Mapiranje zaraze koronavirusom u Hrvatskoj

Živaljić, David

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

DAVID ŽIVALJIĆ

MAPIRANJE ZARAZE KORONAVIRUSOM U HRVATSKOJ

DIPLOMSKI RAD

Split, srpanj 2023.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

MAPIRANJE ZARAZE KORONAVIRUSOM U HRVATSKOJ

DIPLOMSKI RAD

Student:
David Živaljić

Mentorica:
doc.dr.sc. Vesna Gotovac
Đogaš

Split, srpanj 2023.

TEMELJNA DOKUMENTACIJSKA KARTICA

Diplomski rad

Sveučilište u Splitu

Prirodoslovno-matematički fakultet

Odjel za matematiku

Ruđera Boškovića 33, 21000 Split, Hrvatska

MAPIRANJE ZARAZE KORONAVIRUSOM U HRVATSKOJ

David Živaljić

Sažetak:

Cilj ovog rada je bio otkrivanje županija u Hrvatskoj visokog rizika od zaraze koronavirusom u prvoj polovini 2021. godine uz pomoć metoda prostorne statistike. U svrhu procjene rizika od zaraze koristimo standardizirani omjer smrtnosti, te tri različita empirijska Bayesova procjenitelja. Za procjenu smrtnost nam je poslužila stopa smrtnosti. Također, promatramo utjecaj raznih kovarijabli na mortaitet. Dolazimo do sljedećeg zaključka. Što se tiče morbiditeta, najveći rizik ima Primorsko-Goranska županija, a slijede je Međimurska i Varaždinska. Što se tiče mortaliteta, najveći rizik ima Karlovačka županija, a slijede Varaždinska, Grad Zagreb, Ličko-Senjeska i Osječo-Baranjska. Vodeći se dobnom podjelom stanovništva, pokušali smo pokazati da je populacija starijih od 60 statistički značajna varijabla za povećan mortalitet od korone, ali nismo uspjeli postići konvergenciju Markovljevog lanca. Nakon toga smo prikazali prostornu zagađenost zraka u prvoj polovici 2021. godine. Model kojeg smo proveli ukazao je na statističku značajnost uz konvergenciju Markovljevog lanca. U županijama gdje je zagađenost zraka bila najveća, ponalazimo istog glavnog zagađivača, PM2.5. Očito, kako te sitne čestice PM2.5 uzrokuju probleme u dišnom sustavu, uz kombinaciju sa koronavirusom dolazi do više smrtnih slučajeva.

Ključne riječi:

Koronavirus, statistički modeli, mapiranje, distribucije, procjenitelj

Rad je pohranjen u knjižnici Prirodoslovno-matematičkog fakultet, Sveučilišta u Splitu

Rad sadrži:

53 stranice, 31 grafičkih prikaza, 15 literaturnih navoda. Izvornih je na hrvatskom jeziku.

Mentorica: *doc.dr.sc. Vesna Gotovac Đogaš*

Ocjenjivači:

doc.dr.sc. Vesna Gotovac Đogaš

prof. dr. sc. Milica Klaričić Bakula

dr. sc. Ivan Jelić

Rad prihvaćen: srpanj, 2023.

BASIC DOCUMENTATION CARD

Graduation thesis

University of Split

Faculty of Science

Department of mathematics

Ruđera Boškovića 33, 21000 Split, Croatia

MAPPING CORONAVIRUS INFECTION IN CROATIA

David Živaljić

Abstract:

The aim of this work was to detect counties in Croatia with a high risk of infection with the coronavirus in the first half of 2021 with the help of spatial statistics methods. For the purpose of assessing the risk of infection, we use the standardized mortality ratio and three different empirical Bayesian estimators. We used the death rate to estimate mortality. We also observe the influence of various covariables on mortality. We come to the following conclusion. In terms of morbidity, Primorsko-Goranska County has the highest risk, followed by Međimurje and Varaždin. As for mortality, Karlovac County has the highest risk, followed by Varaždin County, the City of Zagreb, Ličko-Senjeska and Osječo-Baranjska. Guided by the age distribution of the population, we tried to show that the population over 60 is a statistically significant variable for increased mortality from corona, but we failed to achieve the convergence of the Markov chain. After that, we presented spatial air pollution in the first half of 2021. The model we implemented indicated statistical significance with the convergence of the Markov chain. In counties where air pollution was the highest, we find the same main pollutant, PM2.5. Obviously, as these tiny PM2.5 particles cause problems in the respiratory system, in combination with the coronavirus, more deaths occur.

Key words:

Coronavirus, statistical models, mapping, distributions, estimator

Thesis deposited in library of Faculty of science, University of Split

Thesis consist of:

53 pages, 31 graphical representation, 15 literary references. Original language: Croatian.

Supervisor: *assisstant professor Vesna Gotovac Đogaš*

Reviewers:

assisstant professor Vesna Gotovac Đogaš

professor Milica Klaričić Bakula

Ivan Jelić, phd

Thesis accepted: July, 2023

Uvod

Nit vodilja prostorne analize je takozvani Toblerov zakon koji glasi ovako: *"Sve je povezano sa svime, ali bliske stvari su povezanije jedna s drugom."*

Prostorna statistika je samostalna grana statistike koja proučava metode prikupljanja i analize podataka koristeći njihova topološka (geografska) svojstva. Široko je primjenjiva i započela je razvoj iz različitih disciplina. Dovoljno je nabrojiti samo tri sljedeća problema u kojoj se primjenjuje kako bi dobili dojam široke rasprostranjenosti: analiza položaja dijelova čipova, analiza i predviđanje položaja galaksija u svemiru, epidemiologija.

U ovom radu razmatrat ćemo primjenu prostorne statistike u epidemiologiji. Preciznije govoreći, razmatrat ćemo mapiranje zaraze koronavirusom po hrvatskim županijama u prvoj polovini 2021. godine. Općenito, cilj mapiranja zaraze je prikazati prostornu distribuciju rizika koji se može manifesirati na dva načina: mortalitet i mobiditet(broj ljudi koji su oboljeli). Nakon prikazane distribucije mogu se učiti regije koje imaju povećani rizik u odnosu na ostale. Sukladno tome možemo poduzeti određene mjere zaštite ili pak otkriti razlog zašto dolazi do povećanog rizika.

Cilj ovog rada nije pružiti detaljan i opsežan opis svih metoda koje se trenutno koriste u prostornoj epidemijiologiji, već pokazati one koje se koriste u širokom obujmu.

Glavni zadatak ovog rada je primjena svih iznesenih metoda uporabom

računala, programskog jezika **R** i prostornih podataka korištenjem takozvanog **GIS-a** [6], računalni sustav koji analizira i prikazuje geografski referencirane informacije.

U prvom poglavlju navodimo osnovne pojmove potrebne za razumjevanje rada.

U drugom poglavlju kratko se navode osnove mapiranja zaraze uz prikaz učitanih podataka te dobnu raspodjelu za Hrvatsku po županijama. Pokazuje se podjela podataka na slojeve čiji cilj može biti uklanjanje zbunjujućih varijabli.

U trećem poglavlju se iznose razne metode te komentiraju njihove prednosti i nedostatke. Nakon toga se pokušava napraviti što zaglađeniji prikaz relativnog rizika. Sve metode provodimo u statističkom programu R.

U četvrtom poglavlju se definira matrica susjedstva županija i uvodi prostorno strukturirani statistički modeli. U ovim modelima rizik svake županije ovisi o susjednim županijama blizu kojih se nalazi. Kratko se opisuje model linearne regresije gdje se relativni rizik modelira kao linearana funkcija kovarijebli (u našem slučaju razmatramo kao kovarijable količinu starije populacije, te zagađenost zraka u županijama), te se pokazuje princip kako otkriti statistički značajne kovarijable. Specifično za Hrvatsku, pokušati će se pronaći statistički značajna kovarijabla koja povećava rizik od bolesti koronavirusa. Simulacije potrebne za Bayesovsku procjenu parametara modela će se provoditi korištenjem Monte Carlo Markovljevih lanaca (vidi [7]) te ćemo ih izvršavati pomoću vanjskog programa **WinBUGS** [15] koji je razvijen na Cambridgeu, a rezultate ćemo spremati u posebne datoteke prilagođene za taj program.

Sadržaj

Poglavlje 1

Osnovne definicije

U ovom poglavlju ćemo iznijeti definicije pojmova koje ćemo koristiti u daljnjem tekstu.

1.1 Slučajne varijable (vektori)

Definicija 1.1 *Neka je* $(\Omega, \mathcal{F}, \mathbb{P})$ *vjerojatnostni prostor,* $(\Omega, \mathcal{F})$ *i* $\left(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k)\right)$, $k \geq 1$ *izmjerivi prostori te* $\mathcal{B}(\mathbb{R}^k)$ *Borelova sigma algebra na* $\mathbb{R}^k$. ***Slučajna varijabla*** *(**Slučajni vektor**) je izmjerivo preslikavanje* $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))\ \left(\mathbb{R}^k, \mathcal{B}(\mathbb{R})^k\right)$, *tj. takvo da vrijedi :*

$$X^{-1}(B) \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R})\ \left(\mathcal{B}(\mathbb{R}^k)\right).$$

Definicija 1.2 *Neka je* $X : \Omega \to \mathbb{R}^k$ *slučajan vektor na* $(\Omega, \mathcal{F}, \mathbb{P})$. *Induciranu vjerojtnost* $\mathbb{P}_X : \mathcal{B}(\mathbb{R}^k) \to [0, 1]$ *na* $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ *definiranu s:*

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$$

nazivamo ***distribucija ili zakon razdiobe od*** X.

1.1. Slučajne varijable (vektori)

Definicija 1.3 *Neka je X slučajan vektor sa zakonom razdiobe $\mathbb{P}_X$. Funkciju $F = F_X : \mathbb{R}^k \to [0,1]$ definiranu s*

$$F_X(x) := \mathbb{P}_X(\langle -\infty, x]), x \in \mathbb{R}^k,$$

nazivamo ***funkcija distribucije od*** X.

Razlikujemo 2 osnovna tipa slučajnih varijabli (vektora).

Definicija 1.4 *Kažemo da je X* ***neprekidna slučajna varijabla*** *ukoliko je njena funkcija distribucije F_X apsolutno neprekidna, tj. ako postoji funkcija $f : \mathbb{R}^k \to [0, +\infty\rangle$ takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(y) d\lambda(y)$$

Funkciju f nazivamo ***funkcija gustoće od*** X.

Definicija 1.5 *Kažemo da je X* ***diskretna slučajna varijabla*** *ukoliko postoji prebrojiv podskup $D \in \mathcal{B}(\mathbb{R}^k)$ takav da je $\mathbb{P}_X(D) = 1$.*

Često se u praksi diskretna slučajna varijabla X definira eksplicitnim navođenjem distribucije P_X koju ćemo navoditi tablično, a neprekidna pomoću funkcije gustoće.

Sada za slučajne varijable uvodimo sljedeće pojmove.
Neka je X slučajna varijabla na $(\Omega, \mathcal{F}, \mathbb{P})$. Vrijedi $X = X^+ - X^-$, gdje je

$$X^+ := \max\{X, 0\} \geq 0, X^- = \max\{-X, 0\} \geq 0$$

Definicija 1.6 *Kažemo da je slučajna varijabla $X = X^+ - X^-$ ima* ***matematičko očekivanje*** *ukoliko je barem jedan od integrala*

$$EX^+ := \int_\Omega X^+ d\mathbb{P}, EX^- := \int_\Omega X^- d\mathbb{P}$$

konačan. U tom je slučaju matematičko očekivanje od X, u oznaci EX, jednako

$$EX := EX^+ - EX^- \in \overline{\mathbb{R}},$$

gdje oznaka $\overline{\mathbb{R}}$ predstavlja prošireni skup realnih brojeva. Ako postoji EX tada definiramo i ***varijancu od*** *X, u oznaci* ***Var***$[X]$*, tako da je*

$$Var[X] := E[(X - E[X])^2].$$

Sada ćemo navesti nekoliko učestalih primjera slučajnih varijabli njihovim funkcijama distribucije koje ćemo koristiti dalje u ovom radu.

Primjer 1.7 (Bernoullijeva razdioba) *Bernoullijeva slučajna varijabla X indicira je li rezultat slučajnog pokusa bio uspjeh ili neuspjeh. Preciznije, X će imati vrijednost 1 ako se dogodio elementaran ishod koji je povoljan za događaj uspjeh, inače će imati vrijednost 0. Označimo sa $\theta = \mathbb{P}(X = 1)$ vjerojatnost uspjeha. Tablica distribucije dana je sa :*

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}$$

$$E[X^2] = 0^2 * (1-\theta) + 1^2 * \theta = \theta$$

$$Var[X] = E[X^2] - E[X]^2 = \theta - \theta^2$$

Primjer 1.8 (Negativna binomna razdioba) *Broj X nezavisnih jednako distribuiranih Bernoullijevi pokusa s vjerojatnosti uspjeha θ do uključivo k-tog uspjeha je slučajna varijabla koja ima negativnu binomnu razdiobu s parametrima k i θ, $0 < \theta < 1$. Njena distribucija dana je sa:*

$$X \sim \begin{pmatrix} k & k+1 & k+2 & \cdots \\ \theta^k & \binom{k}{k-1}\theta^k(1-\theta) & \binom{k+1}{k-1}\theta^k(1-\theta)^2 & \cdots \end{pmatrix},$$

$$E[X] = \frac{k}{\theta},\ Var[X] = k\frac{1-\theta}{\theta^2}.$$

Primjer 1.9 (Poissonova distribucija) *Slučajna varijabla X ima Poissonovu distribuciju s parametrom $\lambda > 0$ ako prima vrijednosti iz skupa $\{0, 1, 2, 3, ...\}$ s vjerojatnostima*

$$p_i = \mathbb{P}\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}$$

Tada pišemo $X \sim \mathcal{P}(\lambda)$.
Vrijedi:

$$E(X) = \lambda,$$

$$Var(X) = \lambda.$$

Poissonova distribucija interpretira se kao slučajan broj događaja u nekom vremenskom periodu, pri čemu je parametar θ stopa učestalosti odnosno prosjek događaja po jedinici vremena.

