Statističko modeliranje ekstremnih vrijednosti - metoda maksimuma blokova

Spajić, Anđela

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Anđela Spajić

STATISTIČKO MODELIRANJE EKSTREMNIH VRIJEDNOSTI — METODA MAKSIMUMA BLOKOVA

Diplomski rad

Voditelj rada:
doc. dr. sc. Hrvoje Planinić

Zagreb, rujan, 2022.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Teorija ekstremnih vrijednosti statistička je disciplina koja se bavi procjenom vjerojatnosti događaja koji su na određeni način ekstremniji od svih prethodno opaženih. Premda njezini začetci sežu na sami početak dvadesetoga stoljeća, tehnike potrebne za primjenu teorije ekstremnih vrijednosti u modeliranju različitih fizikalnih fenomena razvijaju se tek pedesetih godina, nakon čega počinje široka primjena ove discipline u graditeljstvu, financijama, biomedicini, oceanografiji, meteorologiji, seizmologiji i dr., a koja traje sve do danas.
Osnovna ideja pristupa teorije ekstremnih vrijednosti ekstrapolacija je budućih ekstremnih vrijednosti iz dostupnih povijesnih podataka koristeći modele koji se oslanjaju na asimptotske rezultate. U najjednostavnijem slučaju takav pristup podrazumijeva shvaćanje dostupnih povijesnih podataka kao realizacija slučajnih varijabli jednake, ali nepoznate distribucije, grupiranje podataka u skupine, odnosno *blokove*, unaprijed određene veličine te promatranje ponašanja niza koji čine maksimumi (ili minimumi) dobivenih blokova. Teorija ekstremnih vrijednosti tada osigurava okvir unutar kojega je, pod određenim uvjetima i koristeći samo dostupne podatke, moguće aproksimirati distribuciju niza tako definiranih maksimuma. Drugim riječima, uporabom rezultata teorije ekstremnih vrijednosti moguće je donositi predviđanja i zaključke o budućim ekstremnim vrijednostima promatrane pojave.
Središnja je tema ovoga rada metoda maksimuma blokova, koja predstavlja klasični pristup modeliranju ekstremnih vrijednosti, a temelji se na analizi asimptotskoga ponašanja niza nezavisnih i jednako distribuiranih slučajnih varijabli. Glavni je cilj pružiti teorijsku osnovu za razumijevanje metode maksimuma blokova te ilustrirati njezine primjene na podacima iz stvarnoga svijeta. Pristup razvijanju teorije potrebne za modeliranje ekstremnih vrijednosti te ilustracija njezine primjene na stvarnim podacima, koji su izloženi u ovome radu, najvećim se dijelom temelje na Colesovoj knjizi *An Introduction to Statistical Modeling of Extreme Values*, objavljenoj 2001. godine [1].
Poglavlje 1 bavi se modeliranjem niza maksimuma blokova iz dostupnih podataka te proučavanjem njegova asimptotskog ponašanja. Središnji je dio ovoga poglavlja, ali i cijeloga rada, Fisher–Tippett–Gnedenkov teorem koji, u slučaju da promatrani niz maksimuma blokova konvergira po distribuciji, navodi sve funkcije distribucije koje bi mogle biti limes. Definiraju se pojmovi poput povratnoga perioda i razine povrata te opisuje način prilagodbe

pravoga modela podacima, što je vrlo važno u praktičnoj primjeni.
Poglavlje 2 donosi primjenu metode maksimuma blokova na primjerima iz stvarnoga svijeta: u proučavanju maksimalne godišnje razine mora na obalama Port Pirie, u proučavanju izdržljivosti staklenih vlakana te u analizi maksimalnih ljetnih temperatura zraka u gradu Zagrebu. Na kraju poglavlja nalazi se kratak osvrt na eventualne nedostatke metode maksimuma blokova te pregled alternativnih pristupa modeliranju ekstremnih vrijednosti.
Naposljetku, Dodatak sadrži kratak pregled pojmova i rezultata klasične matematičke statistike eksplicitno ili implicitno iskorištenih u ovome radu.

Poglavlje 1

Metoda maksimuma blokova

Pretpostavimo da je u svrhu izgradnje riječnoga nasipa potrebno odrediti visinu vodostaja koju rijeka neće prijeći u idućih sto godina s dovoljno velikom vjerojatnosti, pri čemu su dostupni podaci o visinama vodostaja te rijeke u proteklih deset godina. Budući da ne postoje empirijske ili fizikalne smjernice za ekstrapolaciju tražene vrijednosti, ima smisla pokušati upotrijebiti rezultate teorije ekstremnih vrijednosti. Klasičan pristup rješavanju ovakva problema upravo je metoda maksimuma blokova.

1.1 Oblikovanje i asimptotsko ponašanje modela

Neka je $(X_n)_{n\in\mathbb{N}}$ niz nezavisnih jednako distribuiranih slučajnih varijabli sa zajedničkom funkcijom distribucije F. Metoda maksimuma blokova proučava statističko ponašanje slučajne varijable

$$M_n = \max\{X_1, X_2, \ldots, X_n\},\ \forall n \in \mathbb{N}.$$

U primjenama dostupni podaci, npr. mjerenja visina vodostaja rijeke, predstavljaju realizacije slučajnih varijabli X_n, a M_n maksimalnu izmjerenu vrijednost promatrane pojave tijekom n vremenskih jedinica. Rezultati predstavljeni u okviru ovoga rada vezani su uz asimptotsko ponašanje niza maksimuma, no svi su primjenjivi i na proučavanje niza minimuma uz relaciju $m_n = \min\{X_1, X_2, \ldots, X_n\} = -\max\{-X_1, -X_2, \ldots, -X_n\}$.
Dakle, u svrhu donošenja zaključaka o budućim maksimalnim vrijednostima visine riječnoga vodostaja potrebno je odrediti ponašanje, odnosno distribuciju slučajne varijable M_n. U teoriji je za svaki $n \in \mathbb{N}$ moguće pronaći distribuciju od M_n koristeći nezavisnost i jednaku distribuiranost niza $(X_n)_n$. Naime, vrijedi:

$$\begin{aligned}\mathbb{P}(M_n \leq x) &= \mathbb{P}(\max\{X_1, X_2, \ldots, X_n\} \leq x) \\ &= \mathbb{P}(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x)\mathbb{P}(X_2 \leq x) \cdots \mathbb{P}(X_n \leq x) \\ &= F(x)^n,\end{aligned}$$

pri čemu je $x \in \mathbb{R}$ i treća jednakost slijedi iz nezavisnosti, a posljednja iz jednake distribuiranosti slučajnih varijabli $(X_n)_n$.
Međutim, u radu sa stvarnim podacima ovakav teorijski rezultat nije od prevelike koristi jer je distribucija samih podataka najčešće nepoznata.
Jedan mogući pristup određivanju distribucije od M_n procijeniti je F pomoću empirijske funkcije distribucije dostupnih podataka. No, vrlo mala odstupanja procijenjene funkcije distribucije od F mogu dovesti do značajnih odstupanja pri određivanju vrijednosti F^n pa ovakav pristup zahtijeva veliki oprez.
Drugi je pristup jednostavno prihvatiti da je F nepoznata te pokušati pronaći familiju modela koja će dovoljno dobro aproksimirati F^n, a koju je moguće odrediti samo pomoću već zabilježenih maksimuma blokova. Ovakav pristup predstavlja temelj same metode maksimuma blokova te središnju temu ovoga poglavlja.
Međutim, poteškoća koja se javlja kod drugoga pristupa nemogućnost je promatranja distribucije samih vrijednosti M_n. Naime, za sve $x \in \mathbb{R}$ takve da vrijedi $x < x_+$, pri čemu x_+ označava najmanju realnu vrijednost za koju je $F(x) = 1$, vrijedit će $F(x)^n \to 0$ kada $n \to \infty$, odnosno sva masa distribucije od M_n bit će koncentrirana samo u jednoj točki x_+.
Zato se umjesto distribucije od M_n promatra distribucija njezine transformacije

$$M_n^* = \frac{M_n - b_n}{a_n}, \tag{1.1}$$

gdje su $(a_n)_n$, $a_n > 0$ i $(b_n)_n$ nizovi realnih konstanti koji normaliziraju M_n^*.
Naravno, takvi normirajući nizovi konstanti ne moraju nužno postojati, no ako postoje, Fisher–Tippett–Gnedenkov teorem o distribuciji ekstremnih vrijednosti (engl. *extremal types theorem*), daje sve moguće distribucije za M_n^*.

Teorem 1.1.1 (Fisher–Tippett–Gnedenko)**.** *Neka su $(a_n)_n$ i $(b_n)_n$ nizovi realnih konstanti takvi da vrijedi $a_n > 0$ za sve $n \in \mathbb{N}$ i*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \longrightarrow G(x) \quad \textit{kada} \quad n \to \infty, \, \forall x \in C(G), \tag{1.2}$$

pri čemu je G nedegenerirana funkcija distribucije, a C(G) skup točaka neprekidnosti od G. Tada G pripada jednoj od sljedećih familija distribucija:

- $$G(x) = \exp\left\{ -\exp\left[-\frac{x-b}{a} \right] \right\}, \quad -\infty < x < \infty, \tag{1.3}$$
- $$G(x) = \begin{cases} 0, & x \le b, \\ \exp\left\{ -\left(\frac{x-b}{a} \right)^{-\alpha} \right\}, & x > b, \end{cases} \tag{1.4}$$
- $$G(x) = \begin{cases} \exp\left\{ -\left(-\frac{x-b}{a} \right)^{\alpha} \right\}, & x < b, \\ 1, & x \ge b, \end{cases} \tag{1.5}$$

pri čemu su a, b, α realni parametri takvi da vrijedi a > 0 i α > 0.

Familije distribucija (1.3), (1.4), (1.5) poznate su kao **Gumbelova**, **Fréchetova** i **Weibullova** familija, respektivno.

Napomena 1.1.2. *Budući da su sve funkcije G koje se mogu pojaviti kao limes neprekidne, vrijedi C(G) = ℝ, odnosno konvergencija vrijedi za sve realne vrijednosti x.*

Jednostavnije rečeno, teorem 1.1.1 tvrdi da kada se M_n može stabilizirati prikladnim nizovima $(a_n)_n$ i $(b_n)_n$ tako da normalizirana varijabla M_n^* konvergira po distribuciji, tada je limes nužno funkcija distribucije iz Gumbelove, Frechétove ili Weibullove familije.
Važnost Fisher–Tippett–Gnedenkova teorema ogleda se upravo u tome što bez obzira na početnu distribuciju populacije *F*, postoje samo tri moguća limesa, odnosno familije distribucija, prema kojima normalizirani maksimumi blokova M_n^* mogu konvergirati. Zbog toga se teorem 1.1.1 često smatra analogonom centralnoga graničnog teorema za ekstremne vrijednosti.

1.1.1 Dokaz Fisher–Tippett–Gnedenkova teorema

Teorem 1.1.1 prvi su otkrili Fisher i Tippett 1928., a u potpunosti ga je dokazao Gnedenko 1943. godine. Dokaz teorema predstavljen u ovome radu preuzet je iz [4], a zasniva se na de Haanovu pristupu iz 1976. Alternativni dokazi, s manje tehničkih detalja, mogu se pronaći u [1] te [2].
Ideja samoga dokaza poistovjetiti je klase distribucija ekstremnih vrijednosti, odnosno funkcije iz Gumbelove, Fréchetove ili Weibullove familije s klasama max-stabilnih funkcija, zbog čijih će svojstava onda slijediti tvrdnja teorema. Ostatak ovoga potpoglavlja bavi se teoremima i pomoćnim tvrdnjama potrebnima za dokaz teorema 1.1.1, kojim ovo potpoglavlje i završava. Dokazi navedenih tvrdnji pretežito su tehničke prirode i nisu osobito komplicirani, zbog čega nisu navedeni u okvirima ovoga rada te se mogu pronaći u [4].

Najprije definiramo inverz monotone funkcije te navodimo općenita svojstva inverza koja će biti ključna u dokazu teorema 1.1.1.

Definicija 1.1.3. *Ako je* $\psi\colon \mathbb{R} \to \mathbb{R}$ *neopadajuća i zdesna neprekidna funkcija, tada je njoj inverzna funkcija* ψ^{-1} *definirana na intervalu* $(\inf_{x\in\mathbb{R}}\{\psi(x)\}, \sup_{x\in\mathbb{R}}\{\psi(x)\})$ *kao*

$$\psi^{-1}(y) = \inf\{x \in \mathbb{R}\colon \psi(x) \geq y\}.$$

Lema 1.1.4. *i) Neka je* $\psi\colon \mathbb{R} \to \mathbb{R}$ *neopadajuća zdesna neprekidna funkcija te neka je* $H\colon \mathbb{R} \to \mathbb{R}$ *definirana s* $H(x) = \psi(ax+b)-c$ *za neke realne konstante* $a > 0$, b *i* c. *Tada je i* H *neopadajuća zdesna neprekidna funkcija te vrijedi* $H^{-1}(y) = a^{-1}(\psi^{-1}(y+c)-b)$.

ii) Neka je ψ^{-1} *inverz funkcije* ψ. *Ako je* ψ^{-1} *neprekidna funkcija, tada je* $\psi^{-1}(\psi(x)) = x, \forall x \in \mathbb{R}$.

iii) Neka je G *nedegenerirana funkcija distribucije. Tada postoje* $y_1, y_2 \in \mathbb{R}$, $y_1 < y_2$, *takvi da su* $G^{-1}(y_1)$ *i* $G^{-1}(y_2)$ *dobro definirane konačne vrijednosti za koje vrijedi* $G^{-1}(y_1) < G^{-1}(y_2)$.

Korolar 1.1.5. *Neka je* G *nedegenerirana funkcija distribucije te neka su* $a > 0$, $\alpha > 0$, b *i* β *realne konstante takve da vrijedi* $G(ax+b) = G(\alpha x+\beta)$, $\forall x \in \mathbb{R}$. *Tada je* $a = \alpha$ *i* $b = \beta$.

Teorem 1.1.6 (Hinčin)**.** *Neka je* $(F_n)_n$ *niz funkcija distribucije te* G *nedegenerirana funkcija distribucije. Neka su* $a_n > 0$ *i* b_n *realne konstante takve da vrijedi*

$$F_n(a_n x + b_n) \longrightarrow G(x), \quad n \to \infty, \ \forall x \in C(G),$$

pri čemu $C(G)$ *označava skup točaka neprekidnosti od* G. *Tada za neku nedegeneriranu funkciju distribucije* G^* *i konstante* $\alpha_n > 0$, $\beta_n \in \mathbb{R}$

$$F_n(\alpha_n x + \beta_n) \longrightarrow G^*(x), \quad n \to \infty, \ \forall x \in C(G^*)$$

ako i samo ako

$$a_n^{-1}\alpha_n \longrightarrow a \quad i \quad a_n^{-1}(\beta_n - b_n) \longrightarrow b$$

za neke realne $a > 0$ *i* b *te tada vrijedi*

$$G^*(x) = G(ax + b), \quad \forall x \in \mathbb{R}.$$

Hinčinov teorem o konvergenciji funkcija distribucije ključan je za podjelu funkcija distribucije u klase ekvivalencije.

Definicija 1.1.7. *Nedegenerirana funkcija distribucije* G *je* ***max-stabilna*** *ako za svaki* $n = 2, 3, \ldots$ *postoje konstante* $a_n > 0$, $b_n \in \mathbb{R}$ *takve da vrijedi* $G^n(a_n x + b_n) = G(x)$, $\forall x \in \mathbb{R}$.

Napomena 1.1.8. *U okvirima ovoga rada vrlo je korisna i vjerojatnosna interpretacija max-stabilnih funkcija distribucije. Naime, ako su* $(X_n)_n$ *nezavisne i jednako distribuirane slučajne varijable sa zajedničkom funkcijom distribucije* G*, tada je* G *max-stabilna ako i samo ako* $\forall n \in \mathbb{N}$ *postoje odgovarajuće realne konstante* $a_n > 0$*,* b_n *tako da vrijedi* $\frac{M_n - b_n}{a_n} \stackrel{d}{=} X_1 \sim G$*, odnosno funkcija distribucije od* $\frac{M_n - b_n}{a_n}$ *također je* G*.*

Za dvije funkcije distribucije G_1 i G_2 kaže se da su **istoga tipa** ako je $G_2(x) = G_1(ax + b)$, za sve $x \in \mathbb{R}$ te za neke realne konstante $a > 0$ i b. Alternativno, ako su X i Y slučajne varijable takve da vrijedi $X \sim G_1, Y \sim G_2$, onda su G_1 i G_2 istoga tipa ako postoje $a, b \in \mathbb{R}$ takvi da vrijedi $Y \stackrel{d}{=} \frac{X-b}{a}$. Sada se može reći da je nedegenerirana funkcija distribucije G max-stabilna ako je za svaki n= $2, 3, \ldots$ funkcija distribucije G^n istoga tipa kao kao G.
Također, iz teorema 1.1.6 slijedi da ako vrijedi $F_n(a_n x + b_n) \to G_1, \forall x \in C(G_1)$ i $F_n(\alpha_n x + \beta_n) \to G_2, \forall x \in C(G_2)$ za niz funkcija distribucije $(F_n)_n$ te realne konstante $a_n > 0$, $\alpha_n > 0$, b_n i β_n, tada su funkcije G_1 i G_2 nužno istoga tipa, pod uvjetom da su G_1 i G_2 nedegenerirane. Sada je očito kako se funkcije distribucije mogu podijeliti u klase ekvivalencije na način da svaka klasa sadrži funkcije za koje se može reći da su istog tipa. Tako definirane klase ekvivalencije nazivaju se **tipovima** pa se u literaturi za funkcije distribucije koje su istoga tipa kao funkcije iz Gumbelove familije često kaže da su **tipa I**. Fréchetova je familija **tip II**, a Weibullova **tip III**.
Koristeći definiciju slučajne varijable M_n, izraz (1.2) moguće je zapisati kao

$$F^n(a_n x + b_n) \longrightarrow G(x), \ \forall x \in C(G). \tag{1.6}$$

Definicija 1.1.9. *Ako postoje realni nizovi* $(a_n)_n$*,* $a_n > 0$ *i* $(b_n)_n$ *takvi da vrijedi* (1.6)*, tada* F *pripada* ***(maksimalnoj) domeni atrakcije*** *od* G *i pišemo* $F \in MDA(G)$*.*

Domene atrakcije vrlo su zanimljivo područje teorije ekstremnih vrijednosti o kojem će više govora biti u poglavlju 1.2.

Teorem 1.1.10. *i) Nedegenerirana funkcija distribucije* G *je max-stabilna ako i samo ako postoji niz funkcija distribucije* $(F_n)_n$ *te nizovi realnih konstanti* $a_n > 0$ *i* b_n *tako da vrijedi*

$$F_n(a_{nk}^{-1} x + b_{nk}) \longrightarrow G^{1/k}(x), \quad n \to \infty, \ \forall x \in C(G^{1/k}), \tag{1.7}$$

za svaki k=$1, 2, \ldots$.

ii) Posebno, ako je G *nedegenerirana funkcija distribucije, tada je skup* $MDA(G)$ *neprazan ako i samo ako je* G *max-stabilna. Dakle, klasa nedegeneriranih funkcija distribucije koje se pojavljuju kao limes u* (1.2) *sadrži jednake elemente kao klasa max-stabilnih funkcija distribucije.*

Korolar 1.1.11. *Neka je je G max-stabilna funkcija distribucije. Tada postoje realne funkcije $a(s) > 0$ i $b(s)$ definirane za $s > 0$ takve da vrijedi*

$$G^s(a(s)x + b(s)) = G(x), \quad \forall x \in \mathbb{R},\ s > 0. \tag{1.8}$$

U svrhu dokazivanja Fisher–Tippett–Gnedenkova teorema valja pokazati kako je funkcija distribucije max-stabilna ako i samo ako je istoga tipa kao jedna od tri funkcije distribucije[1] navedene u teoremu 1.1.1.

