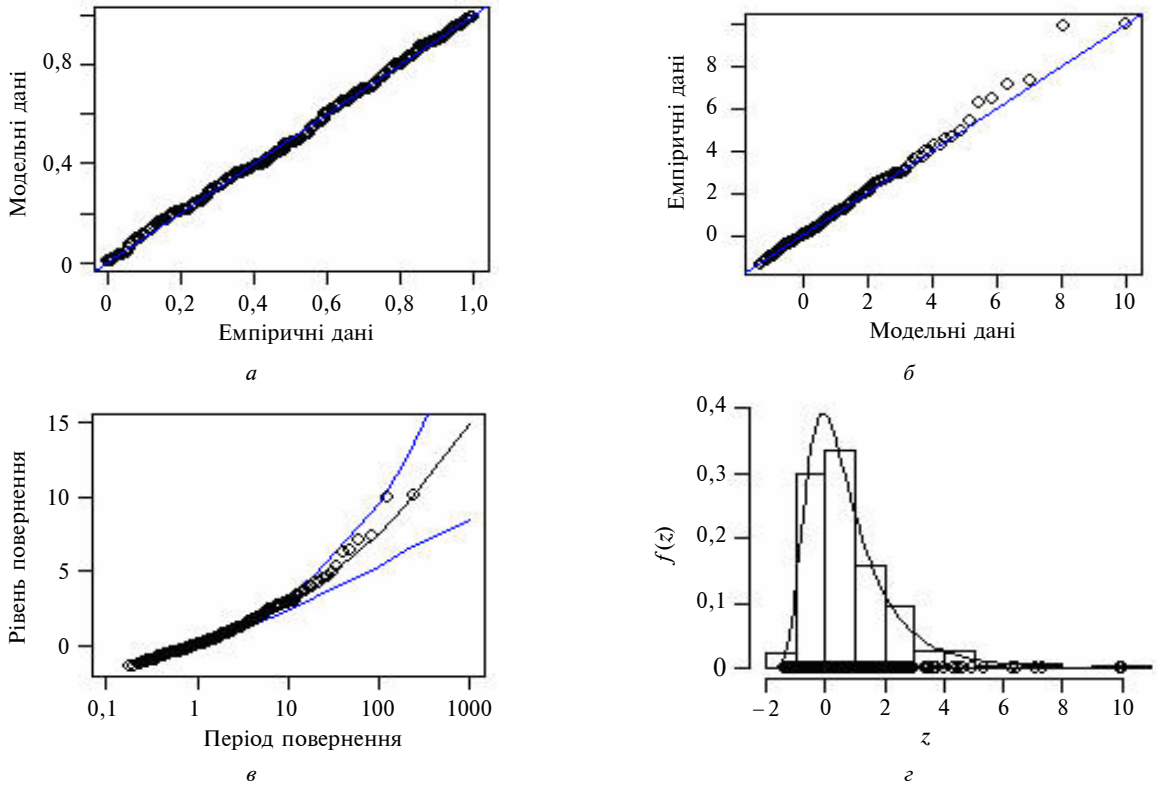

Рис. 7. Графічне зображення оціненої GEV-моделі для 2-го експерименту: *а* – графік розподілу ймовірностей; *б* – графік квантилей; *в* – графік “рівнів повернення”; *г* – графік щільності розподілу

μ. Числові результати оцінок, наведених у таблиці, показують мінімальні відхилення між експериментальними і теоретичними показниками, що свідчить як про високу точність алгоритму генерування псевдовипадкових даних, так і про коректність підібраних значень початкових параметрів.

Висновки

Виконано дослідження стосовно пошуку ефективного алгоритму генерування псевдовипадкових даних та методики обробки екстремальних значень у статистичній вибірці. Розроблено архітектуру СППР для аналізу та генерування псевдовипадкових значень. Запропоновано та експериментально доведено ефективність функціонування створеного багатокрокового підходу з використанням математичного апарату теорії екстремальних значень і алгоритму генерування псевдовипадкових чисел. Розглянутий приклад свідчить про те, що запропонований комплексний підхід стосовно обробки екстремальних значень є ефективним та зручним інструментом аналізу вироджених масивів даних – фактичних і згенерованих.