Primjer 1.10 (Uniformna razdioba) *Kažemo da neprekinda slučajna varijabla X ima **uniformnu razdiobu** na segmentu $[a, b]$, i pišemo $X \sim U(a, b)$, ako je funckija gustoće zadana kao:*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in \langle a, b\rangle\,, \\ 0, & inače. \end{cases}$$

Primjer 1.11 (Normalna razdioba) *Kažemo da neprekinda slučajna varijabla X ima **normalnu razdiobu** s parametrima μ i $\sigma^2 > 0$, i pišemo $X \sim N(\mu, \sigma^2)$, ako je $\operatorname{Im} X = \mathbb{R}$ i gustoća joj je*

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

Vrijedi: $E[X] = \mu$, $Var[X] = \sigma^2$.

Primjer 1.12 (Gama razdioba) *Kažemo da neprekinda slučajna varijabla X ima* ***gama razdiobu*** *s parametrima $\alpha > 0$ i $\lambda > 0$, i pišemo $X \sim \Gamma(\alpha, 1/\lambda)$, ako je $\operatorname{Im} X = \langle 0, +\infty \rangle$ i gustoća joj je*

$$f_X(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & za \quad x > 0, \\ 0 & inače. \end{cases}$$

gdje je $\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt$ Γ-funkcija.
Vrijedi: $E[X] = \frac{\alpha}{\lambda}$, $Var[X] = \frac{\alpha}{\lambda^2}$
Slučajna varijabla sa gore opisanom razdiobom može se intepretirati kao vrijeme čekanja da se dogodi točno α događaja u Poissonovom procesu sa intezitetom λ.

1.2 Osnove statistike

Definicija 1.13 *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i neka je $\mathcal{P}$ familija vjerojatnosti na $(\Omega, \mathcal{F})$. Uređena trojka $(\Omega, \mathcal{F}, \mathcal{P})$ se naziva* ***statistička struktura****.*
Familija $\mathcal{P}$ često je parametrizirana, a zapisuje se na sljedeći način:

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$$

Definicija 1.14 *Na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ definiramo* ***statistiku*** *kao slučajnu varijablu (ili slučajni vektor) $T : \Omega \to \mathbb{R}$ $(\mathbb{R}^k)$ tako da postoji $n \in \mathbb{N}$ i n-dimenzionalni slučajni vektor $(X_1, ..., X_n)$ na $(\Omega, \mathcal{F}, \mathcal{P})$ te izmjeriva funkcija $t : \mathbb{R}^n \to \mathbb{R}(\mathbb{R}^k)$ takva da je $T = t(X_1, ..., X_n)$.*

Neka je $\mathcal{X} = (X_1, ..., X_n)$ slučajan uzorak iz modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}, \Theta \subseteq \mathbb{R}^p$. Na osnovi zadanog uzorka želimo procijeniti vrijednost parametra θ, ili općenito, neke njegove funkcije $\tau(\theta) \in \tau(\Theta) \subseteq \mathbb{R}^k$.

Definicija 1.15 ***(Točkovni) procjenitelj*** *od $\tau(\theta)$ je statistika $T = t(\mathcal{X}) = t(X_1, ..., X_n)$ u $\mathbb{R}^k$.*

Jednostavnije rečeno, točkovna procjena slučajne varijable je postupak procjene vrijednosti ili parametara neke slučajne varijable na temelju dostupnih podataka ili uzoraka. Cilj je dobiti jedan broj, kojeg nazivamo točkovna procjena, koji predstavlja najbolju procjenu nepoznate vrijednosti ili parametra.

Postoji nekoliko različitih metoda za dobivanje točkovnih procjena, a sada ćemo opisati jednu.
Neka je $X = (X_1, ..., X_n)$ slučajni uzorak iz modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$. Ako je $x = (x_1, ..., x_n)$ realizacija slučajnog uzorka X tada definiramo **funkciju vjerodostojnosti** sa

$$L : \Theta \to \mathbb{R},\ L(\theta|x) = L(\theta) := f_X(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Definicija 1.16 (Metoda maximalne vjerodostojnosti (MLE)) $\hat{\theta} = \hat{\theta}(\mathcal{X})$ *je procijenitelj* ***maksimalne vjerodostojnosti*** *za θ ako vrijedi*

$$L(\hat{\theta}|\mathcal{X}) = \max_{\theta\in\Theta} L(\theta|\mathcal{X}).$$

Gore smo definirali točkovnu procijenu, a sada ćemo definirati intervalnu procjenu. Ideja intervalne procjene je konstruirati interval oko točkovne procjene.

Koeficijent pouzdanosti je vjerojatnost da promatrani interval sadrži vrijednost nepozatog parametra promatrane razdiobe i zato taj interval konstruiramo kao slučajni interval, tj. interval čije su granice slučajne varijable.

Definicija 1.17 *Za danu pouzdanost* $1-\alpha \in (0,1)$*, slučajan uzorak* $(X_1, ..., X_n)$ *iz* X *i statistike* $L_n = f(X_1, ..., X_n), D_n = g(X_1, ..., X_n)$ *kažemo da je interval* $[L_n, D_n]$ ***interval pouzdanosti*** *(pouzdanosti* $1-\alpha$*) za parametar* τ *ako je* $\mathcal{P}(L_n \leq \tau \leq D_n) \geq 1-\alpha$*.*

Uočimo da gornja definicija govori da interval $[L_n, D_n]$ sadrži τ u barem $(1-\alpha)100\%$ realizacija slučajnog uzorka $(X_1, ..., X_n)$.

1.3 Osnove Bayesovske statistike

Teorem 1.18 (Bayesova formula) *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor,* $A, B \in \mathcal{F}$ *događaji takvi da je* $P(B) \neq 0$*. Tada vrijedi:*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Kroz ovaj rad koristit ćemo statističke modele koji se temelje na Bayesovoj formuli, a njihova prednost je ta što mogu povezati prostornu varijabilnost neke druge varijable i naše promatrane varijabilnosti morbiditeta ili mortaliteta. U ovom radu ćemo pokušati povezati nekakve druge varijable sa zarazom koronavirusa.

Kako bi mogli opisati rad ovih modela moramo na nekakav način definirati sljedeće pojmove : **aposteriorna distribucija**, **apriorna distribucija**, **funkcija vjerodostojnosti** (engl. likelihood function).

Uočimo da funkcija vjerodostojnosti zapravo govori koliko je vjerodostojna svaka vrijednost parametra θ ako znamo x. Ukoliko se radi o diskretnoj slučajnoj varijabli tada umjesto $f(x|\theta)$ koristimo $P(x|\theta)$ tj. uvjetnu vjerojatnost danih podataka x uz uvjet da je vjerovanje u θ istinito.

1.3.1 Ažuriranje vjerovanja o parametru

Na ovaj način ćemo izvoditi ažuriranje vjerovanja o parametru. Neka je Θ diskretna slučajna varijabla. Sa E označimo skup nekih informacija o Θ.
Od interesa nam je distribucija parametra Θ uz uvjet da znamo informacije E. Dakle, želimo doći do $P(\Theta = \theta|E)$. Koristeći Bayesov teorem dobijamo sljedeće :

$$P(\Theta = \theta|E) = P(\Theta = \theta)\frac{P(E|\Theta = \theta)}{P(E)}$$

Sada ćemo komentirati i imenovati neke pojmove.

- $P(\Theta = \theta)$ je inicijalna funkcija vjerojatnosti od Θ. Ova funkcija je zapravo vjerovanje o parametru Θ prije uzimanja u obzir bilo kakvih drugih informacija, konkretno u našem slučaju bez uzimanja u obzir informacija iz E. Ovu funkciju nazivamo **apriorna funkcija vjerojatnosti**.

- $P(\Theta = \theta|E)$ je funkcija vjerojatnosti Θ uz uvjet da su nam poznate nove informacije iz E. Uočimo da je ovo funkcija koja je ažurirana sa informacijama. Upravo ova funkcija nam je od interesa, nju nazivamo **aposteriorna funkcija vjerojatnosti**.

- $\frac{P(E|\Theta=\theta)}{P(E)}$ je izraz koji pokazuje promjenu od apriori do aposteriori distribucije. Uočimo da je izraz $P(E|\Theta)$ funkcija vjerodostojnosti.

Sada ćemo iskazati strogu matematičku definiciju za aposteriori distribuciju.

Definicija 1.19 *Neka je $\theta \in \Theta$, $\pi : \Theta \to \mathbb{R}$ apriori distribucija parametra θ i x realizacija slučajne varijable (vektora) X. Distribuciju od θ za dani x nazivamo **aposteriori distribucija** i definiramo sa*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)},$$

gdje je $f(x)$ marginalna distribucija od X.
Marginalna distribucija $f(x)$ računa se na sljedeći način:

$$f(x) = \begin{cases} \sum_\theta f(x|\theta)\pi(\theta), & \text{u diskretnom slučaju,} \\ \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta), & \text{u neprekidnom slučaju,} \end{cases} \tag{1.1}$$

gdje je $\pi(x)$ apriorna distribucija.
Navest ćemo sljedeće definicije u svrhu definiranja Bayesovog statističkog modela.

Definicija 1.20 *Bayesovi statistički model slučajne varijable (ili slučajnog vektora)* $X : \Omega \to \mathbb{R}$ $(\mathbb{R}^k)$ *je familija*

$$\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$$

iz definicije 1.13 pri čemu je parametaru θ *pridružena njegova apriorna distribucija s gustoćom* $\pi(\theta), \theta \in \Theta$.

U gore opisanom načinu vidjeli smo kako ažuriramo vjerovanje o parametru kombinirajući obrađene podatake i apriori uvjerenja. Prema tome, i prije skupljanja podataka mi imamo nekakva uvjerenja, odnosno prethodno znanje koje može biti subjektivno, što nas na kraju može dovesti do krivog rezultata. Upravo ovo apriori uvjerenje je najveća kritika u Bayesovom zaključivanju.

Sada ćemo iznijeti primjer za lakše razumjevanje Bayesovih modela i principa računanja.

Primjer 1.21 *Pojavila se smrtonosna bolest* X*, a poznato je da je vjerojatnost da se zarazi tom bolešću 10%. Međutim, poznato je da ako je osoba plućni bolesnik da su tada izgledi za zarazu nešto veći.*
Definiramo sljedeća dva događaja:

- A *... osoba se zarazila bolešću* X*.*
- B *... osoba je plućni bolesnik.*

Dakle u početku imamo apriori vjerovanje da vrijedi sljedecće:

$$P(A) = 0.1 \text{ odnosno } P(A^c) = 0.9.$$

Nadalje, statističkim istraživanjem je utvrđeno da ukoliko se osoba zarazila bolešću X *da je vjerojatnost da je ona plućni bolesnik jedanaka 30%. Istovremeno se pokazalo da ukoliko je poznato da se osoba nije zarazila bolešću* X*,*

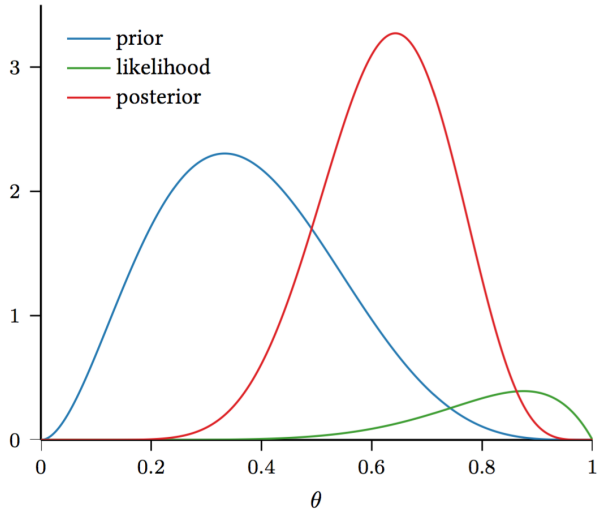
vjerojatnost da je plućni bolesnik je 15%. Dakle, gornja razmatranja možemo napisati ovako:

$$P(B|A) = 0.3 \quad i \quad P(B|A^C) = 0.15.$$

Sada nam je od interesa odrediti $P(A|B)$ odnosto vjerojatnost da će se osoba zaraziti ako je plućni bolesnik. Korištenjem Bayesovog teorema sljedi:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \\ = \frac{0.3 \times 0.1}{0.3 \times 0.1 + 0.15 \times 0.9} \approx 0.18 = 18\%.$$

Uočimo da je izračunata vjerojatnost aposteriori veća nego naša početna apriori. Dakle, prema ovom primjeru, ukoliko je osoba plućni bolesnik veća je vjerojatnost da će se zaraziti bolešću X.