Teorem 1.1.12. *Svaka max-stabilna funkcija distribucije F istog je tipa kao jedna od distribucija ekstremnih vrijednosti iz teorema 1.1.1, odnosno vrijedi $F(x) = G(ax+b), \forall x \in \mathbb{R}$ te za realne konstante $a > 0$ i b gdje je*

- $G(x) = \exp\left(-e^{-x}\right) \quad -\infty < x < \infty,$ (1.9)
- $G(x) = \begin{cases} 0, & x \le 0, \\ \exp(-x^{-\alpha}), & x > 0, \end{cases}$ (1.10)
- $G(x) = \begin{cases} \exp(-(-x)^{\alpha}) & x < 0, \\ 1, & x \ge 0, \end{cases}$ (1.11)

pri čemu je $\alpha > 0$. Obratno, svaka je distribucija ekstremnih vrijednosti max-stabilna.

Dokaz. Neka je G max-stabilna funkcija. Tada iz korolara 1.8 slijedi

$$G^s(a(s)x + b(s)) = G(x), \tag{1.12}$$

za sve realne x i $s > 0$. Pretpostavimo još da vrijedi $0 < G(x) < 1, \forall x \in \mathbb{R}$. Dva puta logaritmiramo izraz (1.12) i dobivamo

$$-\log(-\log(G(a(s)x + b(s)))) - \log s = -\log(-\log(G(x))).$$

Koristeći svojstvo max-stabilnosti za $n = 2$ dobivamo da vrijedi $G^2(ax+b) = G(x)$ za $a > 0$ i b realne, iz čega slijedi da G nema prekid ni u jednoj od konačnih *završnih točaka* x_- i x_+.[2]
Neka je $\psi(x) = -\log(-\log(G(x)))$. Tada je ψ neopadajuća funkcija i vrijedi $\inf_{x\in\mathbb{R}} \psi(x) = -\infty$, $\sup_{x\in\mathbb{R}} \psi(x) = +\infty$ pa ψ ima inverznu funkciju $U(y)$ definiranu za sve realne y. Također vrijedi

$$\psi(a(s)x + b(s)) - \log s = \psi(x),$$

[1] Precizno govoreći, u teoremu 1.1.1 navedene su **familije tipova** funkcija distribucija.

[2] Općenito, završne točke proizvoljne funkcije distribucije G definiraju se kao: $x_- = \sup\{x \in \mathbb{R}\colon G(x) = 0\}$ te $x_+ = \inf\{x \in \mathbb{R}\colon G(x) = 1\}$.

pa po prvom dijelu leme 1.1.4 slijedi

$$\frac{U(y + \log s) - b(s)}{a(s)} = U(y). \tag{1.13}$$

Posebno, za $y = 0$ vrijedi

$$\frac{U(\log s) - b(s)}{a(s)} = U(0). \tag{1.14}$$

Oduzimanjem izraza (1.14) od (1.13) slijedi

$$\frac{U(y + \log s) - U(\log\ s)}{a(s)} = U(y) - U(0), \tag{1.15}$$

odnosno uz $z = \log s$, $\tilde{a}(z) = a(e^z)$, $\tilde{U}(y) = U(y) - U(0)$

$$\tilde{U}(y + z) - \tilde{U}(y) = \tilde{U}(y)\tilde{a}(z), \tag{1.16}$$

za sve realne y i z. Zamjenom y i z te nakon oduzimanja slijedi

$$\tilde{U}(y)(1 - \tilde{a}(z)) = \tilde{U}(z)(1 - \tilde{a}(y)). \tag{1.17}$$

Sada su moguća su dva slučaja.

i) Ako je $\tilde{a}(z) = 1$ za svaki $z \in \mathbb{R}$, iz (1.16) slijedi

$$\tilde{U}(y + z) = \tilde{U}(y) + \tilde{U}(z).$$

Jedina monotona neopadajuća funkcija koja zadovoljava ovakav uvjet jest $\tilde{U}(y) = \rho y$ za neki $\rho > 0$, odnosno $U(y) - U(0) = \rho y$. Tada za inverz funkcije ψ vrijedi

$$\psi^{-1}(y) = U(y) = \rho y + v, \quad v = U(0).$$

Sada iz druge tvrdnje leme 1.1.4 slijedi

$$x = \psi^{-1}(\psi(x)) = \rho\psi(x) + v,$$

odnosno $\psi(x) = (x - v)/\rho$. Naposljetku, koristeći definiciju funkcije ψ, dobivamo

$$G(x) = \exp\Big(- e^{-(x-v)/\rho}\Big) \tag{1.18}$$

za $0 < G(x) < 1$. Budući da G ne može imati prekide u x_- niti x_+, slijedi kako za svaki $x \in \mathbb{R}$ vrijedi (1.18) pa zaključujemo da je G istoga tipa kao (1.9).

ii) Neka sada postoji $z \in \mathbb{R}$ takav da je $\tilde{a}(z) \neq 1$. Tada iz (1.17) slijedi

$$\tilde{U}(y) = \frac{\tilde{U}(z)}{1-\tilde{a}(z)}(1-\tilde{a}(y)) = c(1-\tilde{a}(y)), \quad c = \frac{\tilde{U}(z)}{1-\tilde{a}(z)}. \tag{1.19}$$

Vrijedi $c \neq 0$. Naime, u suprotnom bi vrijedilo $\tilde{U}(z) = 0$, a onda i $\tilde{U}(y) = 0$ što bi povlačilo da je $U(y) = U(0)$ konstantna funkcija. Primjenom (1.19) na (1.16) dobivamo

$$c(1-\tilde{a}(y+z)) - c(1-\tilde{a}(z)) = c(1-\tilde{a}(y))\tilde{a}(z) \tag{1.20}$$

pa vrijedi $\tilde{a}(y+z) = \tilde{a}(y)\tilde{a}(z)$. Budući da je $\tilde{a}$ monotona funkcija, jedino je moguće rješenje funkcija oblika $\tilde{a}(y) = e^{\rho y}$ za $\rho \in \mathbb{R}$, $\rho \neq 0$. Sada iz (1.19) slijedi

$$\psi^{-1}(y) = U(y) = v + c(1-e^{\rho y}), \quad v = U(0).$$

Kako je U rastuća funkcija, mora vrijediti $c < 0$ ako je $\rho > 0$ i $c > 0$ ako je $\rho < 0$. Druga tvrdnja leme 1.1.4 povlači

$$x = \psi^{-1}(\psi(x)) = v + c\Big(1 - e^{\rho\psi(x)}\Big) = v + c(1 - (-\log(G(x)))^{-\rho}.$$

Sada za $0 < G(x) < 1$ vrijedi

$$G(x) = \exp\left\{ -\left(1 - \frac{x-v}{c}\right)^{-1/\rho} \right\}. \tag{1.21}$$

Budući da je G neprekidna u svim konačnim x_- i x_+, iz (1.21) slijedi da je G istoga tipa kao (1.10) ili (1.11) $\forall x \in \mathbb{R}$, uz $\alpha = \pm 1/\rho$ ovisno o tomu je li $\rho > 0$ ($c < 0$) ili $\rho < 0$ ($c > 0$).

Preostaje pokazati obratnu tvrdnju, odnosno da je svaka od distribucija ekstremnih vrijednosti max-stabilna.

i) Neka je $G(x) = \exp(-e^{-x})$, $\forall x \in \mathbb{R}$. Tada vrijedi

$$G^n(a_n x + b_n) = \exp\Big(- ne^{-(a_n x + b_n)} \Big) = \exp\Big(- e^{-(a_n x + b_n) + \log n} \Big).$$

Neka je $a_n = 1$ i $b_n = \log n$ za svaki $n \in \mathbb{N}$. Tada očito vrijedi $G^n(a_n x + b_n) = G(x)$, odnosno funkcije distribucije Gumbelova tipa su max-stabilne.

ii) Neka je sada G iz Fréchetove familije distribucija i to

$$G(x) = \begin{cases} 0, & x \leq 0, \\ \exp(-x^{-\alpha}), & x > 0. \end{cases}$$

Vrijedi

$$G^n(a_n x + b_n) = \begin{cases} 0, & x \leq 0, \\ \exp((-n(a_n x + b_n))^{-\alpha}), & x > 0, \end{cases}$$

pa uz $a_n = n^{1/\alpha}$ te $b_n = 0$ slijedi svojstvo max-stabilnosti.

iii) Naposljetku, neka je G iz Weibullove familije, odnosno

$$G(x) = \begin{cases} \exp(-(-x)^{\alpha}), & x \leq 0, \\ 1, & x > 0. \end{cases}$$

Kao i ranije, iz

$$G^n(a_n x + b_n) = \begin{cases} \exp((-n(-a_n x - b_n)^{\alpha})), & x \leq 0, \\ 1, & x > 0, \end{cases}$$

uz $a_n = n^{-1/\alpha}$ i $b_n = 0$ slijedi da je G max-stabilna. □

Sada tvrdnja Fisher–Tippett–Gnedenkova teorema jednostavno slijedi iz rezultata navedenih u ovom potpoglavlju.

Dokaz teorema 1.1.1. Pretpostavimo da (1.2) vrijedi za neku nedegeneriranu funkciju distribucije G. Tada po prvome dijelu teorema 1.1.10 slijedi da je G max-stabilna, a samim time, prema teoremu 1.1.12 slijedi da je G istoga tipa kao jedna od distribucija iz Gumbelove, Fréchetove ili Weibullove familije pa vrijedi tvrdnja Fisher–Tippett–Gnedenkova teorema. □

1.2 Domene atrakcije

Razvoj metode maksimuma blokova najvećim je dijelom bio motiviran potrebom pronalaska distribucije niza maksimuma izvedenih iz niza nezavisnih jednako distribuiranih slučajnih varijabli čija je distribucija nepoznata. Međutim, vrlo se zanimljivim pokazalo proučavati i slučaj kada je distribucija dostupnih podataka poznata. Drugim riječima, u sklopu teorije ekstremnih vrijednosti razvila se bogata teorija koja se bavi određivanjem kojem tipu distribucije ekstremnih vrijednosti pripada niz maksimuma iz neke poznate distribucije. U tu svrhu definicijom 1.2 uveden je pojam domene atrakcije te je cilj ovoga odjeljka pobliže proučavanje domena atrakcije funkcija iz Gumbelove, Fréchetove i Weibullove familije. Preciznije, glavno je pitanje koje uvjete mora zadovoljavati funkcija distribucije F kako bi za danu funkciju distribucije G (tipa I, II ili III) vrijedilo

$$F^n(a_n x + b_n) \longrightarrow G(x), \ \forall x \in C(G) \tag{1.22}$$

Također, bitna su pitanja kako odabrati prave normirajuće konstante a_n i b_n za koje vrijedi (1.22) te dobivaju li se odabirom različitih normirajućih konstanti različite funkcije G kao limes. Na posljednje pitanje odgovara teorem 1.1.6, iz kojega slijedi da ako niz funkcija distribucije konvergira, limesi su uvijek funkcije istoga tipa.
Općenito, svaki od tipova I, II ili III ekstremnih distribucija odgovara specifičnome ponašanju maksimuma te se u praktičnoj primjeni koriste za modeliranje jako različitih pojava. Odabir pravoga tipa distribucije za maksimum najvećim dijelom ovisi o *repu*[3] početne distribucije, koji je glavni alat u proučavanju domena atrakcije. Tri tipa distribucija ekstremnih vrijednosti najviše se razlikuju upravo po ponašanju svojih repova pa tako, primjerice, funkcije iz Fréchetove domene u pravilu imaju najteže repove te se koriste za modeliranje rizičnih događaja. Funkcije iz Weibullove domene ograničene su odozgo. Gumbelova domena atrakcije obuhvaća najširi spektar funkcija distribucije pa se tako u njoj mogu pronaći funkcije s međusobno vrlo različitim ponašanjima repova.

1.2.1 Fréchetova domena atrakcije

Neka je G_F funkcija distribucije iz Fréchetove familije oblika (1.10). Taylorovim se razvojem dobije

$$1 - G_F(x) = 1 - \exp\big(-x^{-\alpha}\big) \sim x^{-\alpha}, \quad x \to \infty, \quad \alpha > 0,$$

iz čega slijedi da rep Fréchetove distribucije opada kao polinom. Pokazat će se kako Fréchetovoj domeni atrakcije pripadaju funkcije distribucije čiji je desni rep regularno varirajući s indeksom $-\alpha$. Pojam regularno varirajućih funkcija preciziran je definicijom 1.2.1.

[3] Rep funkcije distribucije F definira se kao $\overline{F}(x) = 1 - F(x), \forall x \in \mathbb{R}$.

Definicija 1.2.1. *Izmjeriva je funkcija* $f : [0, +\infty) \to [0, +\infty)$ ***regularno varirajuća*** *s indeksom* $\alpha \in \mathbb{R}$ *i pišemo* $f \in \mathcal{R}_\alpha$ *ako za svaki* $t > 0$ *vrijedi*

$$\lim_{x\to\infty} \frac{f(xt)}{f(t)} = t^\alpha.$$

Posebno, ako je $\alpha = 0$, *vrijedi*

$$\lim_{x\to\infty} \frac{f(xt)}{f(t)} = 1$$

i kažemo da je f ***sporo varirajuća*** *u oznaci* $f \in \mathcal{R}_0$.

Fréchetova klasa funkcija sadrži distribucije teškoga repa (engl. *heavy-tailed distributions*) kojima je desna završna točka $x_+ = +\infty$. U primjeni su takve funkcije često pogodne za modeliranje velikih isplata osiguranja, log-povrata, velikih varijacija u cijenama i slično. U ovom odjeljku navedeni su primjeri funkcija distribucije iz Fréchetove domene atrakcije. Nužni i dovoljni uvjeti za pripadnost Fréchetovoj domeni, zajedno s načinom odabira odgovarajućih konstanti a_n i b_n dani su teoremom 3.1.17.

Primjer 1.2.2 (Paretova distribucija)**.** *Neka je* F *funkcija distribucije Paretove slučajne varijable, odnosno* $F(x) = 1 - \kappa x^{-\alpha}$, $\alpha > 0$, $\kappa > 0$, $x \geq \kappa^{1/\alpha}$. *Za rep distribucije* $\overline{F}$ *očito vrijedi*

$$\overline{F}(x) \sim \kappa x^{-\alpha}, \quad x \to \infty,$$

odnosno $\overline{F}$ *je regularno varirajući s indeksom* $-\alpha$, *u oznaci* $\overline{F} \in \mathcal{R}_{-\alpha}$, *pa po teoremu 3.1.17 slijedi* $a_n = (\kappa n)^{1/\alpha}$, $b_n = 0$ *te*

$$\frac{M_n}{(\kappa n)^{1/\alpha}} \xrightarrow{d} G_F,$$

odnosno $(\kappa n)^{-1/\alpha} M_n$ *konvergira po distribuciji prema nekoj slučajnoj varijabli* Y, *u oznaci* $(\kappa n)^{-1/\alpha} M_n \xrightarrow{d} Y$, *čija je funkcija distribucije* G_F. *Konvergencija po distribuciji definirana je definicijom 3.1.2. Alternativno, pripadnost Paretove distribucije Fréchetovoj domeni atrakcije može se pokazati i direktno. Vrijedi*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n).$$

Neka je $a_n = (\kappa n)^{1/\alpha}$ *i* $b_n = 0$. *Tada* $\forall x > 0$ *kada* $n \to \infty$ *vrijedi*

$$\begin{aligned} F^n(a_n x + b_n) &= \left[1 - \kappa((\kappa n)^{1/\alpha} x)^{-\alpha}\right]^n \\ &= \left[1 - \frac{x^{-\alpha}}{n}\right]^n \\ &\to e^{-x^{-\alpha}}. \end{aligned}$$

Inače, za $x \leq 0$ *vrijedi* $F^n((\kappa n)^{1/\alpha}x) = 0, \forall n \in \mathbb{N}$ *pa Paretova distribucija pripada Fréchetovoj domeni atrakcije.*

Uz pomoć Paretove distribucije često se modeliraju fenomeni poput raspodjele bogatstva među stanovnicima neke države kao i distribucija stanovništva po urbanim i ruralnim sredinama.

Primjer 1.2.3 (Cauchyjeva distribucija)**.** *Neka je F funkcija distribucije standardne Cauchyjeve slučajne varijable. Tada je*

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x),\ \forall x \in \mathbb{R}$$

te za rep distribucije vrijedi $\overline{F}(x) \sim (\pi x)^{-1}$*. Normirajuće su konstante* $a_n = n\pi^{-1}$ *i* $b_n = 0$ *pa, ponovno po teoremu 3.1.17, vrijedi*

$$\frac{\pi M_n}{n} \xrightarrow{d} G_F.$$

Primjer 1.2.4 (Log-gamma distribucija)**.** *Slučajna varijabla X pripada log-gamma distribuciji ako njezin prirodni logaritam pripada gamma distribuciji. Funkcija gustoće takve slučajne varijable s parametrima* $\alpha > 0, \beta > 0$ *jednaka je*

$$f(x) = \frac{1}{\alpha^\beta \Gamma(\beta)} e^{\beta x} e^{e^{-x}\alpha^{-1}}$$

Pokaže se kako za rep log-gamma distribucije vrijedi

$$\overline{F}(x) \sim \frac{\alpha^{\beta-1}}{\Gamma(\beta)}(\log(x))^{\beta-1}x^{-\alpha}, \quad \alpha, \beta > 0,\ x \to \infty.$$

Dakle, $\overline{F} \in \mathcal{R}_{-\alpha}$ *pa log-gamma distribucija pripada Fréchetovoj domeni atrakcije. Vrijedi* $b_n = 0$*, a, uz nešto kompliciraniji račun nego u prethodnim primjerima, pronađe se*

$$a_n \sim ((\Gamma(\beta)^{-1}(\log n)^{\beta-1}n)^{1/n}$$

te

$$((\Gamma(\beta)^{-1}(\log n)^{\beta-1}n)^{-1/n}M_n \xrightarrow{d} G_F.$$

Log-gamma distribucija često se koristi za modeliranje velikih svota koje će osiguravateljske kuće trebati isplaćivati klijentima.

1.2.2 Weibullova domena atrakcije

Neka G_W označava funkciju distribucije iz Weibullove familije oblika (1.11). Zajednička karakteristika svih funkcija distribucije iz Weibullove domene atrakcije postojanje je konačne desne završne točke x_+ pa tako ova klasa sadrži odozgo ograničene funkcije distribucije, poput, primjerice, uniformne. Također, očito je da vrijedi

$$G_W(-x^{-1}) = G_F(x), \quad x > 0,$$

pa su Fréchetova i Weibullova familija usko povezane, kao i njihove domene atrakcije, što potvrđuje teorem 3.1.18, koji govori o nužnim i dovoljnim uvjetima za pripadnost funkcije Weibullovoj domeni atrakcije.