Залучення комбінованих методів до обробки погано структурованих вироджених статистичних даних розкриває нові можливості для дослідження особливостей сучасних методик моделювання та їх практичного використання. Застосування запропонованої процедури обробки екстремальних значень гарантує високу точність наближення даних до розподілів та уникнення шуму. Порівняння результатів оцінювання параметрів моделей з різними значеннями вхідних параметрів показало, що чим точніше підібрано параметри форми та масштабованості, тим швидша збіжність ряду. Також можна зробити висновок, що СППР на основі методів математичного моделювання і прогнозування процесів різної природи, за умови залучення сучасних методів обробки даних, оцінювання моделей та прогнозів, може бути надійним інструментом для підтримки прийняття управлінських рішень.

У подальших роботах необхідно дослідити можливість використання результатів застосування моделей екстремальних значень при побудові прогнозних узагальнених лінійних моделей з використанням даних різного походження (фактичних і згенерованих).

Список літератури

1. *Coles S.* An Introduction to Statistical Modeling of ExtremeValues. – London: Springer-Verlag, 2009. – 209 p.
2. *Smith R.L.* Extreme Value Theory. – Chapel Hill: University of North Carolina, 2009. – 178 p.
3. *Kuhn M., Johnson K.* Applied Predictive Modeling. – New York: Springer, 2013. – 600 p.
4. *Shumway R.H., Stoffer D.S.* Time Series Analysis and its Applications. – NewYork: Springer, 2006. – 598 p.
5. *Romano A., Secundo G.* Dynamic Learning Methods. – New York: Springer, 2009. – 190 p.
6. *McCullagh P., Nelder J.* Generalized Linear Models. – New York: Chapman & Hall, 1989. – 526 p.
7. *Tsay R.S.* Analysis of Financial Time Series. – New Jersey: John Wiley & Sons, Inc., 2010. – 715 p.
8. *Gilks W.R., Richardson S., Spiegelhalter D.J.* Markov Chain Monte Carlo in Pravtice. – Boca Raton (Florida): Chapman & Hall/CRC Press, LLS, 2000. – 486 p.
9. *Бідюк П.І., Коршевнюк Л.О.* Проектування комп'ютерних інформаційних систем підтримки прийняття рішень. – К.: НТУУ "КПІ", 2010. – 340 с.
10. *Holsapple C.W., Whinston A.B.* Decision Support Systems. – New York: West Publishing Company, 1994. – 860 p.
11. *Лоу А., Кельтон Д.* Имитационное моделирование. – СПб: Питер, 2004. – 848 с.
12. *Beirlant J.* Statistics of Extremes: Theory and Application. – New York: John Wiley & Sons, Inc., 2004. – 505 p.
13. *Бідюк П.І., Трухан С.В.* Оцінювання узагальнених лінійних моделей за байєсівським підходом в актуарному моделюванні // Наукові вісті НТУУ "КПІ". – 2014. – № 6. – С. 49–55.

References

1. S. Coles, *An Introduction to Statistical Modeling of ExtremeValues.* London, UK: Springer-Verlag, 2009, 209 p.
2. R.L. Smith, *Extreme Value Theory.* Chapel Hill: University of North Carolina, 2009. – 178 p.
3. M. Kuhn and K. Johnson, *Applied Predictive Modeling.* New York: Springer, 2013, 600 p.
4. R.H. Shumway and D.S. Stoffer, *Time Series Analysis and its Applications.* New York: Springer, 2006, 598 p.
5. A. Romano and G. Secundo, *Dynamic Learning Methods.* NewYork: Springer, 2009, 190 p.
6. P. McCullagh and J. Nelder, *Generalized Linear Models.* New York: Chapman & Hall, 1989, 526 p.
7. R.S. Tsay, *Analysis of Financial Time Series.* New Jersey: John Wiley & Sons, Inc., 2010,715 p.

8. W.R. Gilks *et al.*, *Markov Chain Monte Carlo in Practice*. Boca Raton: Chapman and Hall/CRC, 2000, 486 p.
9. P.I. Bidyuk and L.O. Korshevnyuk, *Design of Decision Support Systems*. Kyiv, Ukraine: NTUU KPI, 2010, 340 p. (in Ukrainian).
10. C.W. Holsapple and A.B. Whinston, *Decision Support Systems*. New York: West Publishing Company, 1994, 860 p.
11. A.M. Law and W.D. Kelton, *Simulation Modeling and Analysis*. New York, McGraw Hill, 2000, 760 p.
12. J. Beirlant, Statistics of Extremes: *Theory and Application*. New York: John Wiley & Sons, Inc., 2004, 505 p.
13. P.I. Bidyuk and S.V. Trukhan, "Estimation of generalized linear models using Bayesian approach in actuarial modeling", *Naukovi Visti NTUU KPI*, no. 6, pp. 49–55, 2014 (in Ukrainian).

М.З. Згуровський, С.В. Трухан, П.І. Бідюк

МЕТОДИКА ЗАСТОСУВАННЯ ТЕОРІЇ ЕКСТРЕМАЛЬНИХ ЗНАЧЕНЬ ДЛЯ АНАЛІЗУ ДАНИХ

Проблематика. Для розв'язання задач моделювання і прогнозування на основі великих масивів (у тому числі вироджених) даних в умовах наявності невизначеності необхідно розробляти інтегровані інформаційні системи підтримки прийняття рішень (СППР). Пропонується методика застосування теорії екстремальних значень для побудови статистичних моделей та СППР на їх основі.

Мета дослідження. Мета роботи полягає у застосуванні теорії екстремальних значень для аналізу й оцінювання параметрів моделей на основі випадкових вибірок даних. Необхідно розробити ефективну методику аналізу псевдовипадкових значень та оцінювання невідомих параметрів статистичних моделей; навести приклади аналізу за допомогою теорії екстремальних значень і створеного програмного інструментарію.

Методика реалізації. Для розв'язання поставлених задач використано такі методи: процедури генерування псевдовипадкових послідовностей, ймовірнісні розподіли теорії екстремальних значень і методи оцінювання невідомих параметрів моделей. Запропоновано багатокрокову методику обробки екстремальних значень і розроблено СППР для аналізу та моделювання випадкових послідовностей.

Результати дослідження. За допомогою створеної СППР і згенерованих статистичних даних, а також запропонованої методики побудовано процедуру аналізу екстремальних значень. Процедура призначена для подальшого застосування при оцінюванні прогнозних моделей процесів різної природи. Виконано порівняльний аналіз характеристик параметрів GEV-розподілів.

Висновки. За допомогою розробленого інструментарію показано, що запропонований підхід до обробки екстремальних значень є зручним для аналізу вироджених масивів даних. Це підтверджується максимальним наближенням емпіричної кривої до теоретичної функції щільності розподілу. Порівняння результатів оцінювання параметрів моделей показало, що уточнення параметрів форми і масштабованості сприяє прискоренню збіжності оцінок.

Ключові слова: теорія екстремальних значень; метод максимальної правдоподібності; імітаційне моделювання; система підтримки прийняття рішень.

М.З. Згуровский, С.В. Трухан, П.И. Бидюк

МЕТОДИКА ПРИМЕНЕНИЯ ТЕОРИИ ЭКСТРЕМАЛЬНЫХ ЗНАЧЕНИЙ ДЛЯ АНАЛИЗА ДАННЫХ

Проблематика. Для решения задач моделирования и прогнозирования на основе больших массивов (в том числе вырожденных) данных в условиях наличия неопределенностей необходимо разрабатывать интегрированные информационные системы поддержки принятия решений (СППР). Предлагается методика применения теории экстремальных значений для построения статистических моделей и СППР на их основе.

Цель исследования. Цель работы состоит в применении теории экстремальных значений для анализа и оценивания параметров моделей на основе случайных выборок данных. Необходимо разработать эффективную методику анализа псевдослучайных значений и оценивания неизвестных параметров моделей; привести примеры анализа с помощью теории экстремальных значений и программного инструментария.

Методика реализации. Для решения поставленных задач использованы такие процедуры и методы: процедуры генерирования псевдослучайных последовательностей, вероятностные распределения теории экстремальных значений и методы оценивания неизвестных параметров моделей. Предложена многошаговая методика обработки экстремальных значений, и разработана СППР для анализа и моделирования случайных последовательностей.

Результаты исследования. С помощью созданной СППР и сгенерированных статистических данных, а также предложенной методики построена процедура анализа экстремальных значений. Процедура предназначена для дальнейшего использования при оценивании прогнозных моделей процессов различной природы. Выполнен сравнительный анализ характеристик параметров GEV-распределений.

Выводы. С помощью разработанного программного инструментария показано, что предложенный подход к обработке экстремальных значений представляет собой удобный инструмент анализа вырожденных массивов данных. Это подтверждается максимальным приближением эмпирической кривой к теоретической функции плотности распределения. Сравнение результатов оценивания параметров показало, что уточнение параметров формы и масштаба способствует ускорению сходимости оценок.

Ключевые слова: теория экстремальных значений; метод максимального правдоподобия; имитационное моделирование; система поддержки принятия решений.

Рекомендована Радою
Навчально-наукового комплексу
"Інститут прикладного системного
аналізу" НТУУ "КПІ"

Надійшла до редакції
18 грудня 2015 року