Slika 1.1: Prikaz apriori, aposteriori i likelihood funkcija

Slika 1.1 dobar je prikaz kako je aposteriori zaključak kojeg dobijemo kompromis između apriori uvjerenja i vjerodostojnosti mjerenja. Vidljivo je da krećemo od apriori uvjerenja te podešavamo "u stranu" vjerodostojnosti mjerenja.

Važno je za istaknuti da u slučaju kad imamo funkciju gustoće $f(x|\theta)$ i apriornu funkciju gustoće parametra $\pi(\theta)$ možemo izračunati aposteriornu distribuciju parametra θ, ali taj izračun često je problematičan.
Kako bi riješili ovaj problem izračuna koristimo se Monte Carlo Markovljevim lancima (MCML). Najjednostavnije rečeno, MCML je metoda generiranja simulacije parametara modela koji nakon prikladnog razdoblja sagorijevanja (engl. burn-in) postaju ostvarenja njihovih posteriornih distribucija. MCML je jako primjenjiva metoda kako u mapiranju bolesti tako i u ekonomiji i raznim drugim područjima.
Softver kojeg ćemo koristiti za korištenje MCML je **WinBUGS** koji je otvorenog koda (engl. open source), a paket u R-u koji će nam povezivati taj vanjski softver sa našim programom je **R2WinBUGS**.

1.4 Monte Carlo Markovljevi lanci

MCML je metoda generiranja slučajnih uzoraka iz dane distribucije. Na početku se zadaje distribucija π koju nazivamo ciljna distribucija i skup stanja S. Ova metoda zatim generira Markovljev lanac čija stacionarna distribucija dobro aproksimira distribuciju π.
Pretražujemo skup S na način da promatramo vjerojatnosti prelaska iz stanja nekog stanja i u stanje j, gdje će stanje j biti najvjerojatnije od svih ostalih mogućih. Tada se stanje j dodaje u lanac. Lanac koji se simulira je ergodski te je njegova stacionarna distribucija jednaka graničnoj.
Dakle, što više simulacija provedemo, bit ćemo bliži ciljnoj distribuciji. Kako je lanac ergotski, on je također i ireducibilan (sva stanja međusobno komuniciraju). Uočimo da nam ovo povlači činjenicu da ćemo posjetiti stanje $j \in S$ sa nekom vjerojatnošću $\pi(j) > 0$ bez obzira iz kojeg stanja krenuli. Ovo će

nam biti jako važno svojstvo u primjeni.
Postoji više algoritama za MCML, ali ovdje ćemo iznijeti najopćenitiji algoritam.

1.4.1 Metropolis-Hasting algoritam

Ovaj algoritam simulira uzorke iz distribucije zadane funkcijom gustoće f kojoj se "približavamo" tako da povećavamo veličinu uzorka. U pokušaju da to napravimo trebamo konstruirati ergotski Markovljev lanac takav da njegova stacionarna distribucija dobro aproksimira distribuciju s funkcijom f kojoj nam je cilj približiti se.
Uvjetna vjerojatnost $q(x, y) = q(y|x)$ označava vjerojatnost prelaska iz stanja x u stanje y, gdje je $x, y \in S$. Uvjetna vjerojatnost $q(x, y)$ će nam koristiti u algoritmu da odlučujemo koje stanje ćemo ili prihvatiti (dodati u lanac) ili odbaciti.
Dovoljno je poznavati f i q do na normalizirajuću konstantu da bi ovaj algoritam mogli izvesti.
Krećemo iz stanja $x_0 \in S$ za koje je $f(x_0) > 0$ i pratimo sljedeći algoritam.

Algoritam 1.22 *Metropolis-Hastingov algoritam:*

0. *Neka je* $X_n = x_n$.

1. *Generiraj* y_n *iz* $Y_n \sim q(y|X_n)$.

2. *Generiraj* u *iz* $U \sim U(0, 1)$.

3. *Izračunaj* $\rho(x_n, y_n)$, *gdje je*

$$\rho(x_n, y_n) = min\left\{ \frac{f(y_n)}{f(x_n)} \frac{q(x_n|y_n)}{q(y_n|x_n)}, 1 \right\}.$$

4. *Definiraj* $x_{n+1} = \begin{cases} y_n, & \text{ako je } u \leq \rho(x_n, y_n), \\ x_n, & \text{inače.} \end{cases}$

5. *Ponavljaj za* $n = n + 1$.

U prvom koraku generiramo novo stanje pomoću prijelazne vrijednosti $q(y|x)$. Nakon toga donosimo kriterij prelaska u novo stanje.

Pretpostavimo da gledamo novčić nepravilnog oblika i neka je u vjerojatnost okretanja glave koja dolazi iz uniformne distribucije na segmentu $[0, 1]$.

Hoćemo li prijeći u novo stanje ovisi o sljedećem omjeru:

$$\frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}.$$

Prvi faktor uspoređuje vjerojatnosti dva stanja, a drugi vjerojatnost prelaska iz predloženog u trenutno stanje i obrnuto.

Ako je gornji izraz veći od 1, to znači da je novo stanje vjerojatnije od trenutnog. Tada će po definiciji od ρ vrijediti $\rho(x_n, y_n) = 1$ što znači da prelazimo u novo stanje jer $u \in [0, 1]$. Ako je manji od 1, bacamo novčić i u slučaju okretanja glave prijeći ćemo u novo stanje, a okretanjem pisma ostajemo u starom. Uočimo da nam ovakva formulacija omogućava da algoritam ne zapne u nekom apsorpcijskom stanju iz kojeg ne može izaći.

Iz ovog kvocijenta sada je jasno zašto je dovoljno poznavati f i q do na normalizirajuću konstantu. Potreban nam je samo njihov omjer.

Vjerojatnost $\rho(x_n, y_n)$ nazivamo **vjerojatnost prihvaćanja**. Ideja ovog algoritma zasniva se na metodi pokušaja i pogreške. Korak 4. implicira da se ovdje radi o konstrukciji Markovljevog lanca jer svako novo stanje ovisi samo o prethodnom stanju.

1.5 Empirijske Bayesovske metode

Empirijske Bayesove metode su postupci za statističko zaključivanje u kojima se apriorna distribucija parametra θ procjenjuje iz slučajnog uzorka slučajne varijable X. Ovaj pristup je u suprotnosti sa standardnim Bayesovim metodama, za koje se apriorna distribucija fiksira prije promatranja bilo kakvih podataka.

Empirijske Bayesove metode mogu se promatrati kao aproksimacija Bayesovog tretmana hijerarhijskog Bayesovog modela. U dvostupanjskom hijerarhijskom Bayesovom modelu, pretpostavlja se da su promatrani podaci $X = \{X_1, X_2, \ldots, X_n\}$ generirani iz distribucije $p(x \mid \theta)$ s neopaženim skupom parametara $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$. S druge strane, parametri θ se smatraju uzorcima iz distribucije $p(\theta \mid \eta)$ koju karakteriziraju hiperparametri η. U hijerarhijskom Bayesovom modelu pretpostavja se da hiperparametri η dolaze iz neparametrizirane distribucije $p(\eta)$.

Informacija o određenom parametru od interesa θ_i stoga ne dolazi samo iz svojstava od uzorka X, već i iz svojstava populacije parametara θ kao cjeline, koji su sažeti u hiperparametrima η.

Koristeći Bayesov teorem dobivamo

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} = \frac{p(x \mid \theta)}{p(x)} \int p(\theta \mid \eta)p(\eta)d\eta. \tag{1.2}$$

U najopćenitijoj situaciji, integral u (1.2) neće moći analitički izračunati i te će se procijenjivati numeričkim metodama. Mogu se koristiti stohastičke ili determinističke aproksimacije. Primjeri stohastičkih metoda su Monte Carlo Markovljevi lanci i Monte Carlo uzorkovanje.

Alternativno, izraz (1.2) se može napisati kao

$$p(\theta \mid x) = \int p(\theta \mid \eta, x)p(\eta \mid x)d\eta = \int \frac{p(x \mid \theta)p(\theta \mid \eta)}{p(x \mid \eta)} p(\eta \mid x)d\eta,$$

te se faktor u integralu može se izraziti kao

$$p(\eta \mid x) = \int p(\eta \mid \theta) p(\theta \mid x) d\theta.$$

Ovo sugeriraju iterativnu shemu za razvoj sukcesivno poboljšanih aproksimacija $p(\theta \mid x)$ i $p(\eta \mid x)$. Prvo se izračuna početna aproksimacija $p(\theta \mid x)$ potpuno zanemarujući ovisnost o η. Zatim se izračuna aproksimacija $p(\eta \mid x)$ na temelju početne aproksimativne distribucije $p(\theta \mid x)$; te se upotrijebi $p(\eta \mid x)$ za ažuriranje aproksimacije za $p(\theta \mid x)$. Nakon toga se ažurira $p(\eta \mid x)$ i tako dalje.

Kada prava distribucija $p(\eta \mid x)$ ima "oštri vrh", integral koji određuje $p(\theta \mid x)$ ne može se puno promijeniti zamjenom distribucije preko η s točkovnom procjenom $\hat{\eta}$ koja predstavlja mod distribucije (ili, alternativno, njenu srednju vrijednost),

$$p(\theta \mid x) \simeq \frac{p(x \mid \theta) p\left(\theta \mid \hat{\eta}\right)}{p\left(x \mid \hat{\eta}\right)}.$$

Spomenimo još da se pojam "Empirijski Bayes" može pokriti široku paletu metoda, ali većina se može smatrati skraćivanjem gornje sheme ili nečim sličnim. Za procjene parametra η se koriste uglavnom točkovne procjene, a ne cijela distribucija. Procjene $\hat{\eta}$ obično se dobivaju bez razmatranja odgovarajuće apriorne distribucije za η.

Primjer 1.23 (Poisson-gamma model) *Pretpostavimo da se podaci X ravnaju po Poissonovoj distribuciji s parametrom θ i neka je apriorna distribucija od θ sada određena konjugiranom apriornom distribucijom, što je u ovom slučaju gamma distribucija ($Ga(\alpha, \beta)$). Lako se pokaže da je aposteriorna distribucija od θ također gamma distribucija. Vrijedi*

$$p(\theta \mid x) \propto p(x \mid \theta) p(\theta \mid \alpha, \beta),$$

gdje $\propto$ *označava proporcionalnost, a marginalna distribucija* $p(x)$ *je izostavljena budući da ne ovisi eksplicitno o* θ*. Uvrštavanjem točnih vrijednosti članova koji ovise o* θ *dobivvamo aposteriornu distribuciju:*

$$\rho(\theta \mid x) \propto \left(\theta^y e^{-\theta}\right)\left(\theta^{\alpha-1} e^{-\theta/\beta}\right) = \theta^{x+\alpha-1} e^{-\theta(1+1/\beta)}.$$

Tako je aposteriorna distribucija također gamma distribucija $Ga\,(\alpha', \beta')$*, gdje je* $\alpha' = y+\alpha$*, i* $\beta' = (1+1/\beta)^{-1}$*. Također, uočimo da marginalnu distribuciju* $p(x)$ *dobivamo kao integral posteriorne po svim* θ*, što daje negativnu binomnu distribuciju.*

Kako bismo primijenili empirijski Bayes, aproksimirat ćemo graničnu vrijednost pomoću metode maksimalne vjerodostojnosti (MLE). Ali budući da je aposteriorna gamma distribucija, MLE marginalne ispada da je samo srednja vrijednost aposteriore, što je točkovna procjena $E(\theta \mid x)$ *koja nam je potrebna. Podsjećamo da je* μ *gamma distribucije* $Ga\,(\alpha', \beta')$ *dana s* $\alpha'\beta'$*, pa imamo*

$$E(\theta \mid x) = \alpha'\beta' = \frac{\bar{x} + \alpha}{1 + 1/\beta} = \frac{\beta}{1+\beta}\bar{x} + \frac{1}{1+\beta}(\alpha\beta).$$

Da bi se dobili vrijednosti α *i* β*, empirijski Bayes propisuje procjenu srednje vrijednosti* $\alpha\beta$ *i varijance* $\alpha\beta^2$ *korištenjem kompletnog skupa empirijskih podataka. Rezultirajuća točkovna procjena* $E(\theta \mid y)$ *stoga je u obliku ponderiranog prosjeka srednje vrijednosti uzorka* $\bar{x}$ *i prethodne srednje vrijednosti* $\mu = \alpha\beta$*. Ispostavilo se da je to opće obilježje empirijskog Bayesa: bodovne procjene za prethodnu (tj. srednju vrijednost) izgledat će kao ponderirani prosjeci procjene uzorka i prethodne procjene (kao i za procjene varijance).*

Poglavlje 2

Osnove pri mapiranju zaraze

Cilj mapiranja zaraze je pružiti prikaz prostorne distribucije bolesti gdje pretpostavljamo da su podaci podjeljeni na disjukntne regije, u našem slučaju županije. Bolest može se reflektirati na dva načina koje nazivamo **morbiditet** i **mortalitet**. Morbiditet označava broj oboljenja dok mortalitet označava broj umrlih od određene bolesti.