Primjer 1.2.5 (Uniformna distribucija na (0,1))**.** *Neka je F funkcija distribucije uniformne slučajne varijable na intervalu (0,1), odnosno* $F(x) = x$, $0 \leq x \leq 1$. *Očito je* $x_+ = 1$. *Vrijedi*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n).$$

Neka je $a_n = n^{-1}$ *i* $b_n = 1$. *Tada* $\forall x < 0$ *kada* $n \to \infty$ *vrijedi*

$$\begin{aligned} F^n(a_n x + b_n) &= \left[1 + \frac{x}{n}\right]^n \\ &\to e^x. \end{aligned}$$

Za $x \geq 0$ *vrijedi* $F^n(1 + (x/n)) = 1$, $\forall n \in \mathbb{N}$. *Dakle, uniformna razdioba na intervalu* $(0, 1)$ *pripada Weibullovoj domeni atrakcije.*
Alternativno, vrijedi $\overline{F}(1 - x^{-1}) = x^{-1}$ *pa po teoremu 3.1.18 slijedi* $a_n = n^{-1}$, $b_n = 1$, *te*

$$n(M_n - 1) \xrightarrow{d} G_W$$

Primjer 1.2.6 (Beta distribucija)**.** *Neka je X funkcija distribucije iz beta razdiobe s realnim i pozitivnim parametrima a i b te gustoćom*

$$f(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad 0 < x < 1, \quad a, b > 0.$$

Može se pokazati kako za rep takve distribucije vrijedi da je funkcija $\overline{F}(1 - x^{-1})$ *regularno varirajuća s indeksom* $-b$ *te*

$$\overline{F}(x) \sim \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b + 1)}(1 - x)^b, \quad x \uparrow 1.$$

Sada, uz

$$a_n = \left(\frac{n\Gamma(a+b)}{\Gamma(a)\Gamma(b+1)}\right)^{-1/\alpha} \quad te \quad b_n = 1$$

vrijedi

$$\left(\frac{n\Gamma(a+b)}{\Gamma(a)\Gamma(b+1)}\right)^{1/\alpha}(M_n - 1) \xrightarrow{d} G_W.$$

Često se kaže kako je beta distribucija *distribucija vjerojatnosti događaja* jer se koristi, primjerice, za modeliranje vjerojatnosti preživljavanja oboljelih od karcinoma, procjenu proporcije korisnika koji će reagirati na određenu marketinšku kampanju, procjenu vjerojatnosti pobjede političara na budućim izborima i slično.

1.2.3 Gumbelova domena atrakcije

Neka G_G označava funkciju distribucije iz Gumbelove razdiobe oblika (1.9). Taylorovim razvojem slijedi

$$1 - G_G(x) \sim e^{-x}, \quad x \to \infty,$$

pa rep Gumbelove distribucije opada eksponencijalno. Funkcije distribucije iz Gumbelove domene atrakcije mogu biti jako različite po ponašanju repa. Primjerice, lognormalna distribucija, za koju se može reći kako je umjereno teškoga repa (engl. *moderately heavy-tailed*), nalazi se u Gumbelovoj domeni atrakcije jednako kao i normalna distribucija koja je distribucija lakoga repa (engl. *thin-tailed*). Također, funkcije distribucije Gumbelove domene mogu biti i ograničene i neograničene odozgo.

Primjer 1.2.7 (Eksponencijalna distribucija)**.** *Neka je F funkcija distribucije eksponenijalne slučajne varijable s parametrom* 1, *odnosno $F(x) = 1 - e^{-x}, x \geq 0$ i $F(x) = 0, x < 0$. Teoremom 3.1.19 dani su nužni i dovoljni uvjeti za pripadnost distribucije Gumbelovoj domeni atrakcije. No, u slučaju eksponencijalne funkcije, nije teško pokazati pripadnost Gumbelovoj domeni direktno. Vrijedi*

$$\begin{aligned}\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) &= F^n(a_n x + b_n) \\ &= \left[1 - \exp\Big(-(a_n x + b_n)\Big)\right]^n\end{aligned}$$

Ako je $a_n = 1$ *i* $b_n = \log n$, *vrijedi*

$$\begin{aligned}\Big[1 - \exp\Big(-(a_n x + b_n)\Big)\Big]^n &= \Big[1 - \exp\Big(-(x + \log n)\Big)\Big]^n \\ &= \Big[1 - n^{-1}\exp(-x)\Big)\Big]^n \\ &\to \exp(-e^{-x}), \quad n \to \infty,\ \forall x \in \mathbb{R},\end{aligned}$$

pa prema definiciji 3.1.2 niz maksimuma dobiven iz početnih podataka s eksponencijalnom distribucijom s parametrom 1 *konvergira po distribuciji k slučajnoj varijabli s funkcijom distribucije iz Gumbelove familije uz normirajuće konstante* $a_n = 1$, $b_n = \log n$, *odnosno*

$$(M_n - \log n) \xrightarrow{d} G_G$$

Eksponencijalna distribucija ima vrlo široku primjenu u teoriji te modeliranju pojava iz svakodnevnoga života. Uz pomoć eksponencijalne distribucije modeliraju se pojave poput vremena kada bi se mogao dogoditi idući potres, broja klijenata koji će nazvati korisničku službu, životnoga vijeka nekog uređaja i mnoge druge.

Primjer 1.2.8 (Normalna distribucija)**.** *Najpoznatija distribucija u vjerojatnosti i statistici zasigurno je standardna normalna distribucija s gustoćom*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}.$$

Može se pokazati kako normalna distribucija pripada Gumbelovoj domeni atrakcije uz normirajuće konstante

$$a_n = (2\log n)^{1/2}, \tag{1.23}$$

$$b_n = (2\log n)^{1/2} - \frac{1}{2}(2\log n)^{-1/2}(\log\log n + \log 4\pi). \tag{1.24}$$

Detaljan izvod konstanti iz primjera 1.2.8, kao i izvodi konstanti i provjere nužnih i dovoljnih uvjeta za pripadnost pojedinoj domeni atrakcije iz ostalih primjera mogu se pronaći u [2] i [4].
Na kraju ovoga odjeljka valja napomenuti bitnu posljedicu teorema 1.1.10, a ta je da svaka od distribucija za ekstremne vrijednosti i sama pripada vlastitoj domeni atrakcije. Normirajuće konstante dobiju se koristeći svojstvo max-stabilnosti na potpuno jednak način kao u dokazu teorema 1.1.12.

Napomena 1.2.9. *i) Funkcije distribucije tipa I pripadaju Gumbelovoj domeni atrakcije uz normirajuće konstante* $a_n = 1$, $b_n = \log n$.

ii) Funkcije distribucije tipa II pripadaju Fréchetovoj domeni atrakcije uz normirajuće konstante $a_n = n^{-1/\alpha}$, $b_n = 0$.

iii) Funkcije distribucije tipa III pripadaju Weibullovoj domeni atrakcije uz normirajuće konstante $a_n = n^{1/\alpha}$, $b_n = 0$.

1.3 Generalizirana distribucija ekstremnih vrijednosti

U odjeljku 1.2 spomenuto je kako se tri tipa distribucija ekstremnih vrijednosti znatno razlikuju po ponašanju svojih repova te se, sukladno tomu, u primjeni koriste za modeliranje veoma različitih pojava. U samom je početku razvoja teorije ekstremnih vrijednosti modeliranje određene pojave podrazumijevalo apriori odabir jednog od tri tipa distribucija za maksimum nakon čega bi se procijenili parametri modela i donosili zaključci. Međutim, takav pristup imao je dva bitna nedostatka. Najprije je trebala postojati neka tehnika kojom bi se odabrala najpogodnija od tri ponuđene familije distribucija. Nadalje, svi kasnije doneseni zaključci podrazumijevali su kako je odabir familije uistinu optimalan, što nije morao uvijek biti slučaj. U svrhu rješavanja tih problema su Gumbelova, Fréchetova i Weibullova familija povezane u jedinstvenu familiju funkcija poznatu kao **generalizirana distribucija ekstremnih vrijednosti** (engl. *Generalized Extreme Value Distribution*, GEV).

Definicija 1.3.1. *Generalizirana distribucija ekstremnih vrijednosti familija je funkcija distribucija oblika*

$$G(x) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\,\xi} \right\} \tag{1.25}$$

definirana na skupu $\{x \in \mathbb{R}\colon 1 + \xi(x-\mu)/\sigma > 0\}$, *pri čemu parametri zadovoljavaju* $-\infty < \mu < \infty$, $\sigma > 0$ *te* $-\infty < \xi < \infty$.

Slučaj $\xi = 0$ u GEV familiji tretira se kao limes izraza (1.25) kada $\xi \to 0$, što zapravo odgovara Gumbelovoj familiji s funkcijom distribucije

$$G(x) = \exp\left\{ \exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, \quad -\infty < x < \infty.$$

Nadalje, slučaj $\xi > 0$ odgovara Fréchetovoj familiji distribucija. Neka je zadana vrijednost parametra $\alpha > 0$ u (1.4) te neka je $\xi = \alpha^{-1}$. Tada je $\xi > 0$ te za izraz (1.25) vrijedi

$$G(x) = \exp\left\{ -\left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\sigma}\right)\right]^{-\alpha} \right\}, \quad \text{kada je } x > \mu - \alpha\sigma,$$

te $G(x) = 0$ za $x \leq \mu - \alpha\sigma$. Sada se raspisivanjem dobije da za $x > \mu - \alpha\sigma$ vrijedi

$$\begin{aligned} G(x) &= \exp\left\{ -\left[1 + \frac{1}{\alpha}\left(\frac{x-\mu}{\sigma}\right)\right]^{-\alpha}\right\} \\ &= \exp\left[-\left(\frac{x-\mu+\alpha\sigma}{\alpha\sigma}\right)^{-\alpha}\right] \\ &= \exp\left[-\left(\frac{x-b}{a}\right)^{-\alpha}\right], \end{aligned}$$

pri čemu je $a = \alpha\sigma > 0$, a $b = \mu - \alpha\sigma \in \mathbb{R}$. Dakle, GEV familija definirana izrazom (1.25) za $\xi = \alpha^{-1} > 0$ može se zapisati kao

$$G(x) = \begin{cases} 0, & x \leq b, \\ \exp\left[-\left(\dfrac{x-b}{a}\right)^{-\alpha}\right], & x > b, \end{cases}$$

što odgovara Fréchetovoj familiji funkcija (1.4). Analogno se pokaže da slučaj $\xi < 0$ odgovara Weibullovoj familiji.

Ovakav jedinstveni zapis familija distribucija za ekstreme uvelike je olakšao implementaciju modela jer se određivanjem parametra ξ iz dostupnih podataka odmah odabere najprikladnija familija distribucija. Također, određivanjem pouzdanosti procijenjenoga parametra ξ utvrđuje se koliko uistinu odabrana familija dobro modelira proučavane podatke, osobito ponašanje repa. Sada je moguće iskazati Fisher–Tippett–Gnedenkov teorem za GEV familiju.

Teorem 1.3.2. *Neka su $(a_n)_n$ i $(b_n)_n$ nizovi realnih konstanti takvi da vrijedi $a_n > 0$ za sve $n \in \mathbb{N}$ i*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \longrightarrow G(x) \quad \textit{kada} \quad n \to \infty, \ \forall x \in C(G), \tag{1.26}$$

pri čemu je G nedegenerirana funkcija distribucije, a $C(G)$ skup točaka neprekidnosti od G. Tada se G nalazi u GEV familiji funkcija distribucija oblika

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

pri čemu je $\{x \in \mathbb{R}\colon 1 + \xi(x-\mu)/\sigma > 0\}$, $-\infty < \mu < \infty$, $\sigma > 0$ te $-\infty < \xi < \infty$.

Ako je distribucija polaznih podataka poznata, bogata teorija domena atrakcije osigurava načine odabira odgovarajućih normirajućih konstanti a_n i b_n, kao što je opisano u odjeljku 1.2. Međutim, u primjeni to najčešće nije slučaj te preostaje problem odabira odgovarajućih konstanti uz koje će uvjeti teorema 1.3.2 biti zadovoljeni. Budući da stvaranje prikladnoga modela svakako podrazumijeva određivanje parametara GEV distribucije, odabir normirajućih konstanti ne predstavlja prevelik problem. Naime, ako vrijedi (1.26), tada je za velike $n \in \mathbb{N}$ te $\forall x \in \mathbb{R}$

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx G(x),$$

odnosno normalizirani maksimumi ponašaju se približno kao nedegenerirana funkcija distribucije G. Tada je ekvivalentan zapis

$$P(M_n \leq x) \approx G\left(\frac{x - b_n}{a_n}\right) = G^*(x),$$

gdje je $G^*(x)$ također funkcija distribucije iz GEV familije. Naime, vrijedi

$$\begin{aligned} G\left(\frac{x - b_n}{a_n}\right) &= \exp\left\{ -\left[1 + \xi\left(\frac{((x - b_n)/a_n) - \mu}{\sigma}\right)\right]^{-1/\xi}\right\} \\ &= \exp\left\{ -\left[1 + \xi\left(\frac{x - b_n - a_n\mu}{a_n\sigma}\right)\right]^{-1/\xi}\right\} \\ &= \exp\left\{ -\left[1 + \xi\left(\frac{x - \mu^*}{\sigma^*}\right)\right]^{-1/\xi}\right\} \\ &= G^*(x), \end{aligned}$$

pa se $G^*(x)$ može zapisati u obliku (1.25), pri čemu su $-\infty < \xi < \infty$, $\sigma^* = a_n\sigma > 0$ i $\mu^* = a_n\mu + b_n$, $-\infty < \mu^* < \infty$ parametri distribucije $G^*(x)$.
Dakle, ako je distribuciju normaliziranih maksimuma $(M_n - b_n)/a_n$ moguće aproksimirati nekim članom GEV familije, tada je distribuciju od M_n moguće aproksimirati nekim drugim članom GEV familije pa se pronalazak normirajućih konstanti svodi na određivanje parametara odgovarajuće distribucije ekstrema.

1.3.1 Povratni period i razina povrata

Teoremom 1.3.2 dana je GEV familija kao prikladan model za distribuciju maksimuma blokova. Primjena GEV familije u modeliranju podrazumijeva raspoređivanje dostupnih poda-

taka, koji predstavljaju realizacije niza nezavisnih i jednakodistribuiranih slučajnih varijabli, u nizove opservacija, odnosno blokove, veličine n, za neku veliku vrijednost n. Na takav se način iz $n \times m$ polaznih podataka dobiva niz maksimuma $M_n^{(k)}$, $k = 1, 2, \ldots, m$, $m \in \mathbb{N}$, kojima valja prilagoditi funkciju distribucije iz GEV familije. U implementaciji je optimalan odabir veličine bloka ključan. Naime, premaleni blokovi mogu dovesti do loše asimptotske procjene, a samim time i do pristranosti u procjenama i ekstrapolaciji. S druge strane, uz preveliku veličinu bloka dobiva se manje maksimuma blokova, što dovodi do velike varijance u procjeni. Zato se odabir optimalne veličine blokova svodi na uspostavljanje ravnoteže između pristranosti i varijance procjene. U primjeni se zbog povijesnih razloga, ali i pragmatičnosti najčešće promatraju godišnji maksimumi, odnosno podaci se grupiraju u blokove tako da veličina bloka n predstavlja broj zabilježenih opservacija u jednoj godini. Ovakav pristup u primjenama se pokazao vrlo dobrim jer je analiza godišnjih maksimuma robusnija nego analiza manjih blokova ako nisu ispunjeni svi uvjeti teorema 1.3.2, kao i zbog činjenice da su za mnoštvo mjerenih pojava zabilježeni upravo samo godišnji maksimumi. Budući da je glavni cilj metode maksimuma blokova ekstrapolacija budućih maksimuma iz dostupnih podataka, u interpretaciji rezultata ključni su pojmovi **razine povrata** te **povratnog perioda**.
Kvantili distribucije maksimuma G dobivaju se invertiranjem izraza (1.25) te su jednaki

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left\{1 - [-\log(1-p)]^{-\xi}\right\}, & \xi \neq 0, \\ \mu - \sigma \log\left[-\log(1-p)\right], & \xi = 0, \end{cases} \tag{1.27}$$

pri čemu je $G(x_p) = 1 - p$. Veličinu x_p naziva se razinom povrata, a $1/p$ povratnim periodom.
Razina povrata x_p vrijednost je za koju se očekuje da će biti nadmašena u prosjeku jednom svakih $1/p$ godina. Preciznije, ako promatramo blokove od godinu dana, godišnji će maksimum biti veći od x_p s vjerojatnošću p svake godine.

1.4 Prilagodba GEV modela

Neka sada vrijednosti $Z_1, Z_2, \ldots, Z_m$, $m \in \mathbb{N}$ označavaju proučavane maksimume blokova. Po pretpostavci modela riječ je o nezavisnim slučajnim varijablama čija se funkcija distribucije nalazi u GEV familiji i ima nepoznate parametre. Postoji mnogo tehnika za procjenu parametara GEV distribucije poput uporabe vjerojatnosnih grafova, metode momenata, uređajnih statistika itd. U ovome radu naglasak je na procjeni parametara metodom maksimalne vjerodostojnosti, koja se u primjeni pokazala optimalnom kada je riječ o kompleksnim modelima i kvantificiranju nepouzdanosti dobivenih procjena i zaključaka.

Definicija 1.4.1. *Neka je* $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$, $k \in \mathbb{N}$ *slučajan uzorak duljine* k *iz statističkoga modela* $\mathcal{P} = \{f(\cdot\colon \theta)\colon \theta \in \Theta\}$, *pri čemu* Θ *označava familiju parametara funkcija*

gustoće promatranoga statističkoga modela, te neka je $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ *jedna njegova realizacija. Tada je vjerodostojnost (engl.* likelihood*) funkcija* $L\colon \Theta \to \mathbb{R}$, *definirana s*

$$L(\theta) = L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{k} f(x_i\colon \theta),\ \theta \in \Theta, \tag{1.28}$$

gdje f *označava funkciju gustoće slučajnih varijabli* X_i, $i = 1, \ldots, k$. *Za zadanu vrijednost* $\theta \in \Theta$, *broj* $L(\theta) = L(\theta \mid x)$ *nazivamo* ***vjerodostojnošću vrijednosti parametra*** θ ***na temelju opaženoga uzorka*** $\boldsymbol{x}$.

Definicija 1.4.2. *Statistika* $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$ *je* ***procjenitelj maksimalne vjerodostojnosti*** *(engl.* maximum likelihood estimator*, MLE) ako vrijedi*

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta \mid \boldsymbol{X}). \tag{1.29}$$

Ponekad se, radi lakšega računanja, umjesto vjerodostojnosti promatra **log-vjerodostojnost** definirana kao

$$l(\theta) = l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{i=1}^{k} \log f(x_i\colon \theta), \tag{1.30}$$

pa se MLE traži maksimiziranjem log-vjerodostojnosti. Ključni pojmovi i rezultati vezani za metodu maksimalne vjerodostojnosti nalaze se u odjeljku 3.1.
Budući da je vjerodostojnost definirana pomoću funkcije gustoće slučajne varijable, potrebna gustoća slučajnih varijabli $Z_1, \ldots, Z_m$, odnosno gustoća GEV distribucije, dana je s

$$g(x) = \begin{cases} \frac{1}{\sigma}\Big[1 + \xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-1-1/\xi} \exp\Big\{-\Big[1 + \xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}, & \xi \neq 0, \\ \frac{1}{\sigma}\exp\Big[-\Big(\frac{x-\mu}{\sigma}\Big)\Big]\exp\Big\{-\exp\Big[-\Big(\frac{x-\mu}{\sigma}\Big)\Big]\Big\}, & \xi = 0. \end{cases} \tag{1.31}$$

Ako su $z_1, \ldots, z_m$ realizacije slučajnih varijabli koje predstavljaju maksimume blokova, koristeći 1.29 i 1.30 dobije se da je funkcija log-vjerodostojnosti za parametre GEV distribucije dana s

$$l(\mu, \sigma, \xi) = -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}, \tag{1.32}$$

kada je $\xi \neq 0$ i vrijedi

$$1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0,\ i = 1, 2, \ldots, m. \tag{1.33}$$

Ako su parametri takvi da uvjet (1.33) nije ispunjen, vjerodostojnost je jednaka 0, a log-vjerodostojnost je $-\infty$.
Kada je $\xi = 0$, funkcija log-vjerodostojnosti dana je s

$$l(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^{m} \exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\}. \tag{1.34}$$

Sada se procjenitelj maksimalne vjerodostojnosti parametara (μ, σ, ξ) dobije maksimizacijom funkcija definiranih s (1.32) i (1.34). Analitičko rješenje problema ne postoji, no u primjeni se problemi maksimizacije log-vjerodostojnosti rješavaju uporabom standardnih numeričkih algoritama za optimizaciju.
Potencijalni problem uporabe metode maksimalne vjerodostojnosti na GEV distribuciji moguće je narušavanje uvjeta regularnosti potrebnih kako bi procjenitelj maksimalne vjerodostojnosti imao standardna asimptotska svojstva. Definicije regularnosti, asimptotske normalnosti procjenitelja te druge definicije i rezultati vezani uz metodu maksimalne vjerodostojnosti nalaze se u odjeljku 3.1. Može se pokazati kako su za $\xi > -0.5$ MLE procjenitelji regularni te imaju uobičajena asimptotska svojstva. Kada je $-1 < \xi < -0.5$ MLE procjenitelji nemaju standardna asimptotska svojstva, ali su općenito održivi. Za $\xi < -1$ nije preporučivo koristiti MLE procjenu. Budući da slučaj $\xi < -0.5$ odgovara distribucijama s ograničenim i lakim repom te da se takve distribucije rijetko koriste u modeliranju ekstremnih vrijednosti, eventualno narušavanje uvjeta regularnosti u pravilu ne stvara probleme u primjeni pa se može reći kako je aproksimativna distribucija procijenjenih parametara $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ multivarijatna normalna s očekivanjem (μ, σ, ξ) i kovarijacijskom matricom jednakom inverzu matrice informacija evaluirane u parametrima modela dobivenima metodom maksimalne vjerodostojnosti. Također, pouzdani intervali slijede direktno iz asimptotske normalnosti procjenitelja.

1.4.1 Zaključivanje o razinama povrata

Nakon pronalaska MLE procjenitelja za parametre modela (μ, σ, ξ) razina povrata i povratni period dobivaju se uvrštavanjem procijenjenih vrijednosti $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ u (1.27),odnosno

$$\hat{x_p} = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \left\{ 1 - [-\log(1-p)]^{-\hat{\xi}} \right\}, & \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma} \log \left[-\log(1-p) \right], & \hat{\xi} = 0. \end{cases} \tag{1.35}$$

Također, korištenjem delta metode dobije se da za varijancu razine povrata vrijedi

$$Var(\hat{x_p}) \approx \nabla x_p^T V \nabla x_p, \tag{1.36}$$

pri čemu je V kovarijacijska matrica od $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$, a ∇x_p^T definira se kao vrijednost izraza

$$\nabla x_p^T = \left[\frac{\partial x_p}{\partial \mu}, \frac{\partial x_p}{\partial \sigma}, \frac{\partial x_p}{\partial \xi} \right]$$

evaluiranoga u $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.
Kao što je već spomenuto, očekuje se kako će vrijednost dobivenoga povratnog perioda biti nadmašena s vjerojatnošću p svake godine, ako promatramo blokove od godinu dana. Stoga je u primjeni obično najzanimljivije promatrati duge povratne periode, koji odgovaraju maloj vrijednosti parametra p. Također, ako je $\hat{\xi} < 0$, mogu se donositi i zaključci o gornjoj završnoj točki distribucije te se na taj način može analizirati i *beskonačni povratni period* koji odgovara slučaju $p = 0$. U tom je slučaju povratni period dan s $\hat{x_0} = \hat{\mu} - \hat{\sigma}/\hat{\xi}$. S druge strane, ako je $\xi \geq 0$, gornja završna točka je u beskonačnosti.
Ipak, treba biti oprezan u shvaćanju zaključaka dobivenih za razine povrata i povratne periode, osobito kada je riječ o dugim povratnim periodima. Jedan je razlog taj što aproksimacija distribucije procjenitelja normalnom distribucijom možda nije optimalna. Također, sve procjene i njihova pouzdanost donose se pod pretpostavkom da je način modeliranja promatrane pojave točan. Naime, modeliranje ekstrema pomoću GEV distribucije podrazumijeva ekstrapolaciju budućih vrijednosti implementacijom limesa kao konačnih aproksimacija. Međutim, nema jamstva kako su stohastički mehanizmi promatranoga procesa dovoljno glatki da bi omogućili ekstrapolaciju predviđenih vrijednosti. Taj problem u vjerojatnosti i statistici poznat je kao paradigma ekstremnih vrijednosti (engl. *extreme value paradigm*). Zbog svega navedenoga, mjere pouzdanosti modela, poput, primjerice, pouzdanih intervala za razine povrata, sigurnije je shvaćati kao ishod u idealnome slučaju te očekivati da su mogući i mnogo širi intervali uzme li se u obzir nepouzdanost u točnost modela.
Također, klasičan način računanja pouzdanih intervala oslanja se na već spomenuta asimptotska svojstva procjenitelja metodom maksimalne vjerodostojnosti, konkretno, na asimptotsku normalnost. Još jedan, u pravilu točniji, način dobivanja pouzdanih intervala je uz pomoć funkcije profil-vjerodostojnosti (engl. *profile likelihood*), čija su definicija i svojstva navedeni u odjeljku 3.1. Kod GEV distribucije, profil-vjerodostojnost svakog od parametara μ, σ, ξ zapravo podrazumijeva fiksiranje vrijednosti parametra za koji se traži profil-vjerodostojnost, primjerice $\xi = \xi_0$, $\xi_0 \in \mathbb{R}$ te maksimizaciju log-vjerodostojnosti (1.30) s obzirom na preostala dva parametra. Takav postupak ponavlja se za nekoliko fiksnih vrijednosti parametra ξ te se na kraju dobivene maksimizirane vrijednosti log-vjerodostojnosti nazivaju profil-vjerodostojnošću od ξ. Pouzdani intervali za razine povrata dobivaju se reparametrizacijom GEV modela tako da x_p bude jedan od parametara. Reparametrizacija poprilično direktno slijedi ako se u (1.27) parametar μ izrazi preko x_p, σ i ξ i potom uvrsti u model. Tada je način izračuna pouzdanih intervala za x_p uz profil-vjerodostojnost dan teoremom 3.1.14.

1.4.2 Provjera modela

Iako je nemoguće unaprijed provjeriti točnost budućih ekstremnih vrijednosti ekstrapoliranih iz prilagođenoga GEV modela, moguće je provjeriti kako se model ponaša za već opažene vrijednosti. Ako je prilagodba modela povijesnim podacima loša, preoptimistično je nadati se kako će imati veliku predikcijsku moć kada su u pitanju budući ekstremi. Prilagođenost GEV modela podacima najčešće se ispituje grafički i to koristeći vjerojatnosni graf, graf kvantila ili graf razine povrata.

Vjerojatnosni graf naziv je za grafičku usporedbu empirijske i modelom prilagođene funkcije distribucije podataka. Ako $z_1, \ldots, z_m$ predstavljaju promatrane maksimume blokova, moguće ih je poredati tako da vrijedi $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(m)}$. Tada za empirijsku funkciju distribucije $\tilde{G}(z_{(i)})$ vrijedi $\tilde{G}(z_{(i)}) = i/(m+1)$, $i = 1, \ldots, m$. Funkcija distribucije procijenjena GEV modelom jednaka je

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

Tada se vjerojatnosni graf sastoji od skupa točaka u ravnini danog s

$$\left\{ \left(\tilde{G}(z_{(i)}), \hat{G}(z_{(i)})\right),\ i = 1, \ldots, m\right\}. \tag{1.37}$$

Ako GEV model dobro aproksimira podatke, točke će se u ravnini grupirati oko pravca $y = x$. Značajna odstupanja sugeriraju nedostatke u GEV modelu. Potencijalni nedostatak vjerojatnosnoga grafa je taj što se i $\hat{G}(z_{(i)})$ i $\tilde{G}(z_{(i)})$ približavaju 1 kako $z_{(i)}$ raste, a ponašanje modela najbitnije je upravo za velike vrijednosti maksimuma. Zato se promatra i **graf kvantila**, odnosno skup točaka

$$\left\{\left(\hat{G}^{-1}(i/(m+1)), z_{(i)}\right),\ i = 1, \ldots, m\right\}, \tag{1.38}$$

gdje je, prema (1.27),

$$\hat{G}^{-1}\left(\frac{i}{m+1}\right) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left\{1 - \left[-\log\left(\frac{i}{m+1}\right)\right]^{-\hat{\xi}}\right\}.$$

Kao i kod vjerojatnosnoga grafa, značajna odstupanja od pravca $y = x$ ukazuju na neuspješnost GEV modela.

Općenito, kvantili omogućuju prezentaciju vjerojatnosnih modela na istoj skali s podacima. Za interpretaciju modela ekstremnih vrijednosti posebno je pogodan **graf razine povrata**.

Ako se u izrazu (1.27), kojim se definira razina povrata, s y_p označi vrijednost $-\log(1-p)$, graf razine povrata skup je točaka dan s

$$\left\{\Big(-\log y_p, x_p\Big)\colon 0 < p < 1\right\}. \tag{1.39}$$

Graf će biti konkavan s asimptotom u $\mu - \sigma/\xi$ kada $p \to 0$ kada je $\xi < 0$, konveksan za $\xi > 0$ i linearan za $\xi = 0$. Budući da su na grafu povrata naglašene procjene razina povrata za duge povratne periode, grafovi povrata koriste se za validaciju, ali i prezentaciju samoga modela. U primjeni je x_p vrijednost jednaka MLE procjeni razine povrata. Također, na graf povrata mogu se dodati i pouzdani intervali te kvantili procijenjeni modelom. Na taj način valjanost procijenjenoga GEV modela vidljiva je iz toga koliko se dobiveni graf povrata dobro *slaže* s empirijskom procjenom funkcije povrata.

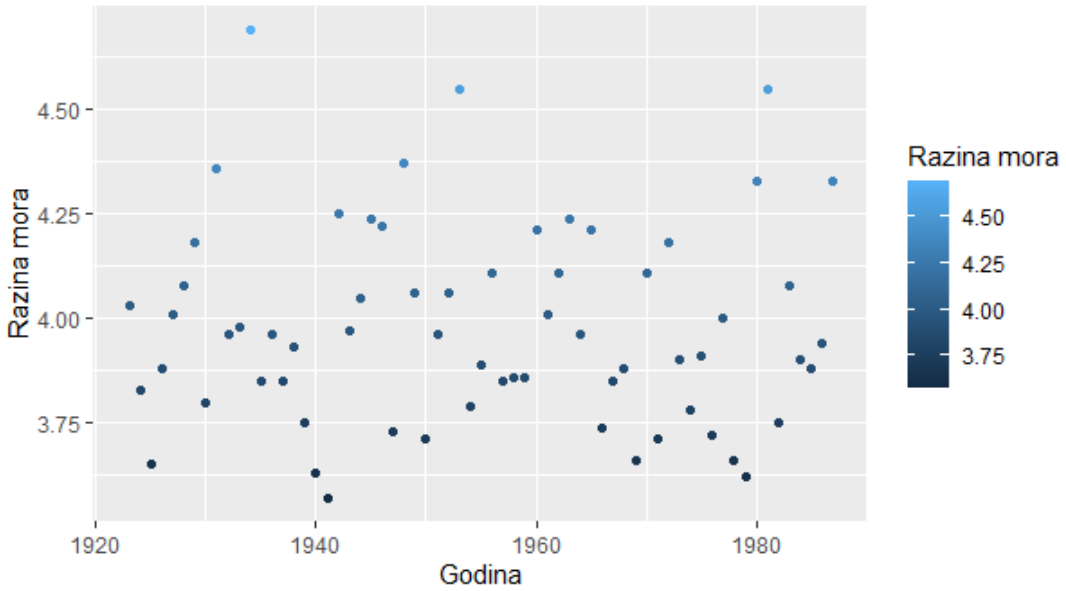
Poglavlje 2

Primjena metode maksimuma blokova

Ovo poglavlje donosi primjenu metoda zaključivanja o budućim ekstremnim vrijednostima obrađenih u prvom poglavlju na primjerima podataka iz stvarnog svijeta. U prvom je primjeru riječ o procjeni budućih maksimalnih razina mora u australskom gradiću Port Pirie koristeći dostupne podatke o maksimalnim morskim razinama u prošlosti. Drugi se primjer bavi izdržljivošću staklenih vlakana. Preciznije, uz pretpostavku da izdržljivost nekog predmeta izrađenoga od staklenih vlakana ovisi o izdržljivosti najslabijega vlakna, cilj metode maksimuma blokova bit će predvidjeti najmanju razinu opterećenja koja bi mogla uništiti stakleno vlakno. Oba primjera obrađena su u [1]. Treći primjer bavi se analizom maksimalnih ljetnih temperatura u Zagrebu. Svi rezultati u ovom poglavlju dobiveni su koristeći programski jezik `R` i u njemu dostupne pakete.

2.1 Razina mora u Port Pirieju

U razdoblju 1923.-1987. bilježene su maksimalne godišnje razine mora u metrima na obali australskoga grada Port Pirie, koje su prikazane na slici 2.1. Najniža razina mora zabilježena je 1941. godine i iznosila je 3.57 metara, dok je najviša iznosila 4.69 metara i zabilježena je 1934. godine. Kada su u pitanju prirodne pojave, ljudima je uvijek bitno pokušati predvidjeti njihovo ponašanje u budućnosti, najčešće kako bi bili spremni na ekstremne situacije koje bi im mogle nanijeti štetu. Tako se može postaviti pitanje kolika je maksimalna razina mora koju stanovnici Port Pirieja mogu očekivati u idućih deset, sto ili pak tisuću godina. Upravo se ovakvim procjenama bavi teorija ekstremnih vrijednosti. No, bitno je napomenuti kako je nemoguće s velikom sigurnošću govoriti o točnosti bilo kakvih procjena jer promatrana pojava ovisi o mnoštvu nepredvidivih faktora, poput, primjerice, budućih klimatskih promjena. Na slici 2.1 nisu vidljivi uzorci u varijacijama morskih razina tijekom godina, odnosno povijesni podaci se čine stacionarnima, ali to ne mora biti slučaj i s budućim podacima. Dakle, o budućim maksimalnim razinama mora do-



Slika 2.1: Maksimalne godišnje razine mora na obalama Port Pirieja.

bivenima metodom maksimuma blokova zapravo ima smisla govoriti kao o razinama koje će se određene godine dostići s nekom vjerojatnošću pod uvjetom da se proces ponaša kao u prošlosti.

Dakle, promatramo maksimume blokova veličine jedne godine. Podaci korišteni u ovom primjeru dostupni su u sklopu paketa `ismev` u R-u. Cilj je odrediti razine povrata (1.27) za povratne periode koji nas zanimaju. U tu svrhu najprije valja prilagoditi podacima odgovarajući GEV model, odnosno pronaći parametre GEV distribucije maksimalnih razina mora oblika (1.25). Procjene parametara dobivaju se metodom maksimalne vjerodostojnosti opisanom u odjeljku 1.4. Takva metoda implementirana je u R-ovu paketu `extRemes` funkcijom `fevd`. Dobiveni su parametri $\hat{\mu} = 3.874$, $\hat{\sigma} = 0.198$ te $\hat{\xi} = -0.0501$. Koristeći asimptotsku normalnost MLE procjenitelja donja i gornja granica 95% pouzdanih intervala za procijenjene parametre dobivaju se tako da se procijenjeni parametar uveća i umanji za umnožak standardne pogreške i odgovarajućega kvantila normalne razdiobe, koji iznosi 1.96 u slučaju 95% pouzdanih intervala. Standardne pogreške iznose 0.0279 za parametar

$\hat{\mu}$, 0.0202 za parametar $\hat{\sigma}$ te 0.0983 za parametar $\hat{\xi}$. Sada su 95% pouzdani intervali jednaki

$$\begin{aligned} [3.8193, 3.9287] \quad & \text{za } \mu, \\ [0.1584, 0.2376] \quad & \text{za } \sigma, \\ [-0.2428, 0.1426] \quad & \text{za } \xi. \end{aligned}$$

Procjena parametra ξ jednaka je -0.0501, što sugerira kako je riječ o funkciji distribucije iz Weibullove familije, a samim time i o funkciji distribucije s konačnom desnom završnom točkom. No, 95% pouzdani interval za ξ uključuje i 0 i vrijednosti veće od 0 pa ne možemo zaključiti kako je distribucija promatranih maksimuma morskih razina uistinu odozgo ograničena.

Jednom kada su parametri modela procijenjeni, moguće je izračunati razine povrata za potrebne povratne periode uvrštavanjem $\hat{\mu}$, $\hat{\sigma}$ i $\hat{\xi}$ u izraz (1.27). Primjerice, pretpostavimo da nas zanima razina povrata za povratni period od 10 godina. Tada je parametar p jednak 0.1 te iz (1.27) slijedi $\hat{x}_{0.1} = 4.2954$. Pouzdani intervali za $\hat{x}_{0.1}$ također se dobivaju koristeći asimptotsku normalnost MLE procjenitelja, pri čemu se standardna pogreška dobije kao korijen vrijednosti varijance razine povrata dane izrazom (1.36). Kovarijacijska matrica procijenjenih parametara jednaka je

$$V = \begin{pmatrix} 0.000780 & 0.000197 & -0.00107 \\ 0.000197 & 0.000410 & -0.000778 \\ -0.00107 & -0.000778 & 0.00965 \end{pmatrix},$$

pa, prema (1.36), vrijedi $\mathrm{Var}(\hat{x}_{0.1}) = 0.003028$ te je 95% pouzdani interval za $x_{0.1}$ jednak $[4.1875, 4.4033]$. Alternativno, razine povrata i pouzdani intervali mogu se izračunati koristeći funkciju `return.level` iz `R`-ova paketa `extRemes`. U tablici 2.1 navedene su razine povrata, zajedno s 95% pouzdanim intervalima, za povratne periode od pet, deset, pedeset i sto godina.

Povratni period	Razina povrata	95% pouzdani intervali (delta metoda)
5 godina	4.16	[4.08, 4.24]
10 godina	4.30	[4.19, 4.40]
50 godina	4.58	[4.34, 4.81]
100 godina	4.69	[4.38, 5.00]

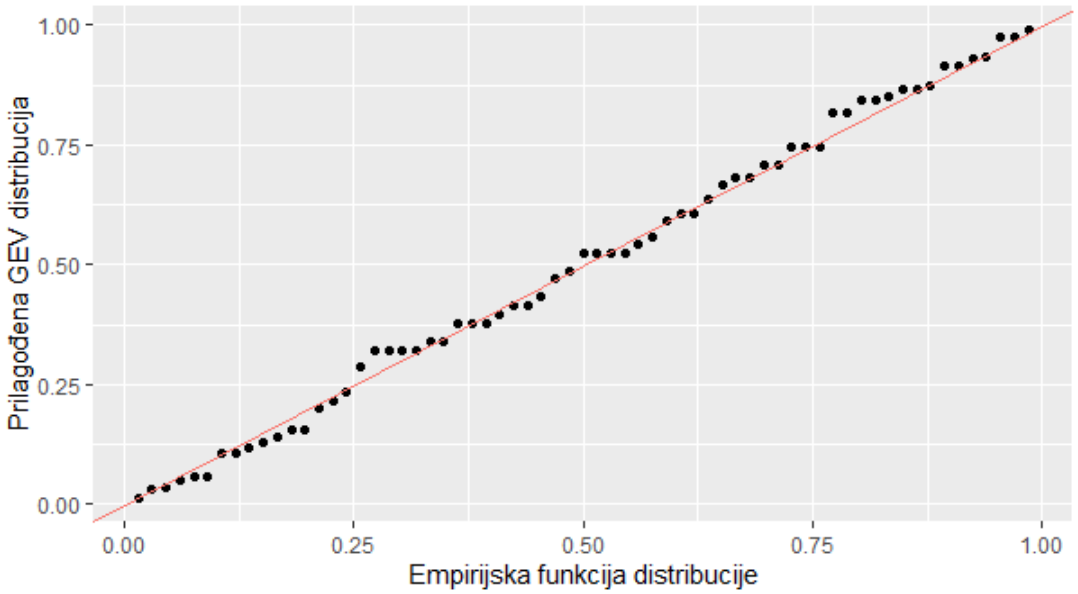
Tablica 2.1: Razine povrata i 95% pouzdani intervali za razine povrata.

U odjeljku 1.4.1 spomenuto je kako se u pravilu precizniji pouzdani intervali za razine povrata dobivaju koristeći metodu maksimizacije profil-vjerodostojnosti umjesto asimptotske normalnosti MLE procjenitelja. Pouzdani intervali dobiveni korištenjem profil-vjerodostojnosti za povratne periode od pet, deset, pedeset i sto godina nalaze se u tablici 2.2, a dobiveni su uz pomoć funkcije `gevrRl`, koju se može pronaći u paketu `eva`.

Povratni period	Razina povrata	95% pouzdani interval (profil-vjerodostojnost)
5 godina	4.16	[4.09, 4.25]
10 godina	4.30	[4.20, 4.45]
50 godina	4.58	[4.42, 4.98]
100 godina	4.69	[4.49, 5.27]

Tablica 2.2: Razine povrata i 95% pouzdani intervali za razine povrata.

Za povratne periode od pet i deset godina pouzdani intervali dobiveni metodom profil-vjerodostojnosti slični su onima dobivenima delta metodom, odnosno koristeći asimptotsku normalnost MLE procjenitelja. S druge strane, za povratne periode od pedeset i sto godina postoje veće razlike među dobivenim pouzdanim intervalima. Razlog je asimetričnost profil-vjerodostojnosti koja se povećava povećanjem povratnoga perioda.
Kao što je već spomenuto, nemoguće je provjeriti točnost budućih maksimalnih razina mora ekstrapoliranih iz prilagođenoga GEV modela te na taj način provjeriti je li dobivena GEV distribucija dobra za modeliranje podataka. Međutim, moguće je provjeriti je li prilagodba modela povijesnim podacima dobra koristeći grafičke prikaze opisane u odjeljku 1.4.2. Na slici 2.2 prikazan je vjerojatnosni graf definiran skupom točaka (1.37). Može se vidjeti kako se točke dobro grupiraju oko dijagonale, koja je na grafu označena crvenom bojom, pa vjerojatnosni graf ne daje razloga za sumnju u prilagođenost GEV modela dostupnim podacima.
Na slici 2.3 prikazan je graf kvantila prilagođenoga GEV modela dan skupom točaka (1.38). Dijagonala je ponovno označena crvenom bojom. Ne primjećuju se znatna odstupanja točaka od dijagonale pa ni graf kvantila ne sugerira lošu prilagođenost promatranoga GEV modela podacima.
Naposljetku, na slici 2.4 nalazi se graf razina povrata definiran kao skup točaka (1.39), a na grafu prikazan crvenom linijom. Točke predstavljaju empirijske razine povrata, a crvene isprekidane linije 95% pouzdane intervale za razine povrata procijenjene modelom. Pouzdani intervali dobiveni su koristeći asimptotsku normalnost MLE procjenitelja. Može se primijetiti kako nema velikih odstupanja u preklapanju empirijskih i modelom procijenjenih razina povrata te da se sve vrijednosti nalaze unutar 95% pouzdanih intervala. Budući da nisu uočena značajna odstupanja na vjerojatnosnome te grafu kvantila i razina povrata, nemamo razloga pretpostaviti da prilagodba dobivenoga GEV modela podacima nije dobra.
Prije pojave GEV familije, odgovarajuća distribucija maksimuma blokova pronalazila se odabirom distribucije iz Gumbelove, Fréchetove ili Weibullove familije kojoj bi se procijenili parametri te bi se zaključci donosili na temelju tako dobivenoga modela. Glavni problem takva pristupa, koji je zapravo bio i glavna motivacija u razvoju GEV familije, bio je što nema jamstva da je odabrana optimalna familija distribucija za maksimum. Stoga

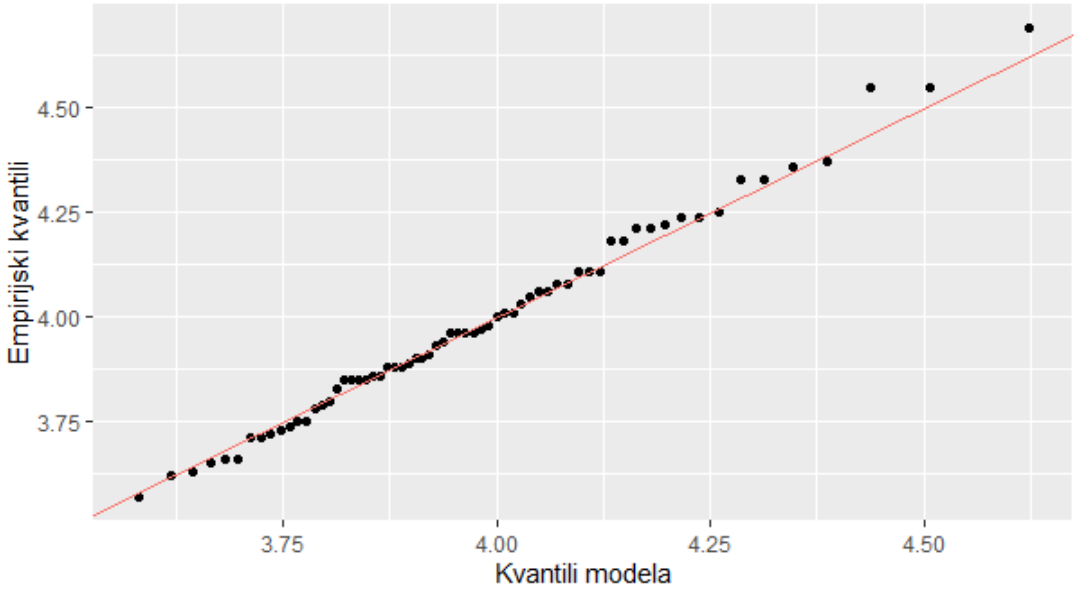


Slika 2.2: Vjerojatnosni graf prilagođenoga GEV modela.

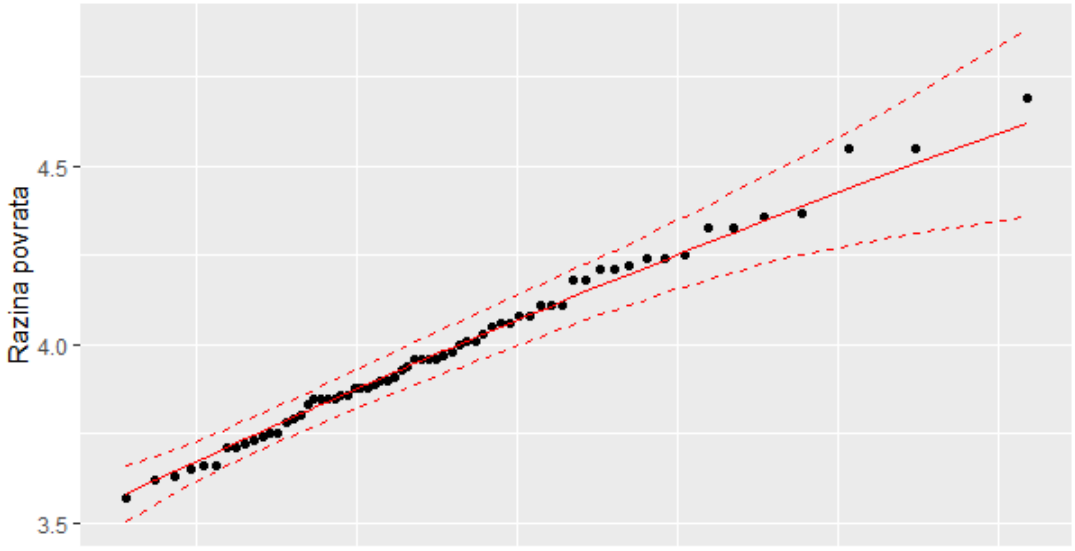
se danas takav pristup više i ne koristi, ali može biti zanimljivo usporediti rezultate tako dobivenoga modela i modela razvijenoga uz pomoć GEV familije.
Primjerice, podacima o razinama mora u Port Pirieju možemo pokušati prilagoditi distribuciju iz Gumbelove familije, koja zapravo odgovara slučaju $\xi = 0$ u GEV familiji. Koristeći funkciju `gevrFit` iz paketa `eva` dobivaju se procjene parametara $\hat{\mu} = 3.87$ te $\hat{\sigma} = 0.19$, čije su standardne pogreške jednake 0.03 i 0.02, respektivno. Tada je 95% pouzdani interval za μ jednak $[3.81, 3.93]$, a za σ $[0.15, 0.23]$.
Kao i uvijek, zanimljivo je promotriti razine povrata za različite povratne periode dobivene prilagođenim modelom. Razine povrata ponovno se računaju koristeći (1.27), ali ovoga puta u slučaju kada je $\xi = 0$. U tablici 2.3 nalaze se razine povrata za povratne periode od pet, deset, pedeset i sto godina zajedno s 95% pouzdanim intervalima dobivenima delta metodom te uz pomoć profil-vjerodostojnosti.

Povratni period	Razina povrata	Delta metoda	Profil-vjerodostojnost
5	4.16	$[4.07, 4.25]$	$[4.08, 4.26]$
10	4.30	$[4.19, 4.43]$	$[4.21, 4.43]$
50	4.63	$[4.45, 4.8]$	$[4.48, 4.82]$
100	4.77	$[4.57, 4.97]$	$[4.6, 4.99]$

Tablica 2.3: Razine povrata i 95% pouzdani intervali za razine povrata.

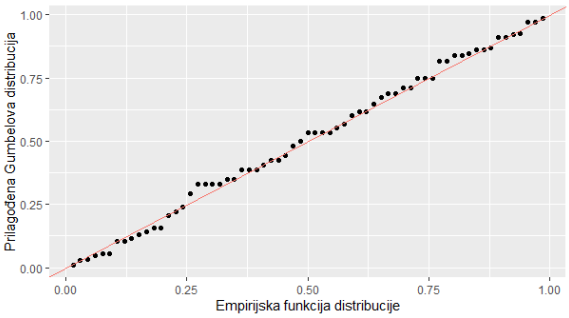


Slika 2.3: Graf kvantila prilagođenoga GEV modela.

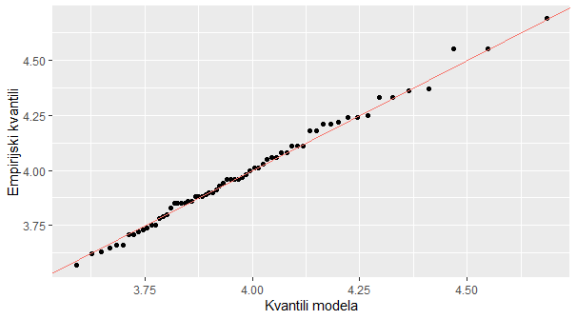


Slika 2.4: Graf razina povrata.

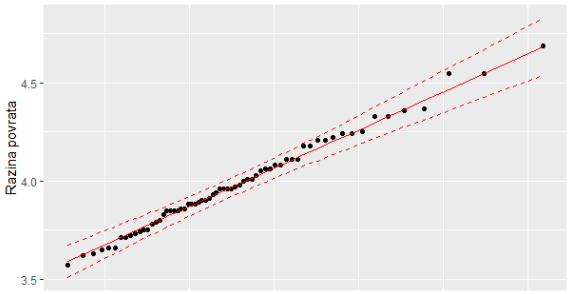
Usporedbom tablice 2.3 s tablicama 2.1 i 2.2 može se vidjeti kako su razine povrata i njihovi pouzdani intervali dobiveni modelom nastalim iz Gumbelove familije slični rezultatima dobivenima uz pomoć GEV familije, ali i da su pouzdani intervali za razine povrata u slučaju Gumbelova modela manjega raspona nego kod GEV familije. Na slici 2.5 prikazani su vjerojatnosni graf, graf kvantila te graf razina povrata prilagođenoga Gumbelova modela. Ni kod vjerojatnosnoga grafa niti kod grafa kvantila nema značajnih odstupanja točaka od dijagonale, koja je na grafovima označena crvenom bojom. Budući da se empirijski kvantili slažu s procijenjenima te da su sve vrijednosti unutar 95% pouzdanih intervala, graf razina povrata također ne sugerira da prilagođenost Gumbelova modela podacima nije dobra. Dakle, i Gumbelov može biti adekvatan za modeliranje maksimalnih morskih razina u Port Pirieju.



(a) Vjerojatnosni graf



(b) Graf kvantila



(c) Graf razina povrata

Slika 2.5: Vjerojatnosni graf, graf kvantila te graf razina povrata prilagođenoga Gumbelova modela.

Najočitija razlika između Gumbelova i GEV modela pouzdani su intervali za razine povrata koji su u slučaju Gumbelova modela manjega znatno manjega raspona, što je vidljivo i na grafu razina povrata sa slike 2.5, gdje su granice 95% pouzdanih intervala označene

isprekidanim crvenim linijama. Budući da je što manji raspon pouzdanih intervala općenito poželjan, prednost u ekstrapolaciji budućih maksimuma imao bi Gumbelov model. No, opet se javlja problem nepouzdanosti odabira prave familije. Gumbelov je model dobro prilagođen podacima, no to ne znači da Fréchetov ili Weibullov model također ne bi bili dobro prilagođeni podacima. Upravo iz tog razloga najbolje je parametar GEV familije ξ procijeniti iz samih podataka te usvojiti zaključke tako dobivenoga GEV modela.

2.2 Izdržljivost staklenih vlakana

Teorija ekstremnih vrijednosti bavi se ekstrapolacijom budućih minimalnih ili maksimalnih vrijednosti neke promatrane pojave koristeći dostupne povijesne podatke. Premda su rezultati u ovome radu predstavljeni u terminima pokušaja procjene budućih maksimalnih vrijednosti, svi su oni primjenjivi i u ekstrapolaciji budućih minimalnih vrijednosti. Kao što je već spomenuto, uz relaciju

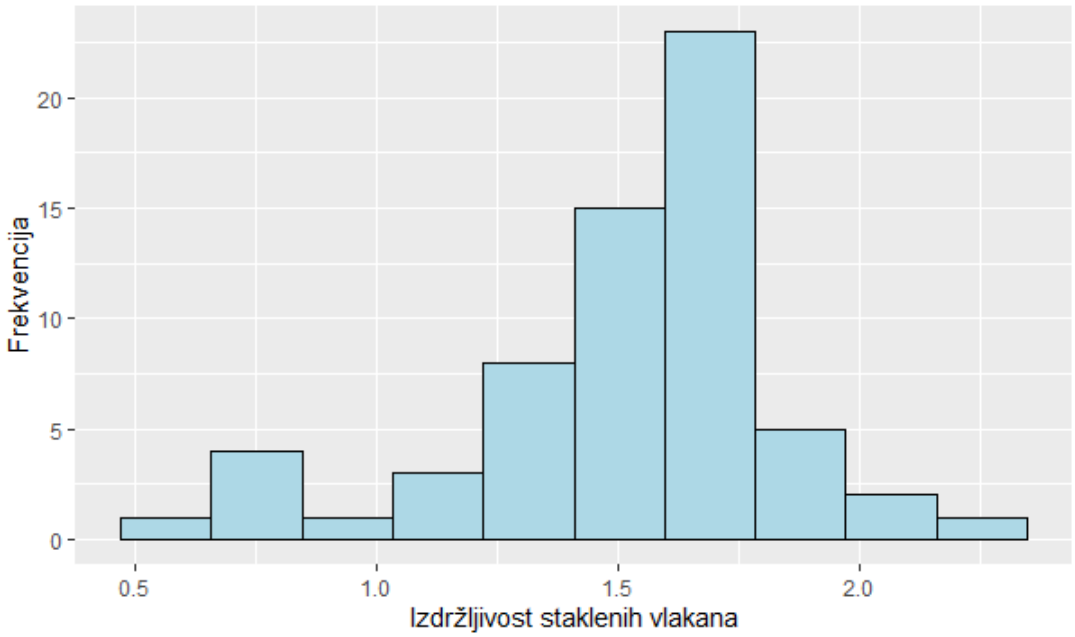
$$m_n = \min\{X_1, X_2, \ldots, X_n\} = -\max\{-X_1, -X_2, \ldots, -X_n\}, \tag{2.1}$$

moguće je umjesto maksimuma promatrati minimume blokova, kao što je slučaj u ovom odjeljku. Alternativno, može se promatrati GEV familija distribucija za minimum, koja je precizno definirana u [1].

Modeliranje minimalnih budućih vrijednosti neke pojave jako je bitno u ispitivanju pouzdanosti nekog sustava, čija izdržljivost ovisi o izdržljivosti njegove najslabije komponente. U tom je slučaju veoma važno procijeniti minimalnu snagu komponenti sustava.

Skup podataka `glass` također se može pronaći u `R`-ovu paketu `ismev`, a predstavlja popis opterećenja pod kojima su se slamala 63 staklena vlakna duljine 1.5 centimetara u eksperimentalnim uvjetima. Ako bilo kakav stakleni predmet shvatimo kao sustav izgrađen od mnoštva staklenih vlakana, tada izdržljivost cijelog predmeta ovisi o izdržljivosti njegova najslabijega vlakna. Izdržljivost, ili snaga, staklenoga vlakna predstavlja broj jedinica opterećenja koje je izazvalo slamanje tog vlakna. Dakle, u svrhu ispitivanja pouzdanosti staklenoga predmeta potrebno je procijeniti minimalnu snagu staklenoga vlakna, što se može napraviti koristeći metodu maksimuma blokova. Naime, moguće je promatrati negirane podatke zabilježene u skupu podataka `glass` te im prilagoditi GEV familiju potpuno analogno kao u odjeljku 2.1 i tako pronaći distribuciju maksimuma negativne snage staklenih vlakana, a onda će, zbog relacije (2.1), biti poznata i tražena distribucija minimuma.

Na slici 2.6 prikazan je histogram izdržljivosti staklenih vlakana zabilježenih u skupu podataka `glass`. Preciznije, na osi x nalaze se vrijednosti opterećenja kojima su vlakna bila izložena, a na osi y su opažene frekvencije slamanja vlakana, odnosno broj staklenih vlakana koja su se slomila u određenom rasponu opterećenja. Minimalna izmjerena snaga staklenoga vlakna iznosi 0.55 jedinica, a maksimalna 2.24 jedinica. S histograma je vid-



Slika 2.6: Histogram izdržljivosti staklenih vlakana.

ljivo kako se najveći broj vlakana slama u rasponu opterećenja od otprilike 1.4 do 1.8 jedinica.

Prilagodbom GEV familije negiranim podacima o izdržljivosti staklenih vlakana dobiju se procjene parametara $\hat{\mu} = -1.642$, $\hat{\sigma} = 0.273$ te $\hat{\xi} = -0.084$, pri čemu standardne pogreške iznose 0.038 za $\hat{\mu}$, 0.026 za $\hat{\sigma}$ te 0.07 za $\hat{\xi}$. Potpuno analogno kao u odjeljku 2.1 dobivaju se 95% pouzdani intervali za procijenjene parametre i to za μ $[-1.716, -1.568]$, za σ $[0.222, 0.324]$ te $[-0.221, 0.053]$ za ξ. Premda je MLE procjena parametra ξ negativna, 95% pouzdani interval za ξ sadrži i pozitivne vrijednosti pa, kao ni u odjeljku 2.1, ne možemo s velikom sigurnošću odrediti koji će tip distribucije za ekstremne vrijednosti imati promatrani maksimumi.

Povratni periodi i razine povrata u ovom slučaju imaju drugačiju interpretaciju nego u odjeljcima 1.4.1 i 2.1. U terminima testiranja pouzdanosti predmeta izgrađenog od staklenih vlakana, vrijednost razine povrata x_p koja odgovara povratnom periodu $1/p$ označava da na $1/p$ promatranih staklenih vlakana možemo očekivati da će jedno imati snagu manju od $-x_p$ jedinica (jer zapravo promatramo minimume). U tablici 2.4 prikazane su razine povrata za odgovarajuće povratne periode zajedno s 95% pouzdanim intervalima dobivenima delta metodom te uz pomoć profil-vjerodostojnosti.

Primjerice, ako promatramo sto staklenih vlakana, očekujemo kako će jedno imati snagu manju od 0.6 jedinica. Također, u tablici 2.4 vidi se kako pouzdani intervali za povratne

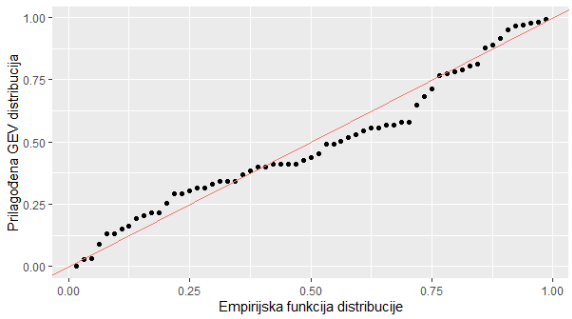
Povratni period	Razina povrata	Delta metoda	Profil-vjerodostojnost
10	-1.08	$[-1.29, -0.87]$	$[-1.12, -0.9]$
50	-0.73	$[-1.14, -0.33]$	$[-0.92, -0.36]$
100	-0.6	$[-1.1, -0.1]$	$[-0.82, -0.11]$
500	-0.32	$[-1.08, 0.43]$	$[-0.65, 0.53]$
1000	-0.21	$[-1.09, 0.66]$	$[-0.59, 0.82]$

Tablica 2.4: Razine povrata i 95% pouzdani intervali za razine povrata.

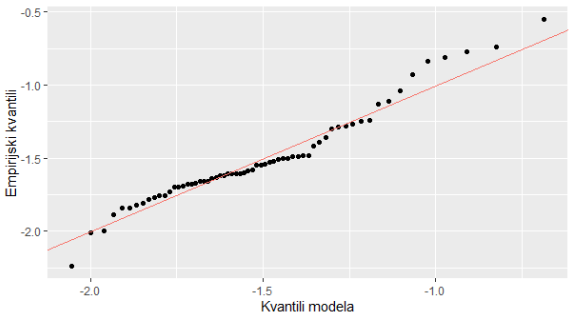
periode 500 i 1000 obuhvaćaju i pozitivne i negativne vrijednosti. Također, procijenjena razina povrata za povratni period od 5000 iznosi -0.02, što se protivi fizikalnim principima jer se staklena vlakna ne može izložiti negativnom opterećenju. Upravo su ovakve pojave vrlo važan argument za racionalnu ekstrapolaciju budućih ekstremnih vrijednosti, osobito za jako velike povratne periode, čija interpretacija mora imati smisla kada se u obzir uzme priroda promatranoga procesa ili pojave.
Naposljetku preostaje provjeriti prilagođenost GEV modela dostupnim podacima. Na slici 2.7 prikazani su vjerojatnosni graf, graf kvantila te graf razina povrata prilagođenoga GEV modela. Na vjerojatnosnome i grafu kvantila mogu se primijetiti veća odstupanja od dijagonale, koja je ponovno prikazana crvenom bojom, nego li na grafovima 2.2 i 2.3, no odstupanja se ne čine značajno velikima pa ne predstavljaju dovoljan razlog za odbacivanje prilagođenoga GEV modela. Graf razina povrata također ne sugerira lošu prilagođenost GEV modela podacima jer se razine povrata procijenjene modelom relativno dobro slažu s empirijskima, a i gotovo sve se vrijednosti nalaze unutar 95% pouzdanih intervala. Dakle, ima smisla pokušati procijeniti buduće minimalne izdržljivosti staklenih vlakana koristeći dobiveni GEV model.

2.3 Temperatura zraka u Zagrebu

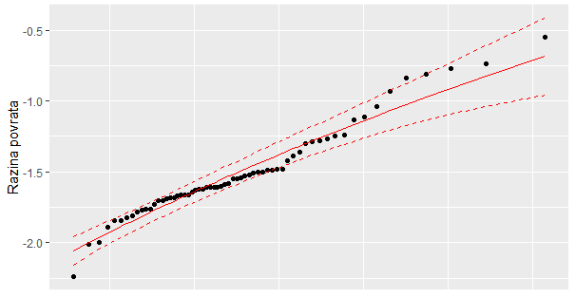
U razdoblju od 1949. do 2019. godine na meteorološkoj postaji Zagreb–Maksimir svaki je dan mjerena maksimalna temperatura zraka u Celzijevim stupnjevima. Pretpostavimo da metodu maksimuma blokova želimo primijeniti u procjeni budućih maksimalnih temperatura tijekom ljeta u Zagrebu. U tu svrhu promatramo maksimalne godišnje temperature u razdoblju 1949.-2019., pri čemu maksimalna godišnja temperatura predstavlja najveću izmjerenu vrijednost dnevnih temperatura u Zagrebu tijekom lipnja, srpnja, kolovoza i rujna. Dobiveni maksimumi prikazani su na slici 2.8.
Na slici 2.8 može se primijetiti kako s godinama rastu i maksimalne temperature, odnosno da postoji rastući trend u podacima, koji je označen crvenom bojom. Dakle, nije smisleno pokušati prilagoditi GEV model originalnim podacima. Umjesto toga, cilj je ilustrirati pristup opisan u [5], prema kojemu se maksimalna temperatura u promatranoj godini t u



(a) Vjerojatnosni graf



(b) Graf kvantila



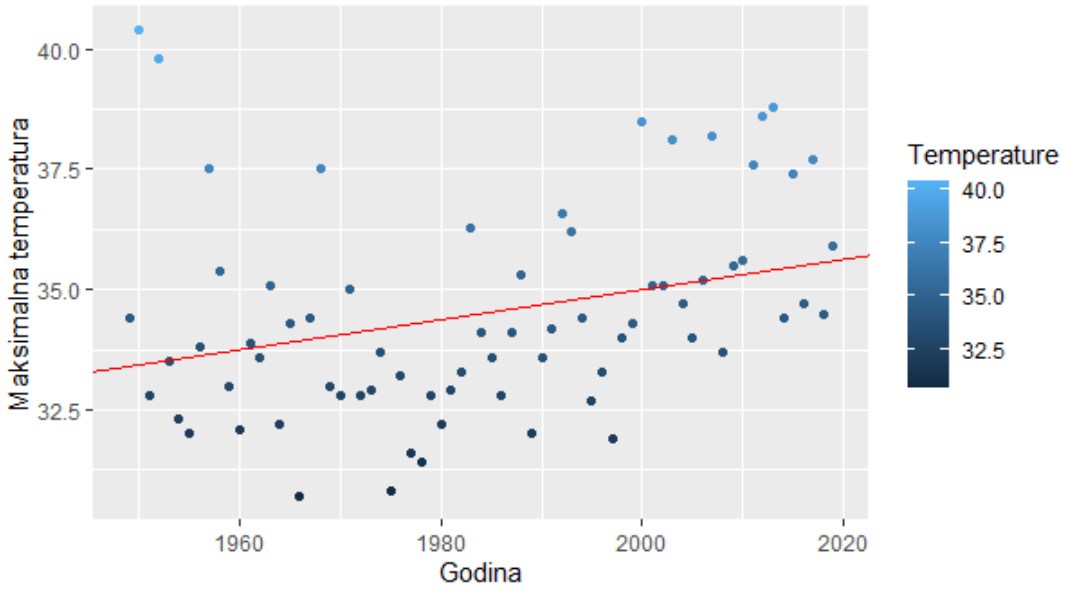
(c) Graf razina povrata

Slika 2.7: Vjerojatnosni graf, graf kvantila te graf razina povrata prilagođenoga GEV modela.

oznaci T_t može zapisati kao

$$T_t = \beta_0 + \beta_1 t + \varepsilon_t, \tag{2.2}$$

pri čemu su β_0 i β_1 parametri odgovarajućega linearnog modela, a $(\varepsilon_t)_t$ shvaćamo kao niz nezavisnih i jednakodistribuiranih slučajnih varijabli čija distribucija pripada GEV familiji. Glavna je ideja metodom najmanjih kvadrata pronaći procjene parametara $\hat{\beta}_0$ za β_0 i $\hat{\beta}_1$ za β_1 te pokušati prilagoditi odgovarajući GEV model rezidualima $\hat{\varepsilon}_t = T_t - \hat{\beta}_0 - \hat{\beta}_1 t$.
Prilagodbom linearnoga modela u R-u dobivaju se procjene parametara $\hat{\beta}_0 = -27.85671$ te $\hat{\beta}_1 = 0.03143$ i reziduali $(\hat{\varepsilon}_t)_t$, koji su prikazani na slici 2.9.
Prilagodbom GEV modela rezidualima dobivaju se MLE procjene parametara, i to $\hat{\mu} = -0.929$, $\hat{\sigma} = 1.585$, $\hat{\xi} = 0.0087$ s pripadnim standardnim pogreškama 0.21, 0.152 te 0.0825 redom. Uporabom delta metode, na standardan se način dobivaju i 95% pouzdani intervali za parametre: $[-1.34, -0.52]$ za μ, $[1.29, 1.88]$ za σ te $[-0.16, 0.17]$ za ξ. Kao i u prethodna dva primjera, ne možemo sa sigurnošću utvrditi pripadnost distribucije Gum-



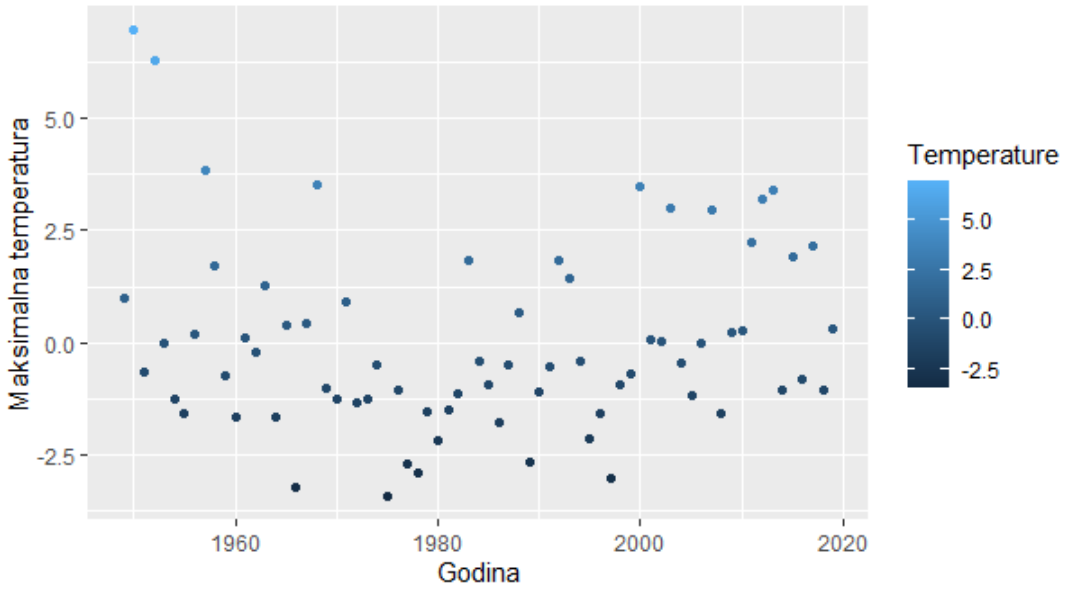
Slika 2.8: Maksimalne ljetne temperature u Zagrebu od 1949. do 2019.

belovoj, Fréchetovoj ili Weibullovoj familiji jer pouzdani interval za ξ sadrži i pozitivne i negativne vrijedosti.

Prilagođenost GEV modela dostupnim rezidualima ispituje se na standardan način: uz pomoć vjerojatnosnoga grafa, grafa kvantila te grafa razina povrata modela, koji su prikazani na slici 2.10.

Na vjerojatnosnom grafu te grafu kvantila sa slike 2.10 nema većih odstupanja od dijagonale, označene crvenom bojom. Na grafu razina povrata vidi se kako modelom procijenjene razine povrata prate empirijske bez većih odstupanja, kao i da se sve vrijednosti nalaze unutar 95% pouzdanih aproksimativnih normalnih intervala. Dakle, grafičko ispitivanje modela ne sugerira da dobiveni GEV model nije dobro prilagođen povijesnim podacima.

Kao i uvijek kod modeliranja ekstremnih vrijednosti, od najvećega je interesa upravo ekstrapolacija budućih ekstrema, u ovom slučaju maksimalnih temperatura koje se mogu očekivati u Zagrebu u nadolazećim godinama. Međutim, kako se distribucija podataka o temperaturama mijenja kroz vrijeme, terminologija povratnih perioda i razina povrata nije baš primjenjiva. No, moguće je odrediti kvantile distribucije od T_t za neki budući t pa se tako $(1-p)-$ kvantil može dobiti kao $\hat{\beta}_0+\hat{\beta}_1 t+x_p$, pri čemu je x_p odgovarajući kvantil GEV modela prilagođenoga rezidualima. Dakle, sada kvantili, zbog prisutnosti trenda, ovise i o promatranoj godini. Vrijednost p predstavlja procijenjenu vjerojatnost da će maksimalna



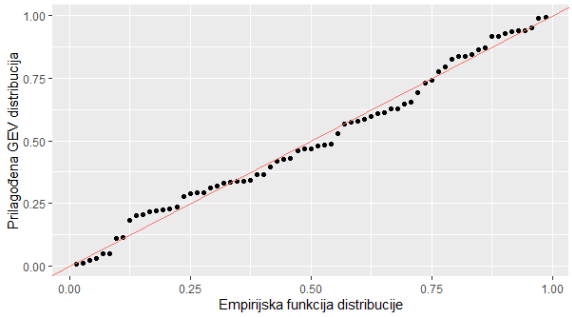
Slika 2.9: Maksimalne ljetne temperature u Zagrebu od 1949. do 2019. (bez rastućeg trenda).

Godina	$p = 0.2$	$p = 0.1$	$p = 0.02$	$p = 0.01$
2025.	37.25	38.46	41.15	42.3
2030.	37.41	38.62	41.31	42.46

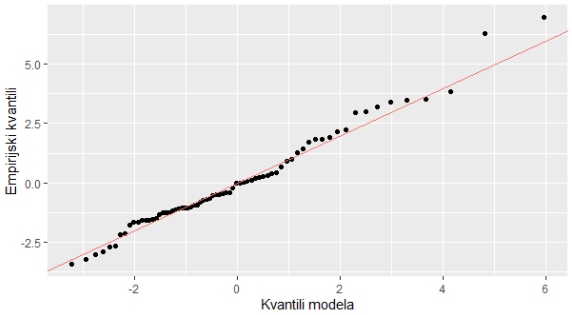
Tablica 2.5: Kvantili distribucije maksimalne ljetne temperature T_t za različite godine t.

ljetna temperatura u godini t biti veća od vrijednosti odgovarajućega kvantila. Tako dobiveni kvantili za različite vrijednosti godine t i vjerojatnosti p prikazani su u tablici 2.5. Primjerice, 2030. godine maksimalna ljetna temperatura mogla bi nadmašiti 37.41 stupnjeva s vjerojatnošću 0.2, a temperaturu od 42.46 stupnjeva s vjerojatnošću 0.01. Ipak, interpretaciji dobivenih vrijednosti valja pristupiti s oprezom jer je pretpostavljen konstantan rastući linearni trend s parametrima procijenjenim na temelju podataka iz prošlosti, koji uopće ne mora ostati nepromijenjen u budućnosti. Dakle, kao i uvijek prilikom primjene metode maksimuma blokova, vrlo je važno interpretirati rezultate vodeći računa o fizikalnim karakteristikama promatranoga procesa.

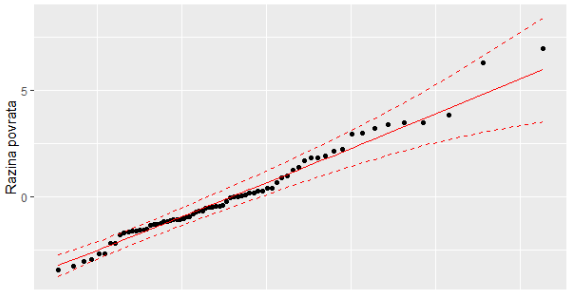
Pouzdane intervale za vrijednosti iz tablice 2.5 može se dobiti na sličan način kao i kvantile, no takva procjena ne bi uzela u obzir nepouzdanosti pri određivanju parametara β_0 i β_1. Jedan od mogućih alternativnih pristupa analizirati je originalne podatke kao da dolaze iz GEV(μ_t, σ, ξ) distribucije, pri čemu je parametar μ_t u trenutku t jednak $\beta_0 + \beta_1 t$, te



(a) Vjerojatnosni graf



(b) Graf kvantila



(c) Graf razina povrata

Slika 2.10: Vjerojatnosni graf, graf kvantila te graf razina povrata prilagođenoga GEV modela.

izračunati MLE procjenitelje za parametre β_0, β_1, σ i ξ. Na taj bi se način odmah dobili i pouzdani intervali za procijenjene kvantile uzimajući u obzir nepouzdanosti pri procjeni parametara β_0 i β_1. No, takva metodologija ipak izlazi iz okvira ovog rada. Više o alternativnim pristupima modeliranju ekstremnih vrijednosti pojava ovakva tipa može se pronaći u [1] te [5].

2.4 Alternativni pristupi modeliranju ekstremnih vrijednosti

Metoda maksimuma blokova predstavlja klasičan pristup modeliranju ekstrema i jednu od najstarijih metoda teorije ekstremnih vrijednosti. Premda su njezin povijesni značaj i korist neupitni, često se izdvajaju dva bitna nedostatka pri uporabi metode maksimuma blokova. Prva je poteškoća to što uporaba metode maksimuma blokova podrazumijeva shvaćanje

povijesnih podataka kao niza nezavisnih i jednako distribuiranih slučajnih varijabli. Iako takva pretpostavka olakšava razvoj teorijske pozadine, nažalost nije realna za većinu praktičnih problema.
Drugi je problem potencijalno traćenje informacija. Primjerice, ako promatramo samo godišnje maksimume, zanemarujemo druge ekstremne vrijednosti koje su možda izmjerene u toj godini, a bitne su za razumijevanje ponašanja proučavanoga procesa. Tako se može dogoditi da se u jednoj godini izmjere vrijednosti maksimuma koje su ekstremnije od, primjerice, maksimuma sljedeće godine, no takva pojava neće imati nikakvu ulogu u analizi maksimalnih vrijednosti jer se proučavaju samo godišnji maksimumi. Također, u povijesti su se iz raznih razloga često bilježile samo ekstremne vrijednosti, kao na primjer u slučaju morskih razina u Port Pirieju, no zahvaljujući ponajprije informatičkoj revoluciji, danas su za većinu procesa proučavanih u okvirima teorije ekstremnih vrijednosti dostupni cijeli vremenski nizovi mjerenja vrijednosti od interesa. Stoga se javila potreba za razvojem modela za ekstrapolaciju budućih ekstrema koji će maksimalno iskoristiti dostupne informacije.
Kako bi se dostupni povijesni podaci maksimalno iskoristili u ekstrapolaciji budućih ekstremnih vrijednosti, razvijeni su modeli *r* statistika najvećega reda (engl. *r Largest Order Statistic Model*) te modeli prekoračenja praga (engl. *Threshold Excess Models*). Obje metode također podrazumijevaju da dostupni podaci čine niz nezavisnih i jednako distribuiranih slučajnih varijabli.
Modeli *r* statistika najvećega reda također podrazumijevaju grupiranje podataka u blokove, no umjesto samo maksimalne (ili minimalne) vrijednosti, promatraju *r* maksimalnih vrijednosti unutar bloka. Analiziranje dodatnih maksimalnih vrijednosti može pomoći u boljoj procjeni varijacije promatranih vrijednosti unutar bloka, kao i u smanjenju varijance procijenjenih razina povrata.
Modeli prekoračenja praga u potpunosti odbacuju grupiranje dostupnih podataka u proizvoljno odabrane blokove čiji će se maksimumi onda dalje proučavati. Umjesto toga u modelima prekoračenja praga neka je vrijednost ekstremna ako prelazi unaprijed određeni prag. Primjerice, ako su dostupna mjerenja vodostaja rijeke za svaki dan, umjesto promatranja godišnjih maksimalnih vodostaja promatraju se svi vodostaji koji su viši od nekog prethodno odabranoga praga. Na taj način u prilagodbi modela sudjeluju sve općenito ekstremne vrijednosti, što na koncu dovodi do veće efikasnosti u procjeni.
Također, postoje i realističnije reprezentacije dostupnih podataka od niza nezavisnih i jednako distribuiranih slučajnih varijabli. Tako postoje tehnike za analizu ekstremnih vrijednosti stacionarnih i nestacionarnih slučajnih procesa. Stacionarni procesi definirani su definicijom 3.1.1. Kod stacionarnih procesa slučajne varijable nisu nužno međusobno nezavisne, ali su njihova stohastička svojstva homogena kroz vrijeme. S druge strane, ponašanje nestacionarnih procesa mijenja se kroz vrijeme. U primjeni je najčešće riječ o podacima s izraženom sezonalnom komponentom ili trendom. Kada je riječ o nestacionarnim pro-

cesima, glavni je cilj pokušati razviti model koji će kvantificirati nehomogenost procesa te procijeniti na koji način ta nehomogenost utječe na ekstrapolaciju budućih ekstremnih vrijednosti.

Teorija ekstremnih vrijednosti relativno je mlada statistička disciplina koja ima široke i izrazito bitne primjene u stvarnom svijetu. Vrlo jasan i minimalno tehnički kompliciran pregled u ovom odjeljku navedenih metoda modeliranja ekstremnih vrijednosti može se pronaći u Colesovoj knjizi *An Introduction to Statistical Modeling of Extreme Values* [1]. S druge strane, detaljne i izrazito matematički argumentirane studije teorije ekstremnih vrijednosti nalaze se u knjigama *Extremes and Related Properties of Random Sequences and Processes* [4], koju potpisuju Leadbetter, Lindgren i Rootzén te *Modelling Extremal Events for Insurance and Finance* [2], čiji su autori Embrechts, Klüppelberg i Mikosch.

Poglavlje 3

Dodatak

U ovome su poglavlju navedeni pojmovi i rezultati iz klasične matematičke statistike koji su eksplicitno ili implicitno korišteni u pregledu teorije o metodi maksimuma blokova. Izlaganje ovoga dijela slijedi [1] te [3], gdje se mogu pronaći svi definirani pojmovi i teorijski rezultati zajedno s dokazima.

3.1 Pregled ključnih teorijskih rezultata

Statistika je, najjednostavnije rečeno, matematička disciplina koja se bavi prikupljanjem i analizom podataka te interpretacijom dobivenih rezultata. Premda veliki dio analize podataka o promatranome procesu, ili promatranoj populaciji, predstavlja proučavanje i sistematiziranje povijesnih podataka, glavni je cilj statistike donositi zaključke o vjerojatnosnim karakteristikama procesa, odnosno proučavanjem dostupnih podataka steći uvid u općenito ponašanje promatranoga procesa te na taj način pokušati procijeniti ponašanje tog procesa u budućnosti.
Slučajni proces zapravo podrazumijeva familiju slučajnih varijabli $(X_n)_n$ definiranih na nekom vjerojatnosnom prostoru, koji standardno označavamo s $(\Omega, \mathcal{F}, \mathbb{P})$. U okvirima ovoga rada bitan je pojam stacionarnoga slučajnoga procesa spomenutoga u odjeljku 2.1.

Definicija 3.1.1. *Slučajni proces $(X_n)_{n\in\mathbb{N}}$ definiran na vjerojatnosnome prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ naziva se **stacionarnim** ako za svaki skup indeksa $\{i_1, i_2, \ldots, i_k\}$ te za proizvoljan $m \in \mathbb{N}$ vrijedi da slučajni vektori $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ i $(X_{i_1+m}, X_{i_2+m}, \ldots, X_{i_k+m})$ imaju jednaku distribuciju.*

Glavni teorijski rezultat iznesen u ovome radu bio je Fisher–Tippett–Gnedenkov teorem, prema kojemu normalizirani maksimumi dani izrazom (1.1) mogu konvergirati po distribuciji prema slučajnoj varijabli čija se funkcija distribucije nalazi u jednoj od samo tri

moguće familije: Gumbelovoj, Fréchetovoj ili Weibullovoj. Konvergencija slučajnih varijabli, odnosno vektora po distribuciji važan je pojam za razumijevanje rezultata u ovome radu, ali i u statistici i vjerojatnosti općenito te je definirana definicijom 3.1.2.

Definicija 3.1.2. *Neka je* $(\mathbf{X}_n)_n$, $n \in \mathbb{N}$ *niz slučajnih vektora u* $\mathbb{R}^k$ *s funkcijama distribucije* $F_1, F_2, \ldots$, *respektivno. Niz slučajnih vektora* $(\mathbf{X}_n)_n$ ***konvergira po distribuciji*** *prema slučajnom vektoru* $\mathbf{X}$ *s funkcijom distribucije* F *u oznaci* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *ako vrijedi*

$$F_n(\mathbf{x}) \to F(\mathbf{x}), \quad n \to \infty,$$

za sve $\mathbf{x} \in \mathbb{R}^k$ *u kojima je funkcija* F *neprekidna.*

Kada je $k = 1$, govorimo o konvergenciji slučajnih varijabli.
Također, spomenuto je kako je Fisher–Tippett–Gnedenkov teorem svojevrsni analogon centralnome graničnom teoremu. Centralni granični teorem zapravo je naziv za više teorijskih rezultata koji se bave proučavanjem graničnoga ponašanja niza parcijalnih suma $(S_n)_{n\in\mathbb{N}}$, $S_n = \sum_{k=1}^{n} X_k$, $n \in \mathbb{N}$, pri čemu je $(X_n)_{n\in\mathbb{N}}$ niz nezavisnih slučajnih varijabli, u smislu konvergencije po distribuciji. U nastavku navodimo centralni granični teorem za niz nezavisnih i jednako distribuiranih slučajnih varijabli poznat i kao Lévyjev centralni granični teorem. Vrlo detaljan pregled rezultata vezanih uz centralni granični teorem, kao i dokaz Lévyjeva teorema, mogu se pronaći u [6].

Teorem 3.1.3 (Lévy)**.** *Neka je* $(X_n)_n$ *niz nezavisnih, jednako distribuiranih slučajnih varijabli s očekivanjem* m *i varijancom* σ^2, $0 < \sigma^2 < \infty$ *te neka je* $S_n = \sum_{k=1}^{n} X_k$, $n \in \mathbb{N}$. *Tada vrijedi*

$$\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \quad kada \quad n \to \infty,$$

Lévyjev teorem zapravo kaže kako niz centriranih i normiranih parcijalnih suma konvergira po distribuciji prema slučajnoj varijabli koja ima normalnu razdiobu s očekivanjem 0 i varijancom 1.
Može se izreći i centralni granični teorem za slučajne vektore, koji je analogan teoremu 3.1.3.

Teorem 3.1.4. *Neka je* $(\mathbf{X}_n)_n \subset \mathbb{R}^k$ *niz nezavisnih, jednako distribuiranih slučajnih vektora s konačnim vektorom očekivanja* $\mathbf{m}$ *i kovarijacijskom matricom* $\mathbf{\Sigma}$ *te neka je* $\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k$. *Tada vrijedi*

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \mathbf{m}\right) \xrightarrow{d} \mathrm{MVN}_k(\mathbf{0}, \mathbf{\Sigma}) \quad kada \quad n \to \infty.$$

Drugim riječima $\sqrt{n}\left(\overline{\mathbf{X}}_n - \mathbf{m}\right)$ konvergira po distribuciji prema slučajnom vektoru s $k-$dimenzionalnom multivarijatnom normalnom razdiobom s vektorom očekivanja $\mathbf{0}$ i kovarijacijskom matricom $\mathbf{\Sigma}$.

Standardni je postupak u statističkome modeliranju donošenje zaključaka o vjerojatnosnim svojstvima populacije na temelju podataka o uzorku iz promatrane populacije. Općenito, **slučajni uzorak** naziv je za niz nezavisnih slučajnih varijabli sa zajedničkom funkcijom distribucije. Izmjerive funkcije slučajnoga uzorka nazivaju se **statistikama**. Vrijednosti koje je proces poprimio na promatranome uzorku nazivaju se realizacijama slučajnoga uzorka. Kada je poznata realizacija slučajnoga uzorka, populacijsku distribuciju promatrane pojave moguće je odrediti procjenom parametara odgovarajućega statističkog modela. Za svrhe ovoga rada dovoljno je promatrati neprekidne slučajne varijable čije funkcije gustoće postoje i pripadaju poznatoj familiji gustoća danoj statističkim modelom $\mathcal{P} = \{f(\cdot\colon \theta)\colon \theta \in \Theta\}$, pri čemu pretpostavljamo da parametar θ zadovoljava *uvjet raspoznavanja* (engl. *identifiability*), odnosno vrijedi $(\forall \theta_1, \theta_2 \in \Theta)\, \theta_1 \neq \theta_2 \implies f_{\theta_1} \neq f_{\theta_2}$. Drugim riječima, za različite vrijednosti parametra θ funkcije f_{θ_1} i f_{θ_2} predstavljaju gustoće različitih distribucija. Dakle, određivanje populacijske distribucije svodi se na određivanje parametra θ iz parametarskog prostora Θ. Parametar θ može biti skalar ili vektor parametara.
Neka je $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ slučajni uzorak duljine $n \geq 1$ iz statističkoga modela

$$\mathcal{P} = \{f(\cdot\colon \theta)\colon \theta \in \Theta\}, \tag{3.1}$$

gdje je gustoća slučajnih varijabli $X_1, \ldots, X_n$ parametrizirana skalarom θ te neka je $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ jedna realizacija slučajnoga uzorka $\mathbf{X}$. Neka je $\tau\colon \Theta \to \mathbb{R}$ izmjeriva funkcija od θ dimenzije 1. Pretpostavimo da želimo procijeniti vrijednost populacijskoga parametra $\tau(\theta)$. Ako je $\tau(\theta) = \theta$, kažemo da želimo procijeniti parametar θ. **Procjenitelj** od $\tau(\theta)$ je statistika $T = t(\mathbb{X})$ iste dimenzije kao i funkcija τ. Dakle, procjenitelj može biti bilo koja statistika iste dimenzije kao funkcija parametra koju želimo procijeniti. Vrijednost $t(\mathbf{x})$ naziva se **procjenom** od $\tau(\theta)$ na osnovi opaženoga uzorka $\mathbf{x}$. Slučajnost u uzorkovanju iz populacije dovodi do slučajnosti kod procjenitelja pa se vjerojatnosna distribucija procjenitelja inducirana ponovljenim uzorkovanjem naziva **distribucijom uzorkovanja** (engl. *sampling distribution*).
Pri procjeni vrijednosti $\tau(\theta)$ cilj je odabrati optimalni, odnosno najprecizniji procjenitelj. Procjenitelji se mogu uspoređivati s obzirom na njihovu srednjekvadratnu pogrešku i pristranost, odnosno nepristranost.

Definicija 3.1.5. *Neka je $T = \tau(\mathbf{X})$ procjenitelj od $\tau(\theta)$ kojemu komponente imaju konačnu varijancu za svaki $\theta \in \Theta$. Tada je* ***srednjekvadratna pogreška procjene*** *(engl.* mean squared error*, MSE) od $\tau(\theta)$ s T broj*

$$\mathrm{MSE}(T) = \mathbb{E}[(T - \tau(\theta))^2].$$

Statistika T najbolji je procjenitelj za $\tau(\theta)$ ako za svaki drugi procjenitelj S vrijedi da je srednjekvadratna pogreška od T manja ili jednaka od srednjekvadratne pogreške od S uniformno po $\theta \in \Theta$. Budući da srednjekvadratna pogreška mjeri varijaciju procjenitelja oko

stvarnih vrijednosti promatrane statistike, manja srednjekvadratna pogreška podrazumijeva da će procjena na osnovi opaženoga uzorka biti bliže pravoj vrijednosti.

Definicija 3.1.6. *Procjenitelj $T = t(\mathbf{X})$ za $\tau(\theta)$ je **nepristran** ako vrijedi*

$$(\forall\theta \in \Theta)\, \mathbb{E}[T] = \tau(\theta).$$

*Za procjenitelj koji nije nepristran kažemo da je **pristran**.*

Nepristran je procjenitelj onaj čije su vrijednosti u prosjeku jednake stvarnim vrijednostima promatrane statistike. Ako procjenitelj već nije nepristran, poželjno je da pristranost, definirana kao $\mathbb{E}[T] - \tau(\theta)$, bude što manja.
Još jedna mjera varijabilnosti procjenitelja je **standardna pogreška** (engl. *standard error*, SE) koja se definira kao standardna devijacija distribucije uzorkovanja procjenitelja i njezina se procjena obično može izračunati koristeći opaženi uzorak. Ako je pristranost procjenitelja zanemariva, standardna pogreška implicitno mjeri njegovu preciznost: manja standardna pogreška označava veću preciznost.
Ipak, najčešća metoda kvantifikacije pouzdanosti procjenitelja u statistici računanje je pouzdanih intervala. Pretpostavimo da želimo procijeniti sam parametar θ, odnosno $\tau(\theta) = \theta$.

Definicija 3.1.7. *Neka je $X_1, X_2, \ldots, X_n$ slučajan uzorak iz statističkoga modela $\mathcal{P} = \{f(\cdot : \theta) : \theta \in \Theta\}$. Slučajan interval $[\theta_L, \theta_U]$ naziva se $(1-\alpha)\cdot 100\%$ **pouzdanim intervalom za** θ ako vrijedi*

$$(\forall\theta \in \theta)\, \mathbb{P}(\theta_L \leq \theta \leq \theta_U) \geq 1 - \alpha.$$

Statistike $\theta_L = \theta_L(X_1, X_2, \ldots, X_n)$ i $\theta_U = \theta_U(X_1, X_2, \ldots, X_n)$ predstavljaju donju i gornju granicu pouzdanog intervala. Vrijedi $\alpha \in (0, 1)$ te se vrijednost $1 - \alpha$ naziva pouzdanošću intervalne procjene za θ.
Pouzdani intervali korisni su jer predstavljaju raspon vrijednosti unutar kojih se pravi parametar nalazi sa željenom pouzdanošću. Malene vrijednosti α podrazumijevaju veliku pouzdanost intervalne procjene, ali i intervale velikoga raspona. S druge strane, velike vrijednosti α daju intervale manjega raspona, ali je onda i pouzdanost intervalne procjene manja. U primjeni se najčešće računaju 95%, 99%, i 99.9% pouzdani intervali.
Postoji više metoda za procjenu nepoznatih parametara promatranih statističkih modela, no u ovom radu naglasak je na procjeni metodom maksimalne vjerodostojnosti koja je i korištena u procjeni parametara GEV modela. Pojam funkcije vjerodostojnosti, koja predstavlja vjerojatnost opaženih podataka kao funkciju u ovisnosti o θ, uveden je definicijom 1.4.1. Vrijednosti θ za koje je vrijednost funkcije vjerodostojnosti velika odgovaraju modelima kod kojih opaženi podaci imaju veliku vjerojatnost. Iz tog se razloga procjene traženih parametara dobivaju maksimizacijom funkcije vjerodostojnosti. Procjenitelj maksimalne vjerodostojnosti definiran je definicijom 1.4.2.

Kao što je već navedeno, MLE procjenitelji parametara GEV modela pokazali su se optimalnim izborom zbog korisnih svojstava metode maksimalne vjerodostojnosti kada je u pitanju kvantificiranje nepouzdanosti dobivenih procjenitelja izračunavanjem standardnih pogrešaka i pouzdanih intervala. Budući da su spomenuta svojstva zapravo asimptotski rezultati koji se dobiju kada veličina uzorka n teži u beskonačnost, za njihovu održivost potrebno je da model bude regularan. Definicija 3.1.8 preuzeta je iz [3] te se odnosi na jednodimenzionalne neprekidne statističke modele parametrizirane jednim skalarom, no lako se generalizira i na općenitije statističke modele.

Definicija 3.1.8. *Statistički model* $\mathcal{P} = \{f(\cdot\colon \theta)\colon \theta \in \Theta\}$ *za jednodimenzionalne razdiobe jest* ***regularan*** *ako su zadovoljeni sljedeći uvjeti*

i) Nosač $\operatorname{supp} f(\cdot\colon \theta) = \{x \in \mathbb{R}\colon f(x\colon \theta) > 0\}$ *ne ovisi o* θ.

ii) Parametarski prostor Θ *otvoreni je interval u* $\mathbb{R}$.

iii) Za sve $x \in \mathbb{R}$ *preslikavanje* $\theta \to f(x\colon \theta)$ *neprekidno je diferencijabilno na* Θ.

iv) Za ***Fisherovu informaciju*** *definiranu kao*

$$I(\theta) := \int_{\mathbb{R}} \left(\frac{\partial}{\partial \theta} \log f(x\colon \theta) \right)^2 f(x\colon \theta) \mathrm{d}x$$

vrijedi $0 < I(\theta) < \infty$ *za sve* $\theta \in \Theta$.

v) Za svaki $\theta \in \Theta$ *vrijedi*

$$0 = \frac{d}{d\theta} \int_{\mathbb{R}} f(x\colon \theta) = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} f(x\colon \theta) \mathrm{d}x.$$

Teorem 3.1.9. *Neka je* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *slučajni uzorak iz statističkoga modela* $\mathcal{P} = \{f(\cdot\colon \theta)\colon \theta \in \Theta\}$ *te* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *jedna njegova realizacija. Nadalje, neka je* $\hat{\theta}$ *MLE procjenitelj parametra modela* θ*, koji je dimenzije* d*. Tada, ako su zadovoljeni potrebni uvjeti regularnosti, vrijedi*

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathrm{MVN}_d\left(0, \left(\frac{I_E(\theta)}{n}\right)^{-1}\right) \quad \textit{kada} \quad n \to \infty,$$

pri čemu je $I_E(\theta)$ *matrica* ***Fisherove informacije*** *s obzirom na uzorak* $\mathbf{X}$ *koja se definira kao*

$$\left[I_E(\theta)\right]_{i,j} = \mathbb{E}\left[- \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{\mathbf{X}}(\mathbf{X}\colon \theta) \right], \quad i, j = 1, 2, \ldots, d,$$

gdje za funkciju gustoće od $\mathbf{X}$ *u oznaci* $f_{\mathbf{X}}$ *vrijedi* $f_{\mathbf{X}}(\mathbf{x}\colon \theta) = \prod_{i=1}^{n} f(x_i\colon \theta)$, $i = 1, \ldots, n$.

Napomena 3.1.10. *i) U okvirima ovoga rada korisna je bila interpretacija teorema 3.1.9 koja glasi da za velike $n \in \mathbb{N}$ vrijedi*

$$\hat{\theta} \overset{d}{\approx} \mathrm{MVN}_d\Big(\theta, (I_E(\theta))^{-1}\Big),$$

odnosno MLE procjenitelj za velike je $n \in \mathbb{N}$ aproksimativno normalan s $d-$dimenzionalnom multivarijatnom normalnom razdiobom očekivanja θ i kovarijacijske matrice jednake $(I_E(\theta))^{-1}$.

ii) Vrijednost $\frac{1}{n} I_E(\theta)$ zapravo je Fisherova informacija za X_1 u oznaci $I_1(\theta)$, koja se definira kao

$$\Big[I_1(\theta)\Big]_{i,j} = \mathbb{E}\Big[- \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(X_1 : \theta)\Big], \quad i, j = 1, 2, \ldots, d.$$

Teorem 3.1.9 koristi se za računanje pouzdanih intervala individualnih komponenti parametra $\theta = (\theta_1, \ldots, \theta_d)$. Ako elemente matrice inverzne Fisherovoj informaciji označimo s $\psi_{i,j}$, iz svojstava multivarijatne normalne razdiobe slijedi da za velike $n \in \mathbb{N}$ vrijedi

$$\hat{\theta}_i \overset{d}{\approx} N(\theta_i, \psi_{i,i}),$$

odnosno procjenitelj $\hat{\theta}_i$ ima aproksimativnu normalnu razdiobu s parametrima θ_i i $\psi_{i,i}$ za velike $n \in \mathbb{N}$. Dakle, granice za aproksimativni $(1 - \alpha) \cdot 100\%$ pouzdani interval za θ_i dobiju se kao

$$\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\psi_{i,i}},$$

pri čemu $z_{\alpha/2}$ označava $(1 - \alpha/2)$ kvantil standardne normalne distribucije.
Budući da prava vrijednost parametra θ najčešće nije poznata, umjesto Fisherove informacije promatra se opažena informacijska matrica I_O čiji su elementi dani kao

$$\Big[I_O(\theta)\Big]_{i,j} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f_{\mathbf{X}}(\mathbf{X} : \theta), \quad i, j = 1, 2, \ldots, d,$$

za $\theta = \hat{\theta}$. Ako elemente inverza opažene informacijske matrice označimo s $\tilde{\psi}_{i,j}$, tada se granice $(1 - \alpha) \cdot 100\%$ pouzdanih intervala za θ_i dobivaju kao

$$\hat{\theta}_i \pm z_{\alpha/2} \sqrt{\tilde{\psi}_{i,i}}.$$

Zanimljivo je kako su intervali dobiveni uz pomoć opažene informacijske matrice često precizniji od onih dobivenih koristeći Fisherovu informaciju.
Osim procjenjivanja samog parametra modela θ, često se javlja potreba za procjenjivanjem i donošenjem zaključaka o raznim funkcijama koje ovise o θ. Teoremi 3.1.11 i 3.1.12 navode korisne rezultate za proučavanje MLE procjena funkcija parametra θ.

Teorem 3.1.11. *Neka je $\hat{\theta}$ procjenitelj metodom maksimalne vjerodostojnosti od θ na Θ i neka je $\tau(\theta)$ funkcija od θ. Tada je $\tau(\hat{\theta})$ procjenitelj metodom maksimalne vjerodostojnosti od $\tau(\theta)$ na $\tau(\Theta)$.*

Teorem 3.1.12. *Neka je $\hat{\theta}$ procjenitelj $d-$dimenzionalnoga parametra modela θ dobiven metodom maksimalne vjerodostojnosti. Ako je $T = \tau(\theta)$ skalarna funkcija, tada za njezin MLE procjenitelj $\hat{T}$ vrijedi*

$$\sqrt{n}(\hat{T} - \tau(\theta)) \xrightarrow{d} N\Big(0, \nabla T^T I_1^{-1}(\theta) \nabla T\Big) \quad kada \quad n \to \infty,$$

pri čemu je

$$\nabla T = \left[\frac{\partial T}{\partial \theta_1}, \cdots, \frac{\partial T}{\partial \theta_d}\right]$$

evaluirana u $\hat{\theta}$.

Teorem 3.1.12 obično se naziva **delta metodom**, a u okvirima ovoga rada bio je osobito koristan u računanju pouzdanih intervala za funkcije parametara modela.
Još jedna metoda ispitivanja preciznosti procjenitelja izračun je pouzdanih intervala uz pomoć funkcije profil-vjerodostojnosti. Neka je $\theta \in \Theta$ vektor parametara, koji parametriziraju statistički model (3.1), dimenzije d. Funkcija log-vjerodostojnosti za θ može se zapisati kao $l(\theta_i, \theta_{-i})$, pri čemu θ_{-i} označava sve komponente od θ izuzev θ_i.

Definicija 3.1.13. *Funkcija **profil-vjerodostojnosti** za θ_i definira se kao*

$$l_p(\theta_i) = \max_{\theta_{-i}} l(\theta_i, \theta_{-i}).$$

Dakle, za svaku komponentu θ_i vektora θ profil-vjerodostojnost zapravo predstavlja maksimiziranu log-vjerodostojnost s obzirom na preostale komponente od θ.
Definiciju 3.1.13 moguće je poopćiti i na slučaj kada je vektor parametara θ potrebno podijeliti u dvije komponente $\left(\theta^{(1)}, \theta^{(2)}\right)$, gdje je $\theta^{(1)}$ vektor parametara dimenzije k koji nas zanima, a $\theta^{(2)}$ predstavlja preostalih $d - k$ komponenti od θ. Tada je profil-vjerodostojnost za $\theta^{(1)}$ jednaka

$$l_p\Big(\theta^{(1)}\Big) = \max_{\theta^{(2)}} l\Big(\theta^{(1)}, \theta^{(2)}\Big).$$

Teorem 3.1.14. *Neka je $X_1, X_2, \ldots, X_n$ slučajni uzorak iz statističkoga modela $\mathcal{P} = \{f(\cdot : \theta) : \theta \in \Theta\}$ te $x_1, x_2, \ldots, x_n$ jedna njegova realizacija. Nadalje, neka je $\hat{\theta}$ MLE procjenitelj*

d−dimenzionalnoga parametra $\theta = \left(\theta^{(1)}, \theta^{(2)}\right)$, *gdje je* $\theta^{(1)}$ *k−dimenzionalni podskup od* θ. *Tada, ako su ispunjeni uvjeti regularnosti, vrijedi*

$$D_p\left(\theta^{(1)}\right) = 2\left(l(\hat{\theta}) - l_p(\theta^{(1)})\right) \xrightarrow{d} \chi^2_k \quad kada \quad n \to \infty,$$

odnosno $D_p\left(\theta^{(1)}\right)$ *konvergira po distribuciji prema slučajnoj varijabli* χ^2 *razdiobe s k stupnjeva slobode kada* $n \to \infty$.

Napomena 3.1.15. *i) Još jedna metoda kvantifikacije nepouzdanosti MLE procjenitelja je funkcija* ***devijance*** *definirana kao* $D(\theta) = 2\left(l(\hat{\theta}) - l(\theta)\right)$, *pri čemu je* θ *parametar promatranoga statističkog modela, a* $\hat{\theta}$ *njegov MLE procjenitelj. Vrijednosti* θ *koje imaju malu devijancu odgovaraju modelima s velikom vjerodostojnošću. Prema [1] za devijancu se može pokazati da za velike* $n \in \mathbb{N}$ *ima aproksimativnu* χ^2 *razdiobu s onoliko stupnjeva slobode kolika je dimenzija parametra* θ. *Premda je samo računanje pouzdanih intervala složenije, dobivene procjene obično su preciznije od onih dobivenih koristeći asimptotsku normalnost MLE procjenitelja.*

ii) Teorem 3.1.14 često se koristi za računanje aproksimativnih pouzdanih intervala jedne komponente θ_i. *Aproksimativni* $(1-\alpha)\cdot 100\%$ *pouzdani interval dan je skupom* $C_\alpha = \{\theta_i \colon D_p(\theta_i) \leq c_\alpha\}$, *pri čemu* c_α *označava* $(1-\alpha)$ *kvantil* χ^2_1 *distribucije.*

U završetku ovoga odjeljka slijede teoremi koji se bave nužnim i dovoljnim uvjetima za pripadnost Fréchetovoj, Weibullovoj ili Gumbelovoj domeni atrakcije, a služe kao nadopuna odjeljku 1.2.

Definicija 3.1.16. *Generalizirani inverz funkcije distribucije F*

$$F^{-1}(q) = \inf\{x \in \mathbb{R} \colon F(x) \geq q\}, \quad 0 < q < 1,$$

naziva se ***kvantilnom funkcijom*** *funkcije distribucije F, a s* $x_q = F^{-1}(q)$ *definira se* ***q-ti kvantil*** *od F.*

Teorem 3.1.17 (Fréchetova domena atrakcije za maksimum)**.** *Funkcija distribucije F pripada Fréchetovoj domeni atrakcije za maksimum ako i samo ako je* $\overline{F}(x) = x^{-\alpha}L(x), \forall x \in \mathbb{R}$, $\alpha > 0$ *i za neku sporo varirajuću funkciju L te pišemo* $F \in MDA(G_F)$. *Ako vrijedi* $F \in MDA(G_F)$, *tada*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G_F,$$

pri čemu je $a_n = F^{-1}(1 - n^{-1})$ *i* $b_n = 0$.

Teorem 3.1.18 (Weibullova domena atrakcije za maksimum)**.** *Funkcija distribucije F pripada Weibullovoj domeni atrakcije za maksimum ako i samo ako je $x_+ < \infty$ i $\overline{F}(x_+ - x^{-1}) = x^{-\alpha}L(x), \forall x \in \mathbb{R}$, $\alpha > 0$ i za neku sporo varirajuću funkciju L te pišemo $F \in MDA(G_W)$. Ako vrijedi $F \in MDA(G_W)$, tada*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G_W,$$

pri čemu je $a_n = x_+ - F^{-1}(1 - n^{-1})$ i $b_n = x_+$.

Teorem 3.1.19 (Gumbelova domena atrakcije za maksimum)**.** *Funkcija distribucije F za koju vrijedi $x_+ \leq \infty$ pripada Gumbelovoj domeni atrakcije za maksimum ako i samo ako postoji neki $z < x_+$ takav da se $\overline{F}$ može zapisati kao*

$$\overline{F}(x) = c(x) \exp\left\{ - \int_z^x \frac{g(t)}{a(t)} \, \mathrm{d}t \right\}, \quad z < x < x_+, \tag{3.2}$$

gdje su c i g izmjerive funkcije za koje vrijedi $c(x) \to c > 0$, $g(x) \to 1$ kada $x \uparrow x_+$, a $a(x)$ pozitivna, apsolutno neprekidna (s obzirom na Lebesgueovu mjeru) funkcija s gustoćom $a'(x)$ tako da $\lim_{x \uparrow x_+} a'(x) = 0$. Kao normirajuće konstante mogu se uzeti

$$b_n = F^{-1}(1 - n^{-1}) \quad te \quad a_n = a(b_n).$$

Za funkciju a može se odabrati

$$a(x) = \int_x^{x_+} \frac{\overline{F}(t)}{\overline{F}(x)} \mathrm{d}t, \quad x < x_+.$$

Bibliografija

[1] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer–Verlag London, 2001.

[2] P. Embrechts, C. Klüppelberg i T. Mikosch, *Modelling Extremal Events for Insurance and Finance*, Springer–Verlag, 1997.

[3] M. Huzak, *Matematička statistika: skripta s predavanja*, 2020., `https://web.math.pmf.unizg.hr/nastava/ms/index.php?sadrzaj=predavanja.php`, (rujan 2022.).

[4] M. R. Leadbetter, G. Lindgren i H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes*, Springer–Verlag New York Inc., 1983.

[5] A. Guillou P. Ribereau i P. Naveau, *Estimating return levels from maxima of non-stationary random sequences using the Generalized PWM method*, (2008.), `https://hal.archives-ouvertes.fr/hal-03193808/document`, (rujan 2022.).

[6] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002.

Sažetak

Cilj ovoga rada predstaviti je teorijske rezultate potrebne za razumijevanje statističkoga modeliranja ekstremnih vrijednosti korištenjem metode maksimuma blokova te ilustrirati primjenu metode maksimuma blokova na primjerima iz stvarnoga svijeta.
Na samom je početku rada opisana reprezentacija dostupnih podataka kao nezavisnih i jednako distribuiranih slučajnih varijabli, kao i postupak pronalaska maksimalnih vrijednosti unutar blokova. Glavni cilj metode maksimuma blokova određivanje je distribucije tako dobivenih maksimuma, o čemu govori i Fisher–Tippett–Gnedenkov teorem, čiji iskaz i dokaz predstavljaju središnji rezultat ovoga rada. Važnost Fisher–Tippett–Gnedenkova teorema ogleda se u tome što navodi sve moguće familije funkcija distribucije kojima može pripadati funkcija distribucije promatranih maksimuma, a to su Gumbelova, Fréchetova i Weibullova familija, i to bez obzira na distribuciju početnih podataka. Dan je i kratak osvrt na Gumbelovu, Fréchetovu i Weibullovu domenu atrakcije zajedno s primjerima poznatih distribucija koje pripadaju pojedinoj domeni.
Nadalje, predstavljen je pojam GEV familije funkcija distribucije kao parametrizacije koja ujedinjuje Gumbelovu, Fréchetovu i Weibullovu familiju te opisan način prilagodbe optimalnoga GEV modela dostupnim podacima. U samoj procjeni parametara modela, kao i donošenju zaključaka o njihovoj preciznosti, ključna je bila metoda maksimalne vjerodostojnosti sa svojim asimptotskim svojsvima.
U ekstrapolaciji budućih ekstremnih vrijednosti iz GEV modela ključni pojmovi bili su razina povrata te povratni period. Razina povrata je vrijednost za koju se očekuje da bi u svakome bloku mogla biti nadmašena s vjerojatnošću p, gdje je p vrijednost recipročna željenom povratnom periodu.
Primjena metode maksimuma blokova ilustrirana je na primjerima iz stvarnoga svijeta, i to u procjeni maksimalnih morskih razina, procjeni minimalne izdržljivosti staklenih vlakana te analizi maksimalnih ljetnih temperatura u Zagrebu.
Naposljetku je dan i pregled osnovnih pojmova klasične matematičke statistike koji su ključni za razumijevanje rezultata iznesenih u radu.

Summary

The prime objective of this thesis is to give an introduction into the statistical theory that is crucial for understanding extreme value modelling via block maxima method, as well as to try and illustrate the use of said block maxima method on real-life data.
At first, it is described how to represent the available data as a sequence of independent, identically distributed random variables and how block maxima are actually found. The main objective of block maxima method consists of determining the distribution of block maxima, which can sometimes be achieved by applying Fisher–Tippett–Gnedenko theorem. Fisher–Tippett–Gnedenko theorem, along with its proof, represents the central theoretical result of this thesis. The importance of Fisher–Tippet–Gnedenko theorem lies in the fact that it provides all distribution families that can appear as limits for the block maxima distributions, regardless of the distribution of the initial data. Said distribution families are widely known as Gumbel, Fréchet and Weibull families. This thesis also contains a brief review of maximum domains of attraction of Gumbel's, Fréchet's and Weibull's distributions, along with examples of some famous distribution functions that belong to one of said maximum domains of attraction.
Next, GEV family is introduced as an unification of Gumbel's, Fréchet's and Weibull's family and a description is given on how to adapt the right GEV model to the available data. The key to estimating model parameters, as well as to determining the confidence of obtained estimates, surely was maximum likelihood estimation with its asymptotic properties.
Moreover, the terms return level and return period are proposed as fundamental in understanding the extrapolated extreme values. Return level denotes a value that might be exceeded in any particular block with probability p, which corresponds to the reciprocal value of the observed return period.
The application of block maxima method is illustrated with the analysis of three real-life datasets: the first one contains data about annual maximum sea-levels, the second dataset consists of measurements of glass fiber breaking strengths, and the third one includes measurements of maximum daily temperatures in Zagreb, Croatia.
At last, this thesis also contains a short overview of classical statistical results that are crucial for understanding the presented extreme value theory results.

Životopis

Rođena sam 21. veljače 1999. u Mostaru. Pohađala sam Prvu osnovnu školu, a nakon toga i Gimnaziju fra Dominika Mandića u rodnome Širokom Brijegu. Tijekom osnovne i srednje škole rado sam sudjelovala i ostvarivala uspjehe na brojnim matematičkim natjecanjima pa sam po završetku srednje škole 2017. upisala Preddiplomski sveučilišni studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Preddiplomski studij završila sam 2020. godine i nastavila obrazovanje upisom diplomskoga sveučilišnoga studija Matematička statistika na istom fakultetu.