МАТЕМАТИЧНЕ ТА КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ

МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

MATHEMATICAL AND COMPUTER MODELLING

УДК 519.766.4

Трухан С. В.[1], Бідюк П. І.[2]

[1]Аспірантка інституту прикладного системного аналізу НТУУ «КПІ», Київ, Україна
[2]Д-р техн. наук, професор кафедри математичних методів системного аналізу НТУУ «КПІ», Київ, Україна

МЕТОДИКА АНАЛІЗУ ЕКСТРЕМАЛЬНИХ ДАНИХ ТА ЇЇ ВИКОРИСТАННЯ ПРИ ОЦІНЮВАННІ ПАРАМЕТРІВ УЗАГАЛЬНЕНИХ ЛІНІЙНИХ МОДЕЛЕЙ

Запропонована методика аналізу екстремальних значень з метою її застосування при оцінюванні невідомих параметрів узагальнених лінійних моделей. В якості математичного апарату використано теорію екстремальних значень, яка є одним із розділів математичної статистики та пов'язана з дослідженням відхилень екстремальних значень від медіани у ймовірнісних розподілах. Також розглянуто методи наближення експериментальних даних до класу узагальнених екстремальних розподілів, методи оцінювання невідомих параметрів та вибору оптимального порогу для екстремальних значень. На основі фактичних статистичних даних із галузі страхування та запропонованого підходу побудовано моделі обробки екстремальних значень для подальшого застосування при оцінюванні прогнозних моделей. Прийнятним для подальшого використання виявилась модель з наближенням даних за допомогою узагальненого розподілу Парето. Це підтверджується незначною похибкою та максимальним наближенням емпіричної кривої до теоретичної функції щільності розподілу. Порівняння результатів оцінювання невідомих параметрів моделі за допомогою методу максимальної правдоподібності та байєсівського підходу показало, що байєсівські методи оцінювання є ефективним підгрунтям для розв'язання задачі вибору кращої моделі на основі множини отриманих альтернатив та значень апріорних параметрів. Можливість використання результатів застосування моделей екстремальних значень при побудові прогнозних узагальнених лінійних моделей є підставою для подальшого дослідження.

Ключові слова: теорія екстремальних значень, узагальнені лінійні моделі, поріг екстремального значення, метод максимальної правдоподібності, байєсівський підхід.

НОМЕНКЛАТУРА

GEV – Generalized extreme value;

GPD – Generalized Pareto distribution;

ММП – метод максимальної правдоподібності;

МКМЛ – метод Монте-Карло для марковських ланцюгів;

ТЕЗ – теорія екстремальних значень;

УЛМ – узагальнені лінійні моделі;

F – функція розподілу випадкової величини;

u – поріг випадкової величини;

$X_1,...,X_n$ – послідовність незалежних випадкових величин;

μ – параметр розподілу;

ξ – параметр форми розподілу;

σ – параметр масштабованості.

ВСТУП

У зв'язку з необхідністю розв'язання нових задач моделювання і прогнозування на основі великих обсягів вироджених вхідних даних, які не можна розв'язати з використанням існуючих методів, виникає потреба у розробці нових інтегрованих інформаційних систем, методів та підходів до обробки таких даних. Одним із таких підходів є ТЕЗ. Вона широко застосовується до розв'язання таких задач як регулювання структури портфелю активів у страхуванні, аналіз виникнення ризикових ситуацій у сфері фінансів та кредитування, прогнозуванні трафіку в галузі телекомунікацій.

Задачею теорії екстремальних значень є цілеспрямований аналіз та 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 об'ємів вибірки. Наприклад, в області страхування будь-якого майна рідкісною, але ймовірною подією є настання страхового випадку, яке повинно супроводжуватись виплатами страхових премій.

DOI 10.15588/1607-3274-2016-1-3

Саме тому для розв'язання задачі прогнозування страхових виплат пропонується ймовірнісна модель, яка будується із застосуванням теорії екстремальних значень. В свою чергу, одним із ключових моментів побудови адекватної моделі досліджуваного процесу є коректний вибір методу оцінювання параметрів математичних моделей за експериментальними (статистичними) даними. Для розв'язання задачі оцінювання невідомих параметрів моделі часто застосовують метод максимальної правдоподібності 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. Перевагою даного підходу є можливість його застосування до обробки статистичних вибірок відносно малих розмірів, а також за наявності пропусків даних [4, 5]. Популярним і відносно універсальним є на сьогодні МКМЛ, який застосовують для оцінювання параметрів лінійних і нелінійних моделей [6–8].

1 ПОСТАНОВКА ЗАДАЧІ

У роботі ставиться за мету застосування теорії екстремальних значень для побудови комплексної моделі обробки екстремальних даних з метою створення УЛМ та оцінювання їх параметрів.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

1) дослідити властивості розподілів екстремальних значень;

2) дослідити методи оцінювання невідомих параметрів моделей екстремальних значень, зокрема можливості використання байєсівського підходу, методу максимальної правдоподібності та ін.;

3) розробити комплексну модель обробки екстремальних значень;

4) навести приклади застосування комплексної моделі для обробки вироджених статистичних даних у страхуванні.

2 ОГЛЯД ЛІТЕРАТУРИ

На ранньому етапі створення статистичної теорії оцінювання найбільша увага приділялась розв'язанню задач наближення кривих розподілу до даних, а значно пізніше – розвитку теорії побудови статистичного висновку. На сьогодні теорія екстремальних значень є складовою частиною багатьох напрямів розвитку практичних наук, таких як гідрологія, астрономія, телекомунікації, економіка та ін. Перші історичні свідчення стосовно існування сімейства розподілів екстремальних значень пов'язані з роботою М. Бернуллі (1709 р.) стосовно визначення середньої тривалості життя. Перші спроби дослідження теорії екстремальних значень грунтувались на використанні нормального розподілу. У 1925 р. Тіппет обчислив ймовірності найбільших значень у нормально розподіленій вибірці із врахуванням різних об'ємів вибірки (до 1000 значень), а також оцінював середній розмах нормально розподілених вибірок (від 2 до 1000 значень). Таблиці Тіппета – це фундаментальний підхід до практичного застосування найбільших величин у вибірці з нормальним розподілом. Саме те, що більшість досліджень грунтувалось на нормальному розподілі, гальмувало розвиток теорії екстремальних значень. Фреше успішно дослідив перший тип розподілу екстремальних даних та отримав граничні розподіли найбільших величин вибірки, запропонував постулат стійкості. Використовуючи даний постулат Фреше та Тіппет винайшли два інші розподіли екстремальних значень та підкреслили повільну збіжність ряду границі розподілів найбільших величин із нормальної вибірки [1].

Проблема недостатніх об'ємів інформації часто зустрічається при дослідженні процесів економічного, фізичного, природничого походження і стає причиною труднощів при розв'язанні задач побудови моделей таких екстремальних даних. Тому виникає потреба у дослідженні інших джерел знань для пошуку оптимальних рішень стосовно обробки даних та побудови моделей. Наприклад, економісту потрібно знайти максимальне значення з деякої вибірки з виродженими даними. Звичайно існує кілька можливих способів, за якими експерт із знаннями досліджуваного процесу може надати інформацію, що має відношення до екстремальної поведінки і яка залежить від наявних даних. Але часто така інформація супроводжується наближеними вимірами, які відрізняються від дійсних значень та роблять хибними майбутні прогнози, що будуються на їх основі.

Тому, виходячи із актуальності задачі обробки екстремальних значень, роботу присвячено дослідженню та розробці комплексної моделі для опису екстремальних значень і оцінюванню невідомих параметрів УЛМ. Такі моделі широко використовують для аналізу страхових випадків, прогнозування продовження старих чи укладення нових страхових договорів, розробці тарифів та андерайтингу, а також у цільовому маркетингу.

3 МАТЕРІАЛИ І МЕТОДИ

Математичну модель екстремальних даних можна представити у вигляді [1]:

$$M_n = \max\{X_1,...X_n\}, \tag{1}$$

де $X_1,...,X_n$ – послідовність незалежних випадкових величин з функцією розподілу F. У виразі (1) величина M_n позначає максимум досліджуваного процесу на інтервалі часу n і має розподіл [1]:

$$\Pr\{M_n \le z\} = \Pr\{X_1 \le z,..., X_n \le z\} = \Pr\{X_1 \le z\} \times$$
$$\times ... \times \Pr\{X_n \le z\} = \{F(z)\}^n. \tag{2}$$

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

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{3}$$

де μ – параметр розподілу; σ – параметр масштабованості; ξ – параметр форми розподілу [2].

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

1) Розподіл Гумбела:

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

2) Розподіл Фреше:

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

3) Розподіл Вейбулла:

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

Для всіх трьох випадків $a>0$, b – дійсне число. Для другої та третьої функції параметр $\alpha>0$. Ці три класи розподілів називають розподілами екстремальних значень, вони зображені на рис. 1.

З рис. 1 видно, що кожен з розподілів має свою форму поведінки хвоста. Наприклад, для розподілу Вейбула хвіст має кінцеву точку $z_{\sup}=\frac{\mu-\sigma}{\xi}$, а для розподілів Фреше та Гумбела $z_{\sup}=\infty$. Крім того, щільність розподілу Гумбела експоненціально затухає, тоді як щільність розподілу Фреше затухає поліноміально. Розподіл Гумбела є наближеним до класу таких відомих як нормальний, лог-нормальний та гамма – розподілів. Розподіл Фреше має тяжкий хвіст, який позначається як $E(X^r)=\infty$ для $r\ge\frac{1}{\xi}$ (що означає нескінченність дисперсії при $\xi\ge 1/2$).

В окремий клас виділяють узагальнений розподіл Парето (*Generalized Pareto Distribution – GPD*), який

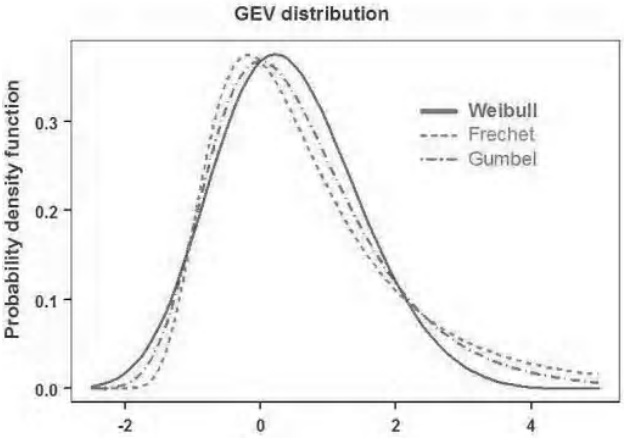

Рисунок 1 – Функції щільності розподілу для трьох типів розподілів

отримуємо за умови: X – це розподіл, що умовно перевищує деякий поріг u:

$$F_u(y)=\frac{F(u+y)-F(u)}{1-F(u)},\qquad(4)$$

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]:

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

1) якщо $\xi>0$, то маємо довгий хвіст $x^{-1/\xi}$, що еквівалентно розподілу Парето;

2) якщо $\xi=0$ та спрямовуючи $\xi\to 0$, отримаємо $G(y,\sigma,0)=1-\exp\left(-\frac{y}{\sigma}\right)$, тобто експоненціальний розподіл з середнім σ;

3) якщо $\xi<0$, то кінцева верхня точка знаходиться на рівні $-\frac{\sigma}{\xi}$.

Також однією із переваг *GEV*-розподілів є інваріантність кожного з розподілів, які належать до даного класу.

Розглянемо методику обробки екстремальних значень. Для обробки статистичного ряду з n-незалежних, однаково розподілених змінних $X_1,...,X_n$ застосовується така послідовність дій.

1. Групування вибірок даних з n спостережень. Такі вибірки повинні містити від 50 до 100 значень.

2. Визначається максимум Z_i для кожного блоку i.

3. 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 i-го року.

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

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

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

$$z_p=\begin{cases}\mu-(\sigma/\xi)\left(1-(y_p)^{-\xi}\right),\ \ \xi\ne 0;\\ \mu-\sigma\log(y_p),\ \ \xi=0;\end{cases}$$

Якщо зобразити z_p в залежності 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].

4. Виконується оцінювання параметрів моделі та розв'язується задача пошуку оптимальної довжини блоку.

Остання 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 році з такими результатами [3]:

– якщо $\xi > -0{,}5$, то оцінювання за ММП носить 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 [5].

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 $i = 1,...,m$. Як тільки остання умова не виконується, то функція правдоподібності дорівнює нулю і логарифмічна функція правдоподібності набуває значення нескінченності.

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

$$l(\mu,\sigma) = -m\log\sigma - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{m}\left(-\frac{z_i-\mu}{\sigma}\right). \quad (5)$$

Після використання методів чисельної оптимізації та максимізації виразу (5), отримуємо оцінку максимальної правдоподібності вигляду $\left(\hat{\mu},\hat{\sigma},\hat{\xi}\right)$ [3, 5].

5. Графічна перевірка наближення *GEV*-моделей.

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

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

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

Апроксимуюча функція щільності розподілу в тій самій точці виглядає так:

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

Для того, щоб отримати найкраще наближення моделі необхідно задовольнити рівність $\widetilde{G}(Z_i) = \hat{G}(Z_i)$. За допомогою цього графіка на практиці часто вдається запобігти ефекту «виродженості». Тобто, коли множина точок $\left\{\widetilde{G}(Z_i),\ \hat{G}(Z_i)\right\},\ i = 1,...,m$ – лежить близько до першої діагоналі в той час, коли 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 у розподілі. На практиці більшість економічних процесів асиметричні, а фінансові ряди мають вироджений хвіст. Саме графік квантилів дає можливість оцінити ступінь довіри для ряду параметричних моделей. Графік квантилів визначається як множина точок [4]:

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

Якщо параметрична надає прийнятне згладжування, то графік має лінійну форму. Тому графік дає можливість порівняти оцінені моделі та вибрати найкращу; оцінити як обрана 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].

6. Визначення порогу 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 розподі-

лених змінних з функцією розподілу F. Тоді умовна ймовірність визначається так:

$$F_u(y) = P(X \le u + y \mid X > u), \text{ або}$$

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

Цей вираз дозволяє визначити ступінь наближення значень ймовірності для великих значень порогу u.

Задача вибору оптимального порога ідентична задачі визначення розміру блока. 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 порогу 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 порогу 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 порогу, а всі інші припустити проміжними при оцінюванні невідомих параметрів моделі. Середнє для обох випадків визначається так [5]:

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

$$e(u) = E(X - u / X > u) = \tilde{\sigma}_u/(1-\xi) = (\tilde{\sigma}_{u_0} + \xi(u - u_0))/(1-\xi). \quad (6)$$

Оскільки $e(u) = E(X - u / X > u)$ – це лінійна функція від u, то враховуючи вираз (6), оцінювання величини порогу можна виконати за такою інструкцією [3, 10]:

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 дослідів, які перевищують u; $x_{\max}$ – верхня межа досліджуваного значення;

2) вибрати порогове значення, над яким графік приймає наближено лінійний характер стосовно u. Застосування довірчих інтервалів допомагає визначити цю точку.

Також, для визначення порогу екстремального значення використовують метод умовно прийнятного вибору, який базується на такому правилі: поріг встановлюється у тому регіоні, де хвіст становить 5–10% від усієї вибірки. Головне припущення: він не повинен бути більшим ніж 10–15%. На практиці 10% межу часто використовували у своїх дослідах Роко (2011), Макнейл і Фрей (2000) [10].

7. Оцінювання невідомих параметрів моделі.

Після кроку визначення порогу потрібно виконати оцінку невідомих параметрів узагальненого розподілу Парето. Як відомо серед методів оцінювання невідомих параметрів моделі поширеним є метод максимальної правдоподібності.

Нехай $y_1, ..., y_k$ – це значення k-залишків з порогу; 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:

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

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 до оцінювання невідомих параметрів було проілюстровано на прикладі узагальнених лінійних моделей [9, 10].

Крім того, даний підхід надає обґрунтовану альтернативу для випадків, коли припущення, необхідні для застосування методу максимальної правдоподібності та ймовірності зважених моментів не виконуються.

4 ЕКСПЕРИМЕНТИ

Експериментальне дослідження ефективності запропонованої методики виконано за допомогою фактичних статистичних даних. Об'єм статистичної вибірки складав 247 вимірів, які включають такі змінні: назва страхової компанії; грошовий еквівалент страхових виплат; статистичний рік; кількість договорів, які уклала конкретна страхова компанія; страхові платежі; кількість страхових випадків на рік. Основна залежна змінна – страхові виплати, яка відображає здійснення грошових переказів при настанні страхового випадку. Решта змінних, які включені до вибірки, є незалежними і беруться до уваги як фактори.

Для виконання попереднього аналізу статистичних даних та реалізації окремих кроків алгоритму обробки екстремальних значень використовувались такі програмні продукти: Microsoft Excel 2010; інструментальне середовище програмування R2.9.2 для статистичної обробки даних та роботи з графікою; економетричний пакет Eviews 8.0 для побудови моделей та попереднього оцінювання невідомих параметрів. В пакеті Eviews 8.0 використано такі модулі: розрахунок описових статистик, побудова УЛМ, метод максимальної правдоподібності для оцінювання параметрів моделі. В середовищі програмування R2.9.2 виконано інтеграцію модулів Rcmdr, extRemes, evdbayes та mcmcPack.

5 РЕЗУЛЬТАТИ

На рис. 2 відображено графік залежності страхових виплат від статистичного року. Різкі зміни величини «*Страхові виплати*» пояснюється коливаннями величини «Кількість страхових випадків» для відповідного періоду. На рис. 3 відображено значення описових статистик. Із рис. 3 помітно, що коефіцієнт асиметрії (Skewness) коливається в межах 2,839 до 8,664. А це в свою чергу свідчить про наявність «правого хвосту» в розподілі. Так, як параметр ексцессу (Kurtosis) має значення більше трьох, то розподіл є гостровершинним.

Рисунок 2 – Залежність страхових виплат від статистичного року

Також, попередній аналіз початкових даних свідчить про сильну виродженість вибірки, яка проявляється у вигляді шуму при побудові моделі, на прикладі рис. 4. Саме тому прийнято рішення про доречність попереднього логарифмування даних.

Аналіз описової статистики та візуальний аналіз логарифмованих даних (рис. 5) дають можливість припустити про наближення даних до *GEV*- або *GPD*-розподілу.

Відносно високий поріг вибирається з метою того, щоб зменшити зміщення моделі, а з іншої сторони – це буде означати, що лише декілька дослідів використовуються для оцінювання параметрів розподілу, тим самим гарантуючи збільшення оцінки дисперсії. Мета вибору величини порогу полягає в тому, щоб уникнути зміщення моделі. Згідно розглянутого вище методу визначення величини порогу для експерименту прийнято значення 6,65. Графік Mean Residual Life Plot відображає залежність порогу від середнього залишку для оціненої моделі. Він слугує важелем перевірки вибраного порогу. З рис. 6 видно, що після значення порогу 6 з'являються помітні відхилення від лінійності.

View	Proc	Object	Print	Name	Freeze	Sample	Sheet	Stats	Spec
			Q_CASES	Q_ARRANG	DAMAGES	CHARGES			
Mean			248.9717	42819.83	2779.190	47154.60			
Median			65.00000	2431.000	780.0000	13017.40			
Maximum			3800.000	2241084.	49575.00	557884.0			
Minimum			2.000000	6.000000	1.200000	469.8000			
Std. Dev.			526.4061	181679.5	6121.650	81935.75			
Skewness			4.241486	8.663552	4.644020	2.838897			
Kurtosis			24.51794	94.69488	29.05691	12.32584			
Jarque-Bera			5505.864	89621.68	7875.493	1226.855			
Probability			0.000000	0.000000	0.000000	0.000000			
Sum			61496.00	10576499	686459.9	11647185			
Sum Sq. Dev.			68167427	8.12E+12	9.22E+09	1.65E+12			
Observations			247	247	247	247			

Рисунок 3 – Описові статистики початкових даних

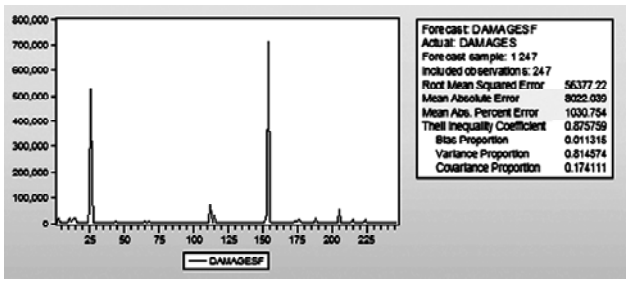

Рисунок 4 – Результати оцінювання моделі без попередньої обробки

Порівняльна характеристика параметрів розподілу представлена в табл. 1. Вона показує, що оптимальним є наближення даних за допомогою *GPD*-розподілу із незначною похибкою та максимальним наближенням емпіричної кривої до теоретичної функції щільності розподілу (рис. 7).

Параметри оцінювання побудованої моделі за допомогою байєсівського підходу зображено на рис. 8. По-

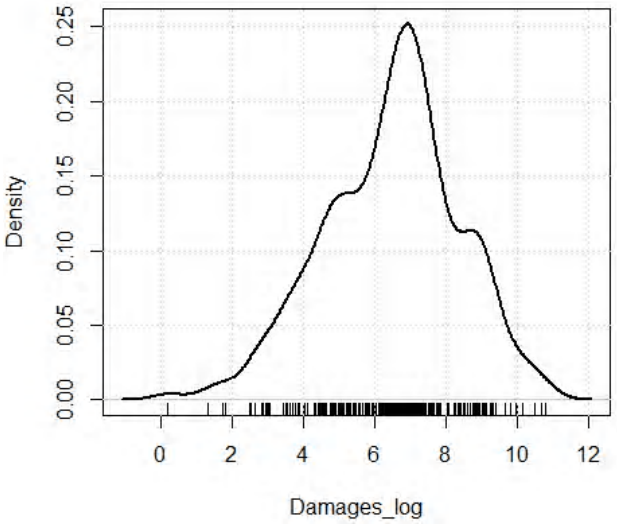

Рисунок 5 – Графік залежності логарифмованих страхових виплат від щільності розподілу

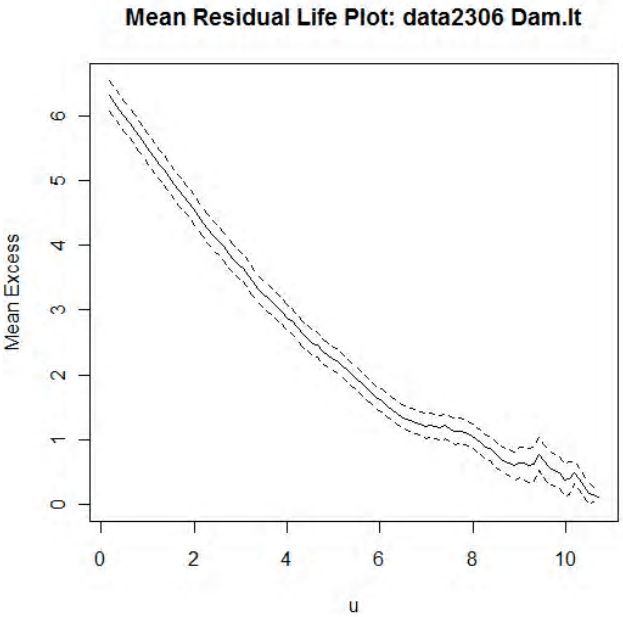

Рисунок 6 – Залежність значення порогу від середнього залишку GPD-моделі

Таблиця 1 – Порівняльна характеристика параметрів розподілів

Тип розподілу	Sigma		Xi		Log-likelihood	Excee-dance rate (per year)	Number of exceedances of threshold
	Maximum likelihood estimation	Std. error	Maximum likelihood estimation	Std. error			
GEV-розподіл	1,953	0,712	–0,650	0,095	487,812	–	–
GPD-розподіл	0,777	0,346	–0,541	0,206	146,369	183,364	124

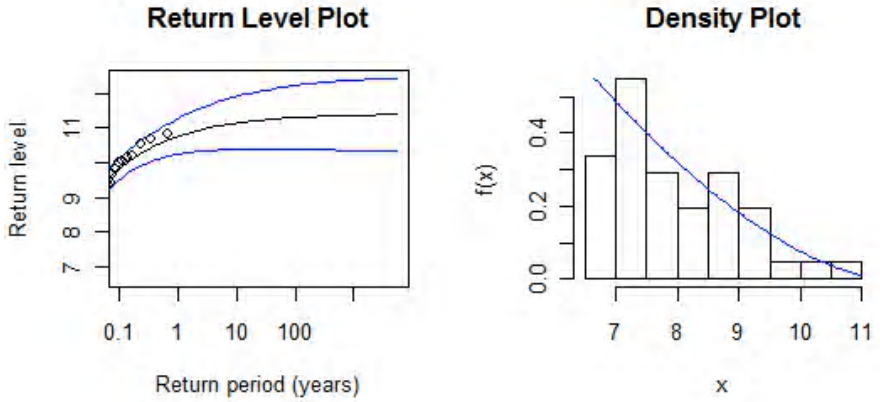

Рисунок 7 – Графічне представлення оціненої GPD-моделі

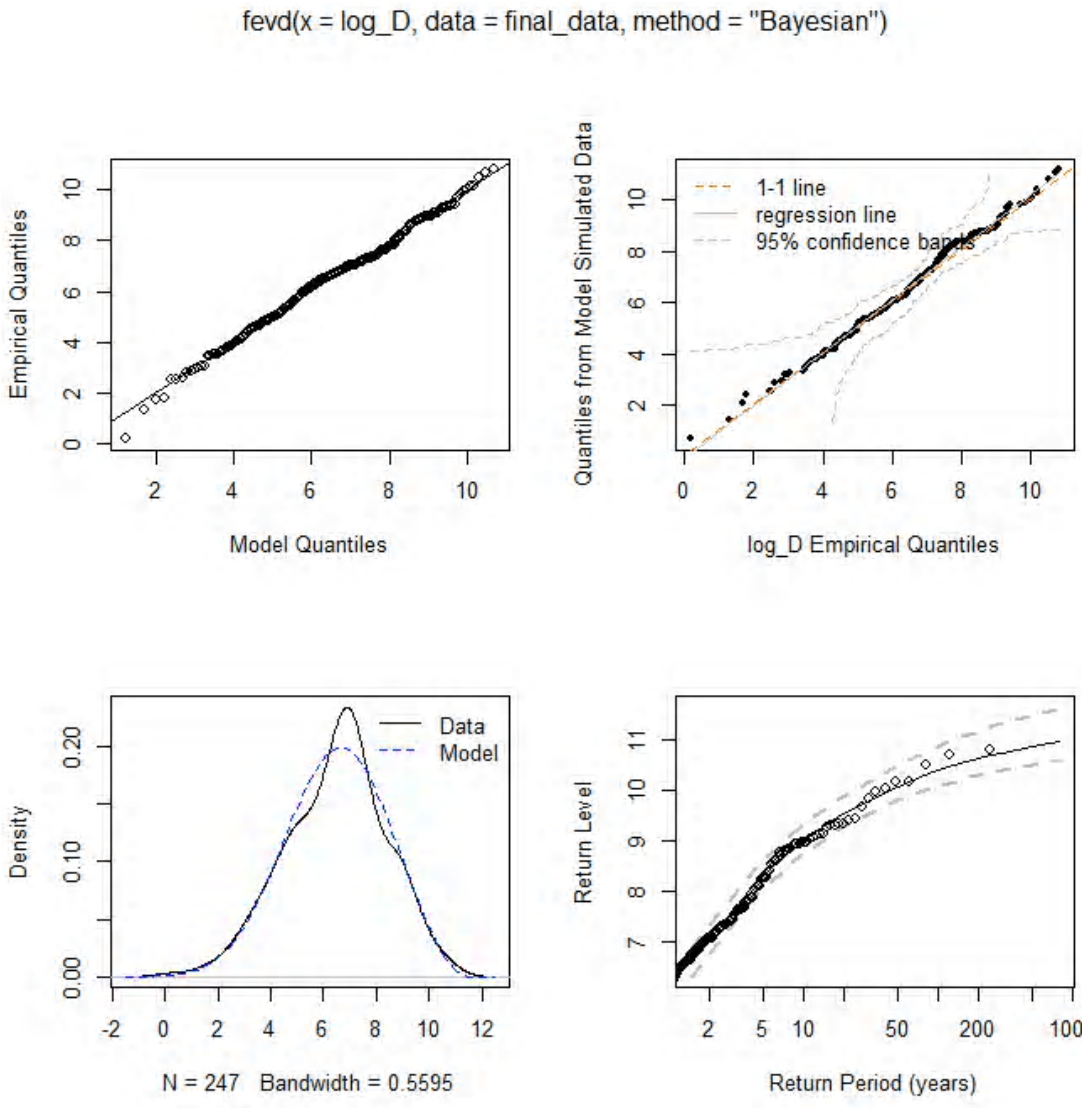

Рисунок 8 – Діагностика наближення моделі до одного з GEV-розподілів

рівнюючи графіки щільності розподілу для побудованої моделі та актуальної вибірки даних помітні значні покращення моделі у термінах належності до одного з GEV-розподілів. Числові значення оцінок невідомих параметрів моделі за допомогою байєсівської методології наведені на рис. 9. Слід зазначити, що на рис. 9 за параметр масштабованості відповідає змінна *scale*, а *shape* – параметр форми. На рис. 10 відображено результати обчислення параметрів апостеріорної вибірки згідно методу Монте-Карло. За допомогою функції «*ci*» було обчислено довірчі інтервали для відповідних параметрів та рівнів повернення (рис. 11). Графічне відображення апріорних оцінок параметрів за методом Монте-Карло та «trace-графіків» наведено на рис. 12.

Порівнюючи результати отриманих оцінок слід зауважити, що байєсівський підхід демонструє кращі результати ніж метод максимальної правдоподібності та сприяє обґрунтованому вибору кращої моделі із запропонованих GEV-розподілів, виходячи зі значень апріорних параметрів, а також алгоритмів вибору кращої моделі.

```
> postmode(fb)
  location      scale      shape
 5.8691196  1.9770899 -0.3610363
```

Рисунок 10 – Результати обчислення параметрів розподілу апріорної вибірки

```
fevd(x = log_D, data = final_data, method = "Bayesian")

[1] "Estimation Method used: Bayesian"


 Acceptance Rates:
log.scale     shape
0.2530506 0.1956391
fevd(x = log_D, data = final_data, method = "Bayesian")

[1] "Quantiles of MCMC Sample from Posterior Distribution"

              2.5% Posterior Mean      97.5%
location  5.5960245      5.8697199  6.1488656
scale     1.7911608      1.9873764  2.2171284
shape    -0.4251922     -0.3529839 -0.2779164
```

Рисунок 9 – Результати оцінювання параметрів моделі за допомогою байєсівської методології

```
> ci(fb)
fevd(x = log_D, data = final_data, method = "Bayesian")

[1] "Quantiles of MCMC Sample from Posterior Distribution"

[1] "Posterior Mean 100-year level: 10.394"

[1] "95% Confidence Interval: (10.0854, 10.8736)"


> ci(fb, type ="parameter")
fevd(x = log_D, data = final_data, method = "Bayesian")

[1] "Quantiles of MCMC Sample from Posterior Distribution"

              2.5% Posterior Mean      97.5%
location  5.5960245      5.8697199  6.1488656
scale     1.7911608      1.9873764  2.2171284
shape    -0.4251922     -0.3529839 -0.2779164
```

Рисунок 11 – Результати обчислення довірчих інтервалів для параметрів форми та масштабованості відповідно та квантилі статистики Монте-Карло для апріорних розподілів

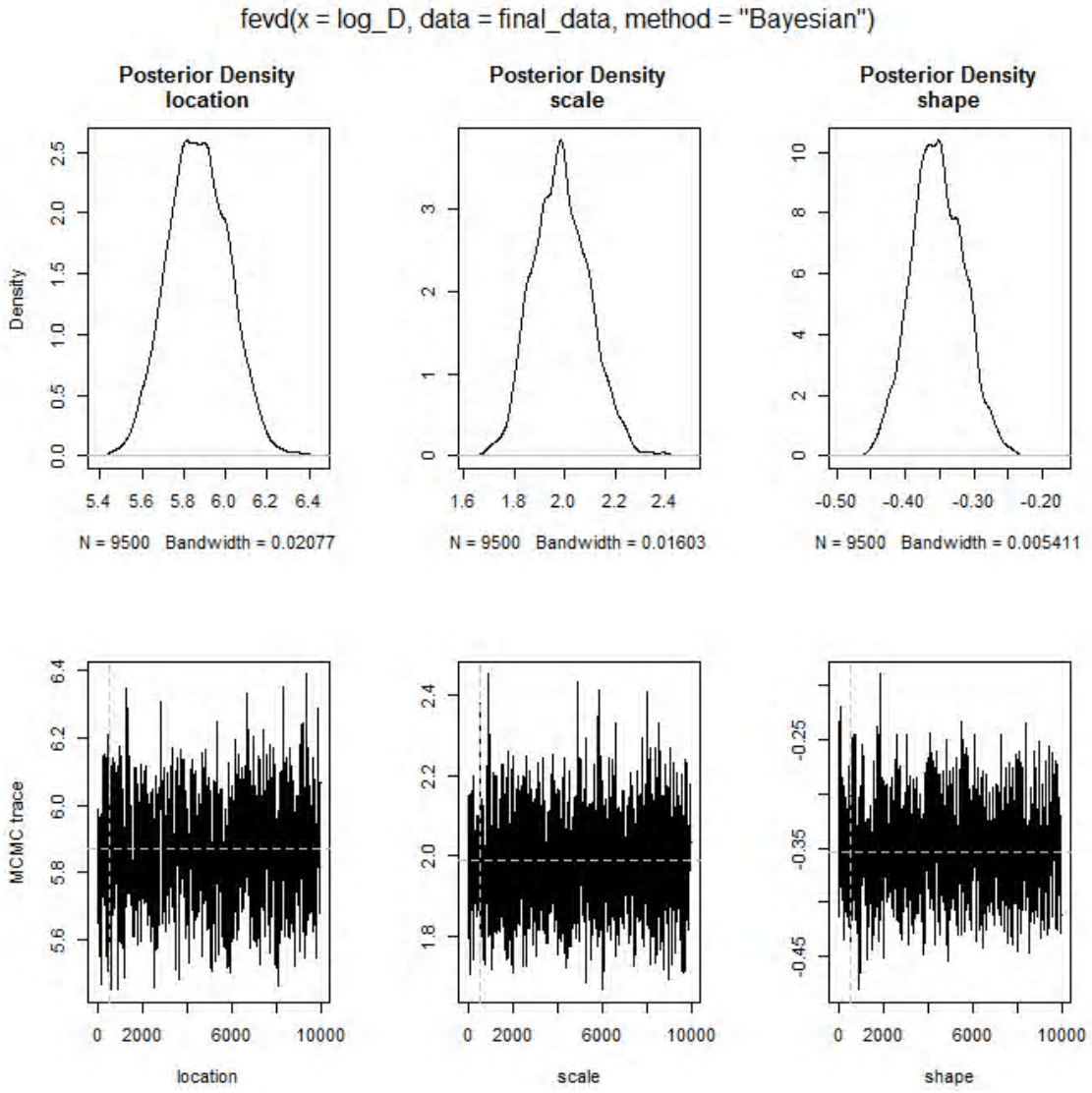

Рисунок 12 – Графічне відображення апріорних оцінок параметрів за методом Монте-Карло та «trace-графіків»

ОБГОВОРЕННЯ

В результаті використання запропонованої комплексної моделі обробки екстремальних статистичних даних вдалося успішно розв'язати проблему невиродженості даних у статистичній вибірці із застосуванням теорії екстремальних значень.

Для оцінювання невідомих параметрів побудованих моделей, які належать до класу GEV-розподілів можна успішно використовувати байєсівський підхід, оперуючи апріорними та апостеріорними розподілами параметрів, а також алгоритмами вибору кращої моделі. Залучення новітніх комбінованих методів до розв'язання задачі обробки екстремальних даних та оцінювання невідомих параметрів, вибору кращої моделі на основі алгоритмів зменшення порогів викидів відкриває нові можливості для дослідження особливостей методів математичного моделювання.

ВИСНОВКИ

Виконано дослідження щодо пошуку ефективної методики обробки екстремальних значень у статистичній вибірці. Запропоновано та експериментально доведено ефективність функціонування створеного багатокрокового підходу із використанням математичного апарату теорії екстремальних значень та методів оцінювання невідомих параметрів моделей. Розглянутий приклад свідчить про те, що запропонований комплексний підхід стосовно обробки екстремальних значень є ефективним та зручним інструментом аналізу вироджених масивів даних та моделювання актуарних процесів. Для оцінювання невідомих параметрів екстремальних моделей зручно використовувати байєсівський підхід, який надає можливість оперувати апріорними та апостеріорними розподілами параметрів і алгоритмами вибору кращої моделі.

Залучення новітніх комбінованих методів до обробки погано структурованих вироджених статистичних даних розкриває нові можливості щодо дослідження особливостей сучасних методик та математичних методів. Надалі необхідно дослідити можливість використання результатів застосування моделей екстремальних значень при побудові прогнозних УЛМ моделей. Застосування запропонованої процедури обробки екстремальних значень гарантує високу точність наближення даних до розподілів та уникнення шуму. Порівняння результатів оцінювання параметрів моделі за допомогою методу максимальної правдоподібності показало, що байєсівські методи оцінювання є кращим підгрунтям для розв'язання задачі вибору кращої моделі на основі множини отриманих альтернатив. Також можна зробити висновок, що сфера страхування, за умови належного менеджменту із застосуванням сучасних матема-

тичних методів обробки даних, оцінювання моделей та прогнозів може бути надійним джерелом стабілізації економіки країни у цілому.

ПОДЯКИ

Роботу виконано відповідно з тематичними планами наукових досліджень Національного технічного університету України «Київський політехнічний інститут». Дослідження виконано в рамках бюджетної НДР, реєстраційний № 0115U000356, тема № 2813–п НТУУ «КПІ»: «Розробка методології системного аналізу, моделювання та оцінювання фінансових ризиків». Страхові дані отримано за сприяння Ліги страхових компаній.

СПИСОК ЛІТЕРАТУРИ

1. Coles S. An Introduction to Statistical Modeling of Extreme Values / S. Coles. – London : Springer-Verlag, 2001. – P. 45–104.
2. Smith R. L. An overview of Extreme value theory / R. L. Smith. – Lausanne : Bernoulli Center, 2009.
3. Mallor F. An introduction to statistical modeling of extreme values / F. Mallor, E. Nualart, E. Omey // Hub research paper. – 2009. – No. 36. – P. 5–31.
4. Shumway R. H. Time series analysis and its applications / R. H. Shumway, D. S. Stoffer. – New York : Springer, 2006. – 598 p.
5. Romano A. Dynamic learning methods / A. Romano, G. Secundo – New York: Springer, 2009. – 190 p.
6. McCullagh P. Generalized Linear Models / P. McCullagh, J. Nelder.– New York : Chapman & Hall, 1989. – 526 p.
7. Tsay R. S. Analysis of financial time series / R. S. Tsay. – New Jersey : John Wiley & Sons, Inc., 2010. – 715 p.
8. Besag J. Markov Chain Monte Carlo for Statistical Inference / J. Besag. – Center for Statistics and the Social Sciences. – 2001. – No. 9. – 25 p.
9. Бідюк П. І. Оцінювання узагальнених лінійних моделей за байєсівським підходом в актуарному моделюванні / П. І. Бідюк, С. В. Трухан // Наукові Вісті НТУУ «КПІ». – 2014. – № 6. – С.49–55.
10. Beirlant J. Statistics of extremes: Theory and application / J. Beirlant. – New York : John Wiley & Sons, Inc., 2004. – 505 p.

Стаття надійшла до редакції 21.10.2015.
Після доробки 28.10.2015.

Трухан С. В.[1], Бидюк П. И.[2]
[1]Аспирантка института прикладного системного анализа НТУУ «КПИ», Киев, Украина
[2]Д-р техн. наук, профессор кафедры математических методов системного анализа НТУУ «КПИ», Киев, Украина

МЕТОДИКА АНАЛИЗА ЭКСТРЕМАЛЬНЫХ ДАННЫХ И ЕЕ ИСПОЛЬЗОВАНИЕ ПРИ ОЦЕНИВАНИИ ПАРАМЕТРОВ ОБОЩЕННЫХ ЛИНЕЙНЫХ МОДЕЛЕЙ

Предложена методика анализа экстремальных значений с целью ее использования при оценивании неизвестных параметров обобщенных линейных моделей. В качестве математического аппарата использована теория экстремальных значений, которая является одним разделов математической статистики и связана с исследованием отклонений экстремальных значений от медианы в вероятностных распределениях. Также рассмотрены методы приближения экстремальных данных к классу обобщенных экстремальных распределений, методы оценивания неизвестных параметров и выбора оптимального порога для экстремальных значений. На основе реальных статистических данных и исследуемого подхода построены модели обработки экстремальных значений для дальнейшего использования при оценивании прогнозных моделей. Допустимой для дальнейшего применения оказалась модель приближения данных с помощью обобщенного распределения Парето. Это обосновывается минимальной величиной погрешности, а также максимальным приближением эмпирической кривой к теоретической функции плотности распределения. Сравнение результатов оценивания неизвестных параметров модели с помощью метода максимального правдоподобия и байесовского подхода показало, что байесовские методы оценивания являются эффективным основанием для решения задачи выбора лучшей модели исходя из множества полученных альтернатив и значений априорных параметров. Для дальнейшего исследования целесообразно рассмотреть задачу применения моделей экстремальных значений при построении прогнозных обобщенных линейных моделей.

Ключевые слова: теория экстремальных значений, обобщенные линейные модели, порог экстремального значения, метод максимального правдоподобия, байесовский подход.

Trukhan S.[1], Bidyuk P.[2]
[1]Post-graduate student of Institute for Applied System Analysis, NTUU «KPI», Kyiv, Ukraine
[2]Dr. Sc., Professor at the Department of Mathematical methods for System Analysis, NTUU «KPI», Kyiv, Ukraine

METHODOLOGY OF EXTREME VALUES ANALYSIS AND ITS APPLICATION FOR PARAMETER ESTIMATION OF GENERALIZED LINEAR MODELS

The article deals with methodology of extreme values treatment for building and estimating unknown parameters of generalized linear models. As a mathematical tool for carrying out the research the extreme value theory was used that creates one of the directions in mathematical statistics, and is related to investigating the extreme deviations from the median values in probability distributions. Also, the methods of approximation statistical data to generalized extreme value distribution, the methods of estimating unknown parameters and selecting an optimal threshold for extreme value models are discussed. The models of treatment extreme values are constructed which are based on actual statistical data and approach is proposed for their future application for estimating predictive models. The model with generalized Pareto distribution turned out to be acceptable for further use, because it has minimum value of observation error and the best approximation of observed curve to theoretical density function. The comparison of evaluation unknown models' parameters using method of maximum likelihood and Bayesian approach leads to next conclusion. The Bayesian methods are efficient way to solve the problem of selection the best model, based on the received alternatives set and prior parameters values. In future studies it will be reasonable to consider the application of extreme value analysis to predicted generalized linear models.

Keywords: extreme value theory, generalized linear models, extreme value threshold, maximum likelihood method, Bayesian approach.

REFERENCES

1. Coles S. An Introduction to Statistical Modeling of Extreme Values. London, Springer-Verlag, 2001, pp. 45–104.
2. Smith R. L. An overview of Extreme value theory. Lausanne, Bernoulli Center, 2009.
3. Mallor F., Omey E. An introduction to statistical modeling of extreme values. Hub research paper, 2009, No. 36, pp. 5–31.
4. Shumway R. H., Stoffer D. S. Time series analysis and its applications. New York, 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. Besag J. Markov Chain Monte Carlo for Statistical Inference, Center *for Statistics and the Social Sciences*, 2001, No. 9, 25 p.
9. Bidyuk, P., Trukhan, S. Estimation of generalized linear models using Bayesian approach in actuarial modeling, *Naukovi Visti NTUU «KPI»*, 2014, No. 6, pp. 49–55.
10. Beirlant J. Statistics of extremes: Theory and application. New York, John Wiley & Sons, Inc., 2004, 505 p.