U ovom radu promatramo obje pojave u slučaju bolesti izazvane koronavirusom (COVID-19) te pokušavamo zaključiti postoji li županija koja ima povećan rizik bilo mortaliteta ili morbiditeta. Kako takvi zaključci mogu biti varljivi, dobro je dane podatke razmatrati podijeljene u više slojeva kako bi izbjegli efekt zbunjujućih faktora[1] (engl. Confounfing factor)

Slojevi mogu biti: spol, godine (gdje možemo na više načina definirati dobne skupine), županije u kojima ljudi žive, nekakvi obrasci ponašanja koji mogu mjerljivi...

Krećemo od oznaka P_{ij} i O_{ij} koji označavaju populaciju i broj promatranih slučajeva zaraze u regiji i i sloju j. Sumiranjem po svim slojevima j možemo

[1] Čimbenik koji zbunjuje, može prikriti stvarnu povezanost ili lažno pokazati prividnu povezanost između varijabli u istraživanju ako ne postoji stvarna povezanost između njih.

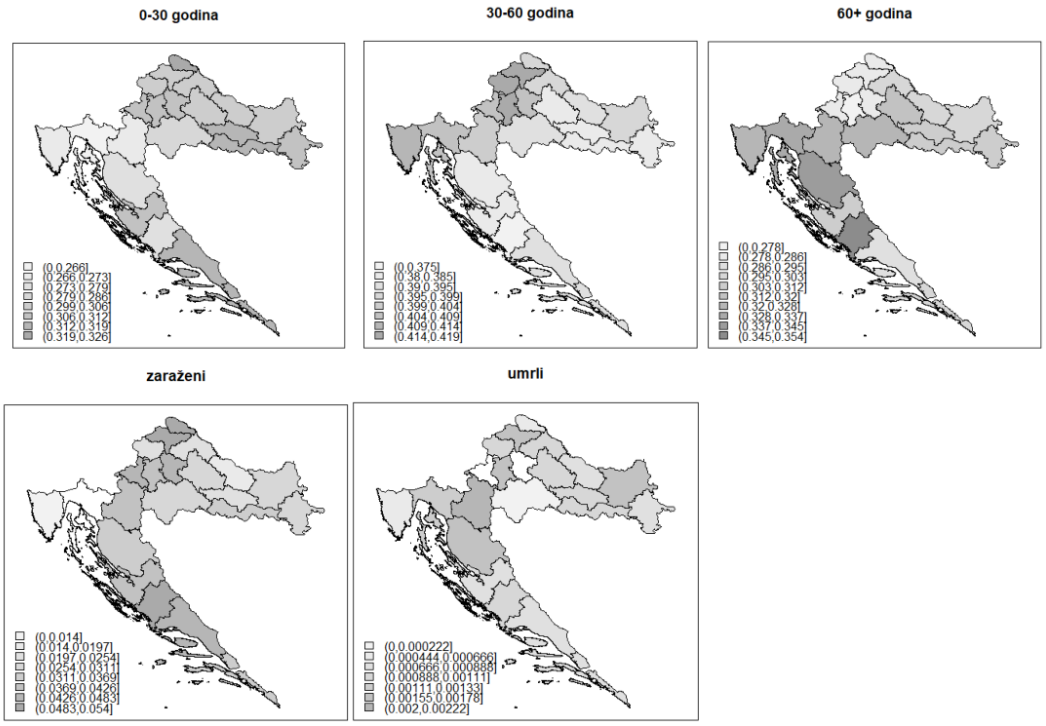
dobiti ukupnu populaciju P_i i slučajeve O_i u nekoj regiji i. Sada na isti način sumiranjem svih P_i dobijemo P_+, analnogno dobijemo O_+.
Dakle, P_+ je ukupan broj populacije koju promatramo, a O_+ je ukupan broj opaženih slučajeva. Kako nam brojevi opaženih slučajeva ne daje puno informacija o samoj distribuciji zaraze potrebno ih je usporediti sa prosječnim brojem slučajeva kojeg ćemo računati na sljedeći način: $E_i = P_i r_+$ gdje je $r_+ = \frac{O_+}{P_+}$. Ovdje se radi o najjednostavnijem slučaju standardizacije kojeg nazivamo *indirektna standardizacija*.
Kada su nam podaci grupirani u slojeve umjesto računanja globalnog omjera $\frac{O_+}{P_+}$ kojeg koristimo za sve regije, za svaki pojedinačni sloj možemo izračunati $r_j = \frac{\sum_i O_{ij}}{\sum_i P_{ij}}$. Ova standardizacije se naziva *interna standardizacija*.
Pogledajmo sada podatke o zarazi koronavirusom Hrvatskoj i njihovu raspodjelu po županijama. Podaci su o slučajevima za prvu polovicu 2021 godine.

Na Slici 2.1 vidimo broj zaraženih po svakoj županiji, ali očigledno je da gledajući samo te podatke ne možemo ništa zaključiti o povećanom riziku zaraze, odnosno smrtnosti u nekoj županiji. U tu svhu smo napravili prikaz na slici 2.2.

Iz ovih prikaza stvari su nešto jasnije nego kada imamo same brojeve slučajeva po županijama. Iz Slike 2.2 bi se dalo naslutiti da je možda broj umrlih i zaraženih pozitivno koreliran sa udjelom starijeg stanovništva od 30 godina. Tu tezu ćemo dalje kroz ovaj rad pokušati potkrijepiti statističkim zaključcima.

U sljedećem prikazu ćemo tamnijim bojama prikazivati županije sa većim brojem slučajeva.

Slika 2.1: Broj zaraženih osoba COVID-19 virusom u period od 1.1.2021 do 1.7.2021. po svakoj županiji u Republici Hrvatskoj



Slika 2.2: Slika 1. red lijevi stupac: Udio zaraženih u dobnoj skupini od 0 do 30 godina, Slika 1. red srednji stupac: Udio zaraženih u dobnoj skupini od 30 do 60 godina, Slika 1. red desni stupac: Udio zaraženih u dobnoj skupini preko 60 godina, Slika 2. red lijevi stupac: Ukupan broj zaraženih, Slika 2. red desni stupac: Ukupan broj umrlih.

Poglavlje 3

Statistički modeli

Uobičajena statistička pretpostavka za modeliranje promatranih slučajeva u regiji i i sloju j je ta da je broj slučajeva izveden iz Poissonove distribucije sa očekivanjem $\theta_i E_{ij}$, gdje je θ_i relativan rizik u regiji i, $\theta_i = O_i/P_i$, a $E_{ij} = P_{ij}r_j$. Kada je relativni rizik 1 to znači da je rizik prosječan, a od interesa će nam biti rizici koji su značajnije veći od 1, što ukazuje povećan rizik. Ovdje ćemo pretpostavljati da nema interakcije između rizika i populacijskog sloja, tj. θ_i ovisi samo o regiji i.

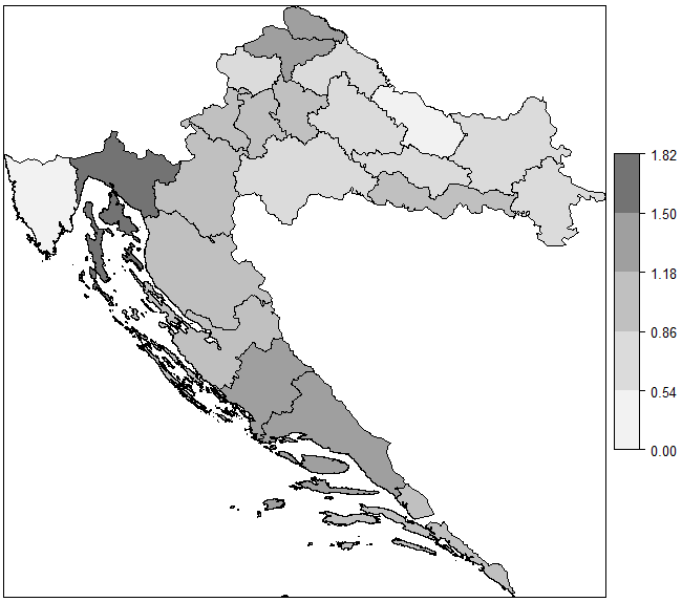
3.1 Standardizirani omjer smrtnosti (engl. Standardised Mortality Ratio SMR)

Jedan način procjene relativnog rizika θ_i bilo smrtnosti ili incidencije kojeg ćemo najviše koristiti u ovom radu definira se na sljedeći način:

$$SMR_i = \frac{O_i}{E_i}$$

kojeg nazivamo *Standardizirani omjer smrtnosti (engl. standardised mortality ratio) kojeg ćemo kroz tekst kraće označavati sa SMR.*

3.1. Standardizirani omjer smrtnosti (engl. Standardised Mortality Ratio SMR)



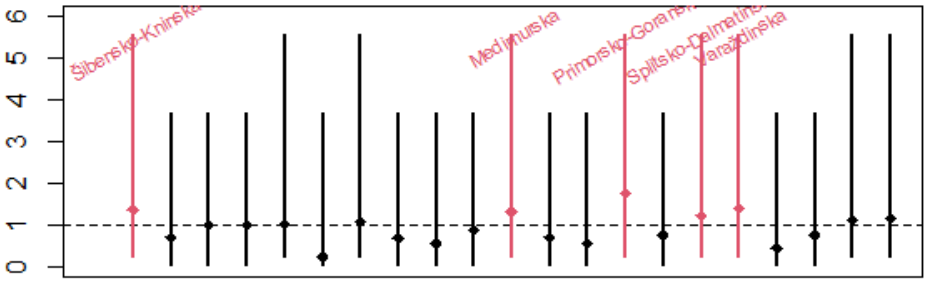
Slika 3.1: SMR

Uz pomoć SMR-a i njegovog grafičkog prikaza na slici 3.1 dobivamo dosta više informacija i bolji dojam o varijabilnosti rizika nego na slici 1.1 gdje koristimo same brojeve slučajeva. Vidljivo je da sa prikazom SMR-a dobijamo dosta benefita, ali on ima neke nedostatke.

SMR u nekim slučajevima i nije baš najbolji prikaz distribucije bolesti ili pak nije dobar za određivanje područja većeg relativnog rizika. Naime, u područjima gdje je populacija dosta manja nazivnik u izrazu SMR_i je mali pa mala varijacija O_i može dramatično promjeniti cijeli razlomak što nas može dovesti do krivog zaključka. Crvenom bojom na slici 3.2 ćemo označiti samo one županije koje imaju veći relativan rizik od 1.2. Sada koristeći činjenicu da je O_i iz Poissonove distribucije možemo izračunati interval pouzdanosti za svaki SMR_i. U tu svrhu u R-u koristimo paket **epitools** i funckiju *pois.exact*.

Povećan SMR_i, u našem slučaju veći od 1, sugerira da je zabilježen broj

smrtnih slučajeva veći nego što bismo prethodno očekivali.



Slika 3.2: 95% intervali pouzdranosti za SMR

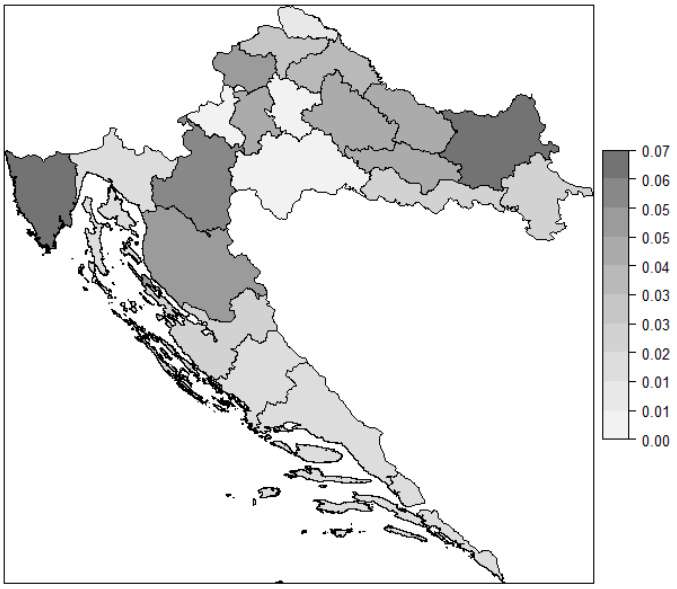
Najveću vrijednost za relativni rizik ima Primorsko-Goranska županija i to 1.769, sljede Varaždinska i Međimurksa sa vrijednostima 1.37 odnosno 1.31. Uočimo, kako je u našem slučaju $SMR_i = \frac{O_i}{E_i} = \frac{O_i}{P_i}\frac{\sum_i P_i}{\sum_i O_i}$ to povlači da u Primorsko-Goranskoj županiji ima oko 77% više zaraženih nego u općoj populaciji jer za taj i vrijedi $\frac{O_i}{P_i} = 1.769\frac{\sum_i O_i}{\sum_i P_i}$.

Prikaz SMR je poprilično "grub" i dopušta velike varijacije među regijama. Javlja se potreba da izradimo nešto zaglađeniji prikaz. U tu svrhu ćemo u nastvaku rada izložiti 3 različita modela, a prije toga ćemo pokazati još jedan standardizirani omjer koji se koristi u praksi kada se promatra određena bolest u određenom vrememenu.

Na slici 3.2 su prikazani intervali 95%-tne pouzdanosti za SMR pomoću funkcije **pois.exact** iz paketa **epitools** u R-u. Vidljivo je da su ti intervali dosta "široki" pa ne možemo sa sigurnošću zaključiti nešto o regijama povećanog rizika.

3.2 Stopa smrtnosti (engl. Case Fatality Rate CFR)

Stopa smrtnosti (u daljnjem teksu ćemo koristiti kraticu CFR) mjeri težinu određene bolesti te predstavlja omjer ukupnog broja smrtnih slučajeva i broja u određenom vremenskom intervalu. Možemo ga definirati na sljedeći način : $CFR_i = \frac{U_i}{O_i}$, gdje je U_i broj umrlih u regiji i, a O_i broj potvrđenih slučajeva zarazom u regiji i. Nakon dijela programa u R-u na slici 3.3 prikazujemo grafiči prikaz od stope smrtnosti od koronavirusa u prvoj polovini 2021. godine.



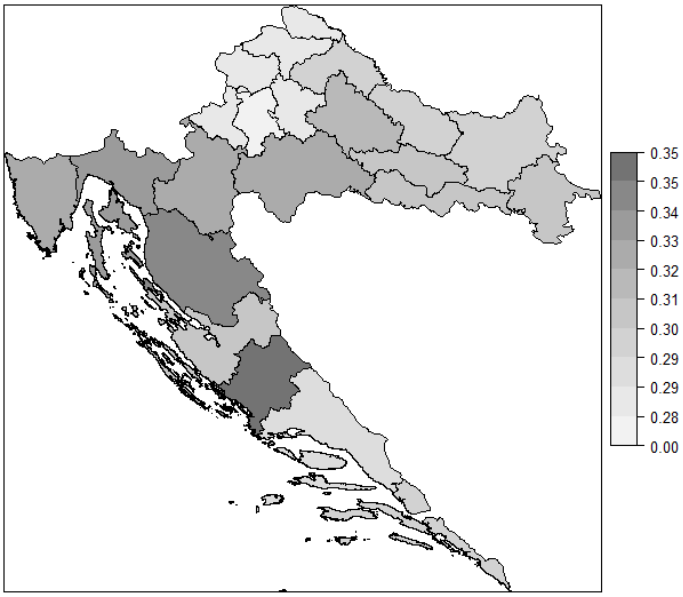
Slika 3.3: CFR bolesti izazvane koronavirusom za prvu polovinu 2021. godine po županijama

Najveće vrijednosti CFR-a imaju županije Istarska, Osječko-Baranjska, Karlovačka i Ličko-Senjska.

Ovakav prikaz je informativan te može ukazivati na razne probleme kao što

je lošiji zdravstveni tretman od prosjeka u državi. No razlozi mogu biti svakakvi, primjerice možda baš u ovim županijama je povećan broj ljudi starijih od 60 godina za koje je poznato da se teže odupiru koronavirusom. Ako ove rezultate usporedimo sa udjelom starijih od 60 godina iz slike 2.2 možemo vidjeti da upravo upravo u tim županijama (osim Osječko-Baranjske) je povećaj udio starijih osoba.
Nedostaci ovog omjera su slični kao oni u SMR-u. Ako iz slike 2.2 uočimo da je broj zaraženih u Istarskoj županiji relativno mali, tada moramo biti oprezni u donošenju zaključaka za Istarsku županiju jer mala promjena u brojniku može znatno utjecati na konačni CRF_i.

Globalni CFR za cijelu hrvatsku u prvoj polovini 2021. godine je 2.8% dok je na svijetskoj razini 2.1%.
Učitali smo podatke za prvu polovinu 2022. godine te napravili sličan postupak. Dobijamo sljedeće rezultate: Uspredimo li ove rezultate sa udjelom populacije starije od 60 godina dobijamo još bolje poklapanje. Šibensko-Kninska i Ličko-Senjska sada imaju najveći CFR_i, a ujedno su to županije sa najvećim udjelom starijih od 60.
Globalni CFR za Hrvatsku u prvoj polovici 2022. godine iznosi 0.082%. Mnogo faktora je moglo utjecati na ovoliku razliku u stopi smrtnosti. Mnoge pristranosti su se mogle dogoditi u odnosu na 2021. Na primjer veći broj testiranja (i točniji testovi) gdje otkrivamo sve vrste zaraze za razliku od prije. No, ovo je možda samo indikacija da postižemo kolektivni imunitet. Još jedan razlog za ovo može biti procjepljnost stanovništva koja je dosta veća u 2022. nego u 2021. godini. No, ovdje ne možemo donijeti siguran zaključak jer je moguće da je bolest lakša kod novih sojeva virusa.



Slika 3.4: CFR bolesti izazvane koronavirusom za prvu polovinu 2022. godine po županijama

3.3 Poisson-Gamma Model

Korištenje Poissonove distribucije ima ograničenje i zahtjeva pretpostavke koje u praksi neće uvijek biti zadovoljene. Glavni nedostatak je taj što zahtjeva da očekivanje i varijanaca od O_i budu isti što većinom nije slučaj. Uglavnom podaci budu "prerasprženi", tj. varijanca bude veća od očekivanja. U pokušaju da riješimo taj problem koristimo negativnu binomnu distribuciju umjesto Poissonove. Negativna binomna distribucija može se smatrati mješovitim modelom u kojem relativni rizik slijedi gama distribuciju za svaku regiju. Formulacija Poisson-Gamma (PG) modela strukturira se na sljedeća dva nivoa:

$$O_i|\theta_i, E_i \sim Po(\theta_i E_i),$$

$$\theta_i \sim Ga(\nu, \alpha),$$

gdje $Po(\theta_i E_i)$ označava Poissonovu razdiobu s parametrom $\theta_i E_i$ i $Ga(\nu, \alpha)$ Gamma razdiobu s parametrima ν, α. U ovom modelu smatramo da je relativan rizik θ_i slučajna varijabla dobivena iz gamma distribucije sa očekivanjem ν/α i varijancom ν/α^2. Uočimo da je distribucija od O_i uvjetovana sa θ_i. Lako se dobije da je neuvjetovanu distribucija svakog O_i negativna binomna distribucije sa parametrima ν i $\frac{\alpha}{\alpha+E_i}$. Dobije se da je aposteriornu distribuciju od θ_i uz poznavanje O_i gamma distribucija sa parametrima $\nu + O_i$ i $\alpha + E_i$. Grubo govoreći, informacije $\{O_i\}_{i=1}^n$ koje smo prethodno skupili ukomponiramo sa prijašnjim znanjem da bi dobili što bolje rezultate. Vrijedi

$$E[\theta_i|O_i] = \frac{\nu + O_i}{\alpha + E_i},$$

što možemo prikazati na sljedeći način kao "kompromis" između očekivanja apriorne distribucije relativnog rizika i SMR_i:

$$E[\theta_i|O_i] = \frac{E_i}{\alpha + E_i} SMR_i + (1 - \frac{E_i}{\alpha + E_i})\frac{\nu}{\alpha}$$

Trebalo bi istaknuti dva nedostatka ovakvog procjenitelja. Prvi nedostatak temelji se na nedostatku SMR-a kojeg smo gore već spomenuli. Dakle, E_i u malim regijama zna biti relativno mali pa mala varijacija O_i znatno utječe na vrijednost SMR_i što za posljedicu može imati premali ili prevelik utjecaj na model u odnosu na prethodno očekivanje koje smo imali bez unošenja tih "dodatnih" podataka. Drugi nedostatak je taj što su vrijednosti ν i α iste za sve regije pa time i očekivanje naše procjene što nije realan slučaj. Taj problem možemo riješiti definiranjem susjedstva regija, no to ćemo razmotriti nešto kasnije u ovom radu.

Kako su nam parametri ν i α obično nepoznati trebamo ih na neki način procijeniti. Za tu svrhu koristimo paket **DCluster** i funkciju **empbaysmoth**. Ovdje se koristi empiriska Bayesova metoda za procjenu tih parametara, a više detalja se može naći u [4]. Iz ovoga vidimo da je apriorno očekivanje

```
> eb<-empbaysmooth(shp$Observed,shp$Expected)
> shp$EBPG<-eb$smthrr
> eb$nu
[1] 6.427749
> eb$alpha
[1] 6.911916
> eb$nu/eb$alpha
[1] 0.9299519
```

Slika 3.5: Primjer koda u R-u gdje korištenjem empbaysmooth funkcije dobijamo procjenjene parametre ν i α.

od θ_i jednako 0.93 što je blizu 1.

3.4 Log-normali Model

Sljedeći procjenitelj kojeg ćemo razmatrati temelji se na pretpostavci da logaritam relativnih rizika ($\beta_i = \log(\theta_i)$) ima višedimenzionalnu normalnu distribuciju (generalizacija jednodimenzionalne normalne distribucije) sa očekivanjem ϕ i varijancom σ^2. Valja napomenuti da u slučaju procjenjivanja log-relativnih rizika ne uzimamo $\log(O_i/E_i)$ nego $\log((O_i + 1/2)/E_i)$ zbog problema definiranosti logaritma kada je broj slučajeva u nekoj regiji 0.

Pri procjeni parametara za normalnu distribuciju koristimo se koristimo se empiriskim Bayesovim procjeniteljem od β_i na sljedeći način:

$$\hat{\beta}_i = b_i = \frac{\hat{\phi} + (O_i + \frac{1}{2})\hat{\sigma}^2 \log[(O_i + \frac{1}{2})/E_i] - \frac{\hat{\sigma}^2}{2}}{1 + (O_i + \frac{1}{2})\hat{\sigma}^2}$$

gdje su $\hat{\phi}$ i $\hat{\sigma}^2$ prethodne procjene očekivanja i varijance, redom. Do njih dolazimo na sljedeći način:

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} b_i = \overline{b}$$

i

$$\hat{\sigma}^2 = \frac{1}{2}\left\{\hat{\sigma}^2 \sum_{i=1}^{n}[1 + \hat{\sigma}^2(O_i + 1/2)]^{-1} + \sum_{i=1}^{n}(b_i - \hat{\phi})^2\right\}$$

Procjene b_i se postepeno ažuriraju sve do konvergencije. Procjenitelj za θ_i je $\hat{\theta}_i = exp\{\hat{\beta}_i\}$. Primjetimo da se sada informacije pri procijeni posuđuju od parametara ϕ i σ^2, tj. rezultirajuća procjena je kombinacija lokalne procjene od log relativnog rizika i ϕ. Ovaj model je puno kompleksniji i ne možemo ga jednostavnije zapisati kao u prethodnom slučaju. Više detalja o ovom modelu se može naći u [4].

```
########################## Empirical Bayes Smoothing USING LOG-NORMAL
ebln<-lognormalEB(shp$Observed,shp$Expected)
shp$EBLN<-exp(ebln$smthrr)
```

Slika 3.6: Primjer koda u R-u za Log-normali model

3.5 Marshallov globalni EB procjenitelj

Marshall je u radu [10] razvio novi empiriski Bayesov procjenitelj pretpostavljajući da relativni rizici θ_i imaju zajedničku apriornu srednju vrijednost μ i varijancu σ^2, ali bez navođenja bilo kakve distribucije. Dobivamo sljedeće:

$$\hat{\theta}_i = \hat{\mu} + C_i(SMR_i - \hat{\mu}) = (1 - C_i)\hat{\mu} + C_i SMR_i$$

gdje je

$$\hat{\mu} = \frac{\sum_{i=1}^{n} O_i}{\sum_{i=1}^{n} E_i},$$

i

$$C_i = \frac{s^2 - \hat{\mu}/\overline{E}}{s^2 - \hat{\mu}/\overline{E} + \hat{\mu}/E_i}.$$

Ovdje $\overline{E}$ označava srednju vrijednost E_i-ova, a s^2 uobičajenu nepristranu procjenu varijance od SMR_i-ova. Loša strana pri ovakvom izračunu je ta što možemo dobiti negativne procjene kada je $s^2 < \hat{\mu}/\overline{E}$, no u tom slučaju uzimamo $\hat{\theta}_i = \hat{\mu}$.

Ovaj procjenitelj opet uvelike ovisi o vrijednostima E_i. Ukoliko je ta vrijednost velika, SMR_i će biti pouzdanija procjena, a C_i će biti blizu 1 što će rezultirati da naš procjenitelj pridoda veću "težinu" na SMR_i. S druge strane, ukoliko je E_i relativno mali, naš procjenitelj će pridodati sada veću "težinu" na prethodnu procjenu srednje vrijednosti, no tada "posuđujemo" više informacija od svih ostalih područja.

Sada ćemo prikazati rezultate gore opisanih metoda te pokušati uočiti neke razlike pomoću grafike.

3.5. Marshallov globalni EB procjenitelj

```
########################## Marshall global EB Estimator
EBMarshall<-EBest(shp$Observed,shp$Expected)
shp$EBMarshall<-EBMarshall[,2]
```

Slika 3.7: Primjer koda u R-u za Marshallov globalni model

```
> podaci[,1]
 [1] "Sibensko-kninska"       "Bjelovarsko-bilogorska" "Brodsko-posavska"      "Dubrovacko-neretvanska" "Grad Zagreb"            "Istarska"
 [7] "Karlovacka"             "Koprivnicko-krizevacka" "Krapinsko-zagorska"    "Licko-senjska"          "Medimurska"             "Osjecko-baranjska"
[13] "Pozesko-slavonska"      "Primorsko-goranska"     "Sisacko-moslavacka"    "Splitsko-dalmatinska"   "Varazdinska"            "Viroviticko-podravska"
[19] "Vukovarsko-srijemska"   "Zadarska"               "Zagrebacka"
> shp$SMR
 [1] 1.3623524 0.7004648 0.9897953 0.9712727 1.0068245 0.2227262 1.0618362 0.6758638 0.5435408 0.8755795 1.3082827 0.6879644 0.5607611 1.7689305 0.7499208
[16] 1.2143541 1.3724894 0.4385883 0.7495577 1.1032341 1.1627212
> shp$EBLN
 [1] 1.3617168 0.7007120 0.9896441 0.9711233 1.0067959 0.2236183 1.0615792 0.6761583 0.5440232 0.8754914 1.3077524 0.6880710 0.5616023 1.7685719 0.7500375
[16] 1.2142449 1.3720999 0.4398076 0.7496717 1.1030225 1.1625858
> shp$EBPG
 [1] 1.3615183 0.7008830 0.9897100 0.9712062 1.0068058 0.2234002 1.0616178 0.6763297 0.5441366 0.8758152 1.3076165 0.6881384 0.5618283 1.7683430 0.7501597
[16] 1.2142294 1.3719743 0.4398835 0.7497907 1.1030325 1.1625770
> shp$EBMarshall
 [1] 1.3615225 0.7011130 0.9898126 0.9713276 1.0068225 0.2236059 1.0617146 0.6765696 0.5443765 0.8762197 1.3076381 0.6882309 0.5622685 1.7682909 0.7503149
[16] 1.2142424 1.3719745 0.4403452 0.7499418 1.1030915 1.1626015
```

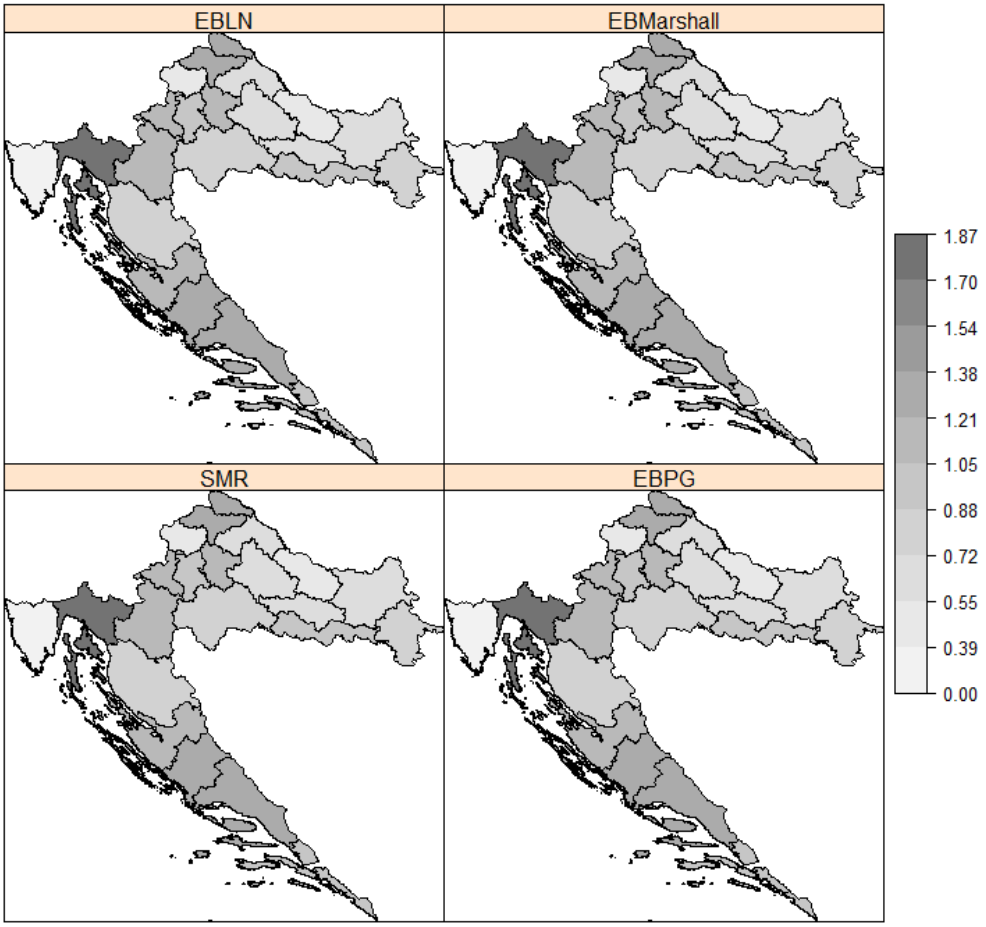
Slika 3.8: Ispis procijenitelja po županijama u Republici Hrvatskoj koje smo računali u R-u (SMR, Possion Gamma, Log-normalni te Marshallov EB)

Na slici 3.8 vidimo da se sve vrijednosti različite od početnog SMR, ali ne razlikuju se previše. Vidljivo je da su sve pomaknute prema globalnoj sredini što je bio i cilj kojeg smo pokušali napraviti. Nažalost, na slici 3.9 nije vidljiva razlika među različitim modelima.

Svi ovi procijenitelji pokušavaju napraviti zaglađene procijene na način da posuđuju informacije od svih ostalih regija. Kako se bavimo problemom mapiranja zaraze, a posebno u ovom slučaju mapiranjem zaraze koronavirusom očigledno je da posuđivanje informacija od svih regija nema pretjeranog smisla. Znamo da se koronavirus prenosi bliskim kontaktom pa na primjer kada računamo relativan rizik za Splitsko-dalmatinsku županiju razumno je vrijenosti iz Slavonije ne uzimati u obzir.

Sada ćemo posuđivati informacije samo od nekog podskupa svih regija koje su relativno blizu jedna drugoj. Uobičajeno se posuđuju informacije samo od regija koje su susjedne, tj. imaju zajedničku granicu. U ovom radu ćemo to definirati na nešto drugačiji način. Nažalost, ovakvi modeli su dosta kompleksniji, zahtjevaju korištenje dodatnih programa te lako takvi izračuni

3.5. Marshallov globalni EB procjenitelj



Slika 3.9: Grafički prikaz procijenitelja po županijama sa sijenčenjem od 10 razina kako bi uočili razlike među županijama

mogu postati problematični.

Što bi se dogodilo da smo ovde proizvoljno ispermutirali županije?Bi li rezultati bili drugačiji?

Uočimo da bi svi ovi modeli dali iste rezultate do na proizvoljnu permutaciju regija. Definiranjem susjedstva to više nece biti slučaj. Sada će prostorana lokacija pojedine regije biti od značaja za razliku od prije.

Poglavlje 4

Prostorno strukturirani statistički modeli

Iako posuđivanje informacija od svih regija u nekim slučajevima ima smisla, obično je bolje posuđivati informacije samo od nekog podskupa koji je definiran na dobar način. Za tu svrhu potrebno je definirati matricu susjedstva županija. Ne postoji način koji je najbolji u definiranju takvog susjedstva. Uzimajući u obzir geografiju, centraliziranost Hrvatske, odsječenost Dubrovačko-neretvanske županije definiramo susjedstva na način prikazan na slici 4.1. Valja napomenuti da se ovim vezama među županijama mogu pridodijeliti različite "težine", no mi smo ih ovdje definirali binarno radi jednostavnosti.

$$w_{ij} = \begin{cases} 1, & \text{ako je } i \text{ susjed od } j, \\ 0, & \text{inače.} \end{cases} \tag{4.1}$$

Veći broj veza oko Zagreba je opravdan iz razloga što županije na tom području jesu relatvno bliže jedna drugoj nego ostale iz Hrvatske, a također je povećana njihova gravitacija prema Zagrebu što je pogodno za širenje zaraze.

Slika 4.1: Susjedstva županija u Republici Hrvatskoj

Za stvaranje ovih prostornih "težinskih" veza između regija koristili smo se paketom **spdep**, funkcijama **knearneigh** i **dnearneigh**. Funkcija **knearneigh** nam služi da dobijemo k najbližih susjeda dok funkcija dnearneigh poveže sve regije čiji se centri nalaze unutar granica koje zadamo koristeći se Euklidskom udaljenošću d_2. Služili smo se kombinacijom ovih dviju funkcija s ciljem da svaka regija ima barem jednog susjeda. Bitno je napomenuti da uz različite definicije susjedstva, a uz isti model, dobijamo različite rezultate.

Kako iz prethodnih modela nismo dobili dovoljno dobru zaglađenu procjenu relativnog rizika u potrazi smo za boljim rješenjem. Više standardnih procjenitelja zaglađenog rizika koji posuđuju informacije na lokalnoj razini može se dobiti korištenjem prostorne autoregresivne i uvjetne autoregresivne specifikacije (CAR). Ideja ovih modela je da uvjetuju da relativni rizik u nekom području bude sličan vrijednostima susjednih područja.

```
> summary(nb_5)
Neighbour list object:
Number of regions: 21
Number of nonzero links: 136
Percentage nonzero weights: 30.839
Average number of links: 6.47619
Link number distribution:

 1  3  4  6  7  8 10 12
 2  3  3  1  1  6  4  1
2 least connected regions:
3 5 with 1 link
1 most connected region:
14 with 12 links
```

Slika 4.2: Sažetak objekta tipa "nb" u R-u koji definira susjedstva županija u Republici Hrvatskoj.

4.0.1 Uvjetna autoregresivna specifikacija (CAR)

Za skup slučajnih varijabli $\{v_i\}_{i=1}^{n}$ uvjetna autoregresivna specifikacija (engl. conditional autoregressive specification, CAR) se može opisati kao

$$v_i|v_{-1} \sim N\left(\sum_{j\sim i} \frac{w_{ij}v_j}{\sum_j w_{ij}}, \hat{\sigma_v}^2 / \sum_j w_{ij}\right)$$

gdje su w_{ij} težine veza koje mjere snagu odnosa (susjeda) između regija i i j,$j \sim i$ skup svih susjeda od i, a $\hat{\sigma_v}^2$ označava uvjetnu varijancu od CARa.

CAR se često koristi kao apriorna distribucija slučajnih prostornih efekata i ona može voditi ka aposteriornoj distribuciji pod nekim ograničenjima.

Dakle, umjesto posuđivanja informacija od svih područja, posuđujemo samo težinske informacije od susjednstva normalizirane sumom svih korištenih veza.

U R-u koristimo funkciju **car.normal** za dobiti prostorne efekte na gore opisan način.

4.1 Poisson-gamma model sa MCML

Sada ćemo primjeniti MCML algoritam u svrhu generiranja aposteriorne distibucije relativnog rizika i parametara ν i α iz Poisson-gamma modela korištenjem WinBUGS programa. Sve podatke koje imamo moramo prvo

```
model
{
for(i in 1:N)
{
observed[i]~dpois(mu[i])
mu[i]<-theta[i]*expected[i]
theta[i]~dgamma(nu, alpha)
}
nu~dgamma(.01, .01)
alpha~dgamma(.01, .01)
}
```

Slika 4.3: PG-model za WinBUGS gdje su apriore distribucije od ν i α gamma s parametrima 0.01,0.01

pretvoriti u oblik pogodan za korištenje WinBUGS-a. Također prije pokretanja programa moramo unijeti i inicijalne vrijednosti za parametre.

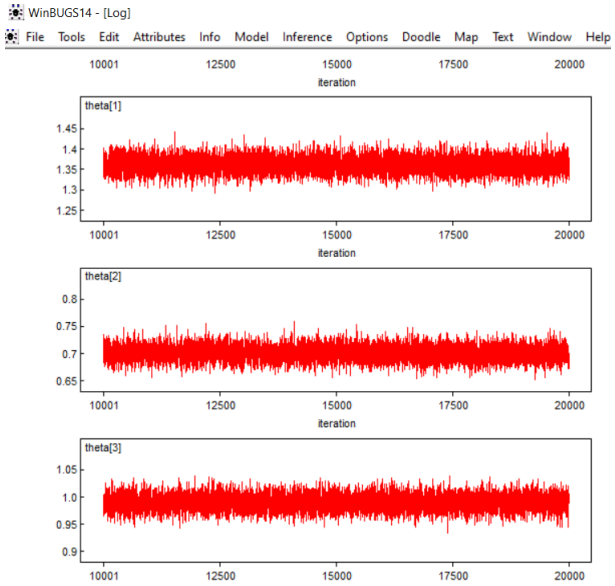
Pokrenut ćemo 200000 simulacija gdje ćemo izbaciti prvih 100000 (burn-in) i zadržat ćemo svaku desetu vrijednost kako bi izbjegli autokorelaciju i poboljšali konvergenciju. Iznosimo samo glavnu liniju koda:

```
MCMCres<- bugs(data=d, inits=list(list(nu=1, alpha=1)),
               working.directory=wdir,
               parameters.to.save=c("theta", "nu", "alpha"),
               n.chains=1, n.iter=200000, n.burnin=100000, n.thin=10,
               model.file=pgmodelfile,
               bugs.directory=BugsDir,debug = TRUE)
```

Slika 4.4: Kod u R-u sa funkcijom bugs, za inicijalne vrijednosti ν i α stavljamo 1 i korsitimo gore opisan PG model

4.1. Poisson-gamma model sa MCML

Dobivamo rezultate i vizualne prikaze simulacija. Vrijednosti koje smo



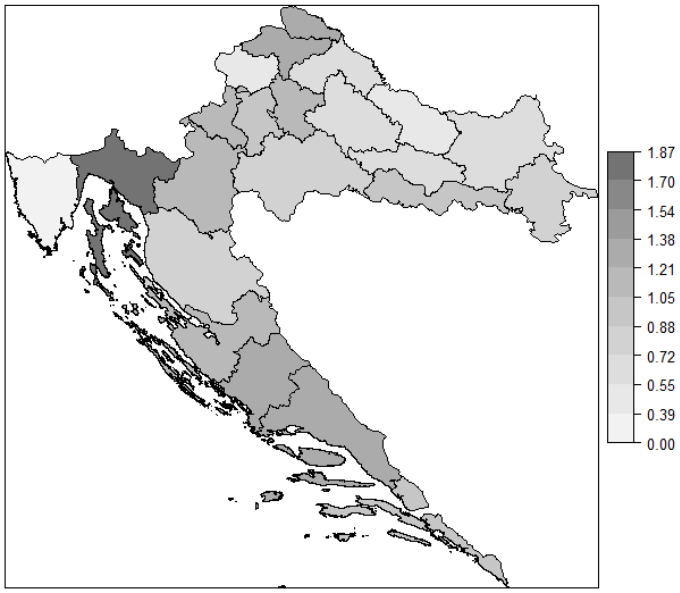
Slika 4.5: Kod u R-u sa funkcijom bugs

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	5.866	1.869	0.02676	2.735	5.697	9.978	10001	10000
deviance	236.8	6.501	0.05897	226.0	236.0	251.5	10001	10000
nu	5.451	1.661	0.02434	2.702	5.293	9.124	10001	10000
theta[1]	1.361	0.01942	1.978E-4	1.324	1.361	1.4	10001	10000
theta[2]	0.7007	0.01369	1.402E-4	0.6741	0.7008	0.7275	10001	10000
theta[3]	0.9899	0.01426	1.33E-4	0.9623	0.9896	1.018	10001	10000
theta[4]	0.9713	0.01495	1.317E-4	0.9422	0.9713	1.001	10001	10000
theta[5]	1.007	0.006021	5.864E-5	0.995	1.007	1.018	10001	10000
theta[6]	0.2232	0.005589	5.428E-5	0.2123	0.2232	0.2343	10001	10000
theta[7]	1.061	0.01579	1.733E-4	1.03	1.061	1.093	10001	10000
theta[8]	0.6763	0.01347	1.216E-4	0.65	0.6761	0.7032	10001	10000
theta[9]	0.5439	0.01099	1.091E-4	0.5227	0.5438	0.5656	10001	10000
theta[10]	0.876	0.02328	2.262E-4	0.8301	0.8758	0.9215	10001	10000
theta[11]	1.308	0.01796	1.794E-4	1.273	1.308	1.343	10001	10000
theta[12]	0.688	0.008592	8.624E-5	0.6712	0.688	0.7045	10001	10000
theta[13]	0.5619	0.01529	1.451E-4	0.5322	0.5619	0.5923	10001	10000
theta[14]	1.768	0.01344	1.335E-4	1.742	1.768	1.795	10001	10000
theta[15]	0.7502	0.01197	1.256E-4	0.7271	0.7502	0.7742	10001	10000
theta[16]	1.214	0.00871	9.347E-5	1.197	1.214	1.231	10001	10000
theta[17]	1.372	0.01531	1.615E-4	1.342	1.372	1.402	10001	10000
theta[18]	0.4394	0.013	1.271E-4	0.4145	0.4393	0.4652	10001	10000
theta[19]	0.7497	0.01181	1.177E-4	0.7267	0.7497	0.773	10001	10000
theta[20]	1.103	0.01365	1.56E-4	1.077	1.103	1.13	10001	10000
theta[21]	1.163	0.01029	9.4E-5	1.143	1.163	1.183	10001	10000

Slika 4.6: Kod u R-u sa funkcijom bugs

dobili su parametri α i ν te θ_i gdje je $i = 1, ..., 21$. Uočimo da smo za svaku od ovih vrijednosti također i dobili 95% interval pouzdanosti. Također, dostupne su nam i vrijednosti *mean* i *median* za svaki parametar.

Pogledajmo sada grafički median od θ_i za svaku županiju. Slično kao i prije, uočavamo najveći rizik u Primorsko-Goranskoj županiji, a potom u



Slika 4.7: Median od θ_i po županijama

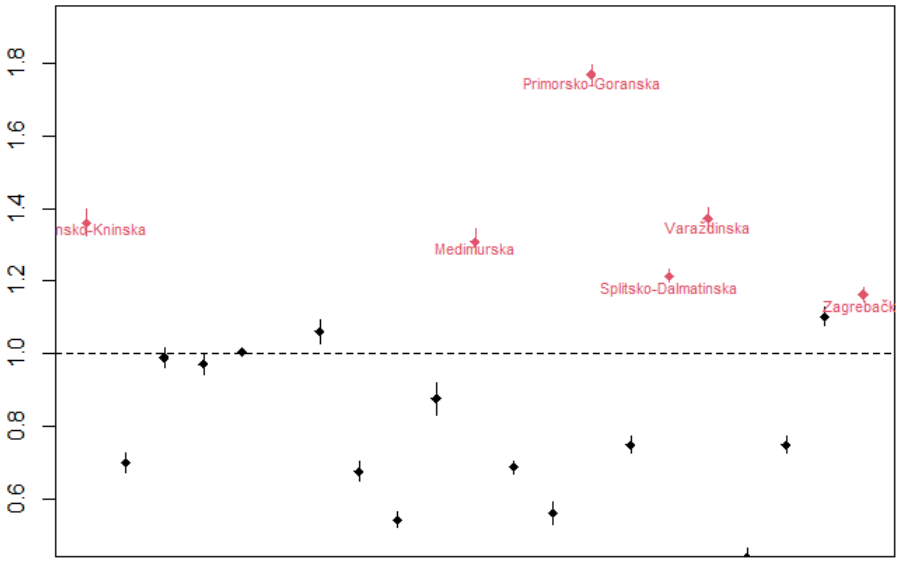
Splitsko-Dalmatinskoj županiji, Međimurkoj i Varaždinskoj.

Iako su točkovne procjene relativnih rizika obično vrlo korisne, za većinu primjena bolje je dati interval pouzdanosti, jer se može koristiti za otkrivanje područja značajno velikog rizika, ukoliko cijeli interval pouzdanosti leži poviše 1.

Stoga ćemo na slici 4.8 pokazati interval pouzdanosti za ove vrijednosti. Sa crnim točkama ćemo prikazati medijan, a okomita crta će označavati 95% interval.

Možemo uočiti da su intervali pouzdanosti dosta manji nego kod SMR-a na slici 2.2. Razlog tome je postignuto zaglađenje i veća pouzdanost modela kojim procjenjujemo relativan rizik. Iz ovakvog prikaza, sa većom sigurnošću možemo zaključiti u kojim županijama je povećan rizik od zaraze koronavirusom. Na kraju, rezultati koje smo dobili zajedno sa sumarnom statikom i grafovima su spremljeni unutar radnog direktorija kojeg smo unaprijed pos-

tavili u log datoteke. Stvaraju se dvije datoteke: ODC i ASCII datoteka. Korištenjem paketa **coda** možemo pristupiti podacima bez da ponovo pokrećemo model i program WinBUGS.

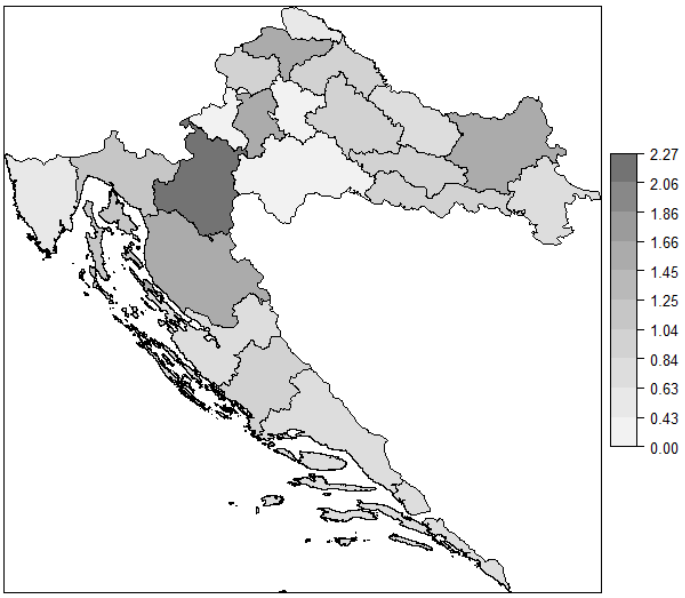


Slika 4.8: Intervali pouzdanosti po županijama

Napomena 4.1 *Nužno je provjeriti je li Markovljev lanac konvergirao. To ćemo raditi pomoću Gewekeovog kriterija [5] koji se temelji na jednakosti srednjih vrijednosti prvog i zadnjeg dijela Markovljevog lanca. Uobičajeno je da se gleda jednakost srednje vrijednosti prvih 10% i zadnjih 50% lanca. Ako su uzorci izvučeni iz stacionarne distribucije lanca, dvije srednje vrijednosti su jednake i Gewekeova statistika ima asimptotski standardnu normalnu distribuciju. Rezultat koji dobijemo ovim testom je broj, a ukoliko se on nalazi u intervalu* $\langle -1.96, 1.96 \rangle$ *to ukazuje na konvergenciju. Svaka vrijednost izvan toga ukazuje na nedostatak konvergencije.*

4.1. Poisson-gamma model sa MCML

Sada ćemo identičan postupak provesti za umrle od koronavirusa. Rezultati su sljedeći:



Slika 4.9: Zaglađene procjene rizika od smrti po županijama u Republici Hrvatskoj

```
> geweke.diag(ncoutput[,c("deviance", "alpha","nu", "theta[1]","theta[11]","theta[20]")])

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

 deviance     alpha        nu  theta[1] theta[11] theta[20]
  -1.1434    1.4193    1.6003    0.3207    1.4508   -1.1620
```

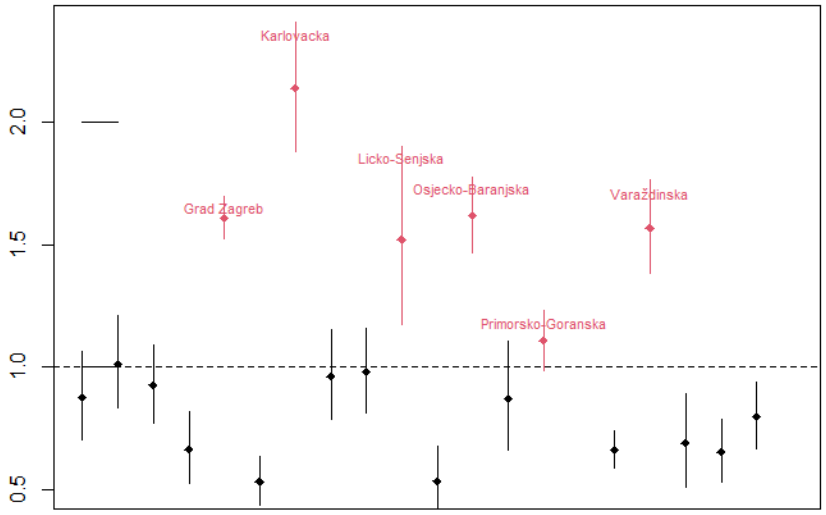
Slika 4.10: Gewekeov test konvergencije lanca

Županija koja imaja najveći rizik smrtnosti je Karlovačka županija, a slijede je Grad Zagreb, Ličko-Senjska, Osječko-Baranjska i Varaždinska. Takav poredak lakše je vidjeti na slici 4.11 gdje se koriste intervali pozudanosti.

Uočimo da je Primorsko-Goranska imala najveći rizik što se tiče morbiditeda, no kada gledamo mortalitet tek je sedma po redu.

Slika 4.10 ukazuje da naš lanac prolazi Gewekeov test što je dobro za pouzdanost dobivenih rezultata.

4.1. Poisson-gamma model sa MCML



Slika 4.11: Intervali pouzdanosti po županijama u Republici Hrvatskoj

Karlovačka županija ima medijalnu vrijednost rizika od smrti 2.14. Ovaj broj je dosta udaljen od 1 koja je prosjek za cijelu Hrvatsku i ukazuje da je smrtnost u toj županiji za 114% veća od prosjeka. Ostale gore spomenute županije sa velikim rizikom imaju otprilike 60% veću smrtnost od prosjeka.

U nastavku ovog rada ćemo pokušati naći uzrok povećane smrtnosti baš u tim županijama.

4.2 Bayesova linearna regresija sa MCML

U ovom dijelu koristi ćemo se sa MCML uz model na slici 4.12. Dakle,

```
model
{

  for(i in 1:N)
  {
     observed[i] ~ dpois(mu[i])
     log(theta[i]) <-  alpha + beta*pp30_60[i] + u[i] + v[i]
     mu[i] <- expected[i]*theta[i]

     u[i] ~ dnorm(0, precu)
  }

  v[1:N] ~ car.normal(adj[], weights[], num[], precv)

  alpha ~ dflat()
  beta ~ dnorm(0,1.0E-5)
  precu ~ dgamma(0.001, 0.001)
  precv ~ dgamma(0.1, 0.1)

  sigmau<-1/precu
  sigmav<-1/precv
}
```

Slika 4.12: Model sa prostorno ovisnim i neovisnim učincima

modeliramo

$$
\begin{aligned}
O_i &\sim Po(\mu_i), \\
\mu_i &= \theta_i E_i, \\
\theta_i &= \exp\left\{\alpha + \beta * kovarijabla_i + u_i + v_i\right\}.
\end{aligned}
\tag{4.2}
$$

Opažene frekvencije u ovom modelu imaju Poissonovu distribuciju sa ovakvim parametrom μ_i, u_i je prostorno nezavisni dio sa normalnom distribucijom s očekivanjem 0, a v_i prostorno zavisni učinak te koristimo CAR specifikaciju i susjedstva županija koja smo definirali na početku Poglavlja 4 i prikazali na slici 4.1.

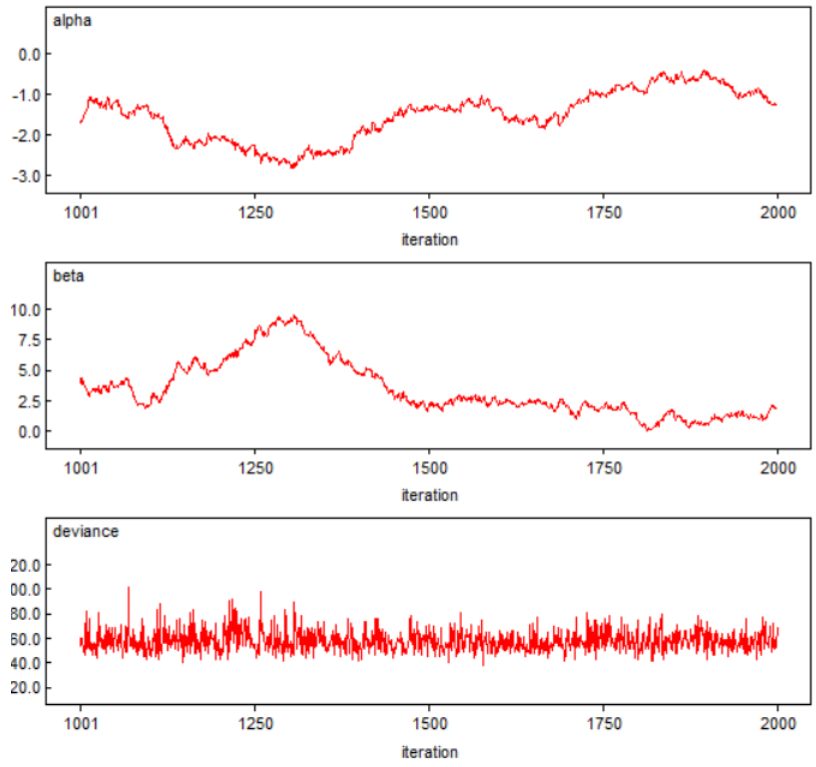
U ovakvom modelu ćemo uključivati nekakve kovarijable čiji će smisao biti u procijenjivanju i uklanjanju učinaka potencijalnih zbunjujućih čimbenika (engl. confounder). Važnost kovarijable ćemo gledati na način da gledamo procjenjenu vrijednost koeficijenta koji joj je pridružen. U praksi ćemo koristiti 95%-tni interval pouzdanosti za koeficijent te u slučaju da ne sadrži vrijednost 0 zaključujemo da je promatrana kovarijabla značajna. Ukoliko je 95%-tni interval iznad 0 zaključujemo da postoji pozitivna korelacija između rizika od bolesti i promatrane kovarijable.

Na osnovu razmatranja situcije o zarazi do sad logično je za pretpostaviti da udio starijih od 60 godina ima smisla biti kovarijabla za mortalitet od koronavirusa te s toga sa prop60 ćemo označavati udio starijih od 60 godina po svakoj županiji. Pokrećemo MCMC sa gore opisanim modelom uz 20000 simulacija da bi dobili aposteriorne distribucije parametara modela: α, β, precu i precv. Uočimo da za apriorne vrijednosti stavljamo $0, 0, 0.001, 0.001$ redom kojeg smo naveli gore.

```
d <- list(N = N, observed = shp$umrli, expected = shp$exp_umrli,pp30_60 = shp$prop60,
          adj = shp$nb$adj, weights = shp$nb$weights,num = shp$nb$num)
inits <- list(u = rep(0, N), v = rep(0, N), alpha = 0,beta = 0, precu = 0.001, precv = 0.001)
wdir <- paste(getwd(), "/BYM3", sep = "")
if (!file.exists(wdir)) { dir.create(wdir) }
MCMCres_3 <- bugs(data = d, inits = list(inits),working.directory = wdir,
                  parameters.to.save = c("theta","alpha", "beta", "u", "v", "sigmau", "sigmav"),n.chains = 1,
                  n.iter = 20000, n.burnin = 10000, n.thin = 10, model.file = bymmodelfile,
                  bugs.directory = BugsDir,debug=TRUE)
```

Slika 4.13: Model sa kovarijablom prop60

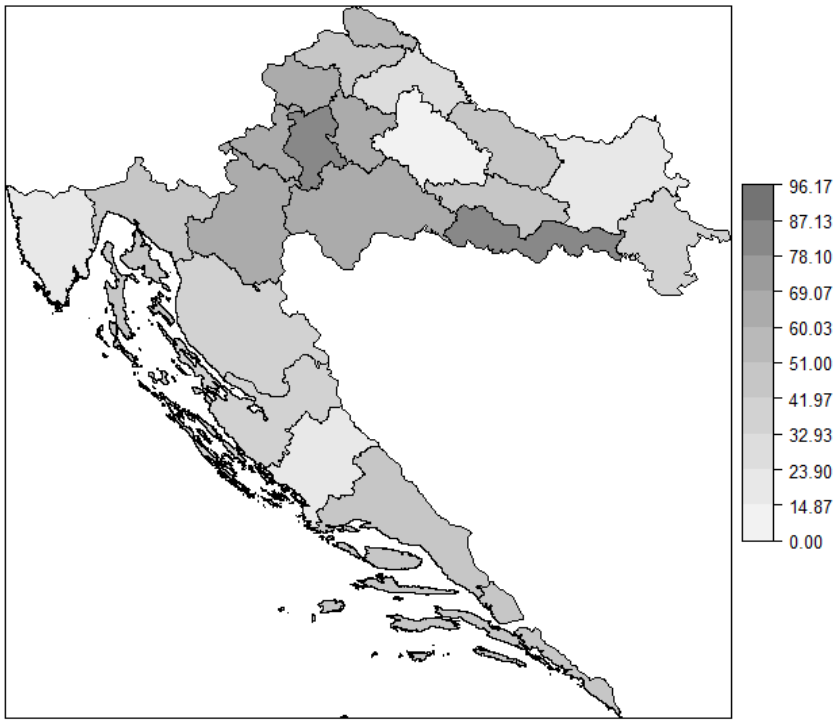
Interval 95%-tne pouzdanosti za β je $\langle 0.52, 8.99\rangle$ dok je medijan 2.56 što bi ukazivalo na pozitivnu korelaciju između udjela starijih od 60 i umrlih od koronavirsa. No, kada napravimo Gewekeow test dobijamo rezultat koji govori da lanac nije konvergirao ka stacionarnoj distribuciji pa ne možemo sa sigurnošću donijeti nikakav zaključak. S obzirom na definiciju Gewekeovog testa, sličan zaključak možemo donijeti promatrajući sliku 4.14. Dakle, ovim putem nismo dobili statistički značajan zaključak.



Slika 4.14: Grafički prikaz vrijednosti MCML simulacija za parametre modela (4.2).

4.2. Bayesova linearna regresija sa MCML

U potrazi za statističkim značajnim kovarijablama koje možemo povezati sa povećanim rizikom mortaliteta dolazimo do sljedećeg. Na slici 4.15



Slika 4.15: Grafički prikaz kvalitete zraka po županijama u Republici Hrvatskoj

je prikazana kvaliteta zraka po županijama u prvoj polovici 2021. godine. Nakon mnogo različitih pokušaja izdvajamo povezanost kvalitete zraka sa povećenom stopom smrtnosti od koronavirusa.

Podatci su preuzeti u dogovoru sa Švicarskom tehnološkom tvrtkom za kvalitetu zraka IQAir i raspoloživim stanicama za mjerenje po Hrvatskoj. Internet stanicu tvrtke IQAir možete pronaći na linku. Početna ideja je bila ispitivanje točno određenih zagađivača zraka kao što je PM2.5, CO, NO_2, CO_2, no zbog nedostatka informacija u svim postajama istraživanje provodimo sa indeksom US AQI koji predstavlja sve zagađivače ukomponirane u

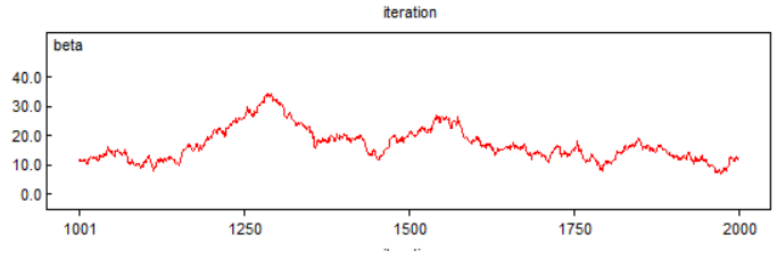
jedan broj. Na slici 4.15 tamnije obojane županije su one gdje je zrak više zagađen.

Sada u modelu (4.2) za kovarijablu stavljamo kvalitetu zraka. Dobijamo rezultate:

Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha	-1.579	0.4556	0.07996	-2.254	-1.636	-0.8142	1001	1000
beta	16.92	5.713	0.9899	8.747	15.69	31.34	1001	1000
deviance	156.5	9.035	0.3849	143.9	155.4	175.1	1001	1000
sigmau	3.369	1.368	0.09685	1.424	3.119	6.909	1001	1000
sigmav	0.6351	0.7047	0.09285	0.0556	0.3792	2.802	1001	1000
theta[1]	0.8697	0.09503	0.003103	0.7017	0.8641	1.065	1001	1000
theta[2]	1.016	0.09527	0.00298	0.8342	1.015	1.203	1001	1000

Slika 4.16: Rezultati za parametre



Slika 4.17: Vrijednost parametra β kroz simulacije

Uočimo da je 95%-tni interval pouzdanosti $\langle 8.7, 31.3 \rangle$ te da mu je medijan 15.7. Ovakve vrijednosti ukazuju na visoku korelaciju između loše kvalitete zraka i rizika od smrtnosti.

Treba još provjeriti konvergenciju Markovljevog lanca.

```
> geweke.diag(ncoutput[,c("deviance", "alpha", "beta", "theta[5]", "theta[10]", "theta[15]")])

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

 deviance     alpha      beta  theta[5] theta[10] theta[15]
   0.5682    2.0843   -1.2642    1.0301    0.2072   -0.1185
```

Slika 4.18: Provjera konvergencije lanca sa Gewekeovim kriterijem

Na slici 4.18 vidimo da parametar beta prolazi Gewekeov test, slično smo mogli zaključiti gledanjem slike 4.17

4.2. Bayesova linearna regresija sa MCML

Analiziranjem podataka o zagađenju zraka po županijama te povećanim rizikom mortaliteta sa slike 4.11 možemo primjetiti da je glavni zagađivač zraka PM2.5. Naime, PM2.5 su male čestice u zraku koje su manje od 2.5 mikrometra. Te male čestice mogu se dugo zadržati u zraku i lako uđu u pluća i dišni sustav što otežava disanje te pogoršaje plućne bolesti. Kako je koronavirus bolest koja najviše pogađa dišni sustav te očito uz pomoć čestica PM2.5 može lakše uzrokovati smrtnost.

Cijeli kod koji je korišten pri izradi ovog diplomskog rada zajedno sa svim potrebnim podacima za učitati te slikama može se pronaći na github profilu na ovom linku.

Također tu možete pronaći i kod u pythonu koji je korišten kako bi se lakše preradili strojno čitljivi podaci sa stranice koronavirus.hr.

Zaključak

Izvorni cilj ovog rada je bio upoznavanje osnovih metoda prostorne statistike te otkrivanje županija u Hrvatskoj visokog rizika od zaraze koronavirusom. Koristili smo 5 različitih metoda i iznijeli sve prikaze koje smo stvarali uz pomoć programskog jezika R.

Vodeći se tim prikazima zaključili smo sljedeće: što se tiče morbiditeta, najveći rizik ima Primorsko-Goranska županija, a slijede je Međimurska i Varaždinska. Što se tiče mortaliteta, najveći rizik ima Karlovačka županija, a slijede Varaždinska, Grad Zagreb, Ličko-Senjeska i Osječo-Baranjska.

Također, koristili smo Bayesovu statistiku i Monte Carlo Markovljeve lance za simuliranje aposteriornih distribucija parametara naših modela kako bismo potkrijepili početne slutnje.

Pokušali smo pronaći kovarijable koje mogu utjecati na relativni rizik od mortaliteta. Prvo, koristeći dobnu podjelu stanovništva, pokušali smo pokazati da je populacija starijih od 60 statistički značajna varijabla za povećan mortalitet od korone, ali nismo uspjeli postići konvergenciju Markovljevog lanca. Nakon toga smo prikazali prostornu zagađenost zraka u prvoj polovici 2021. godine te ispitali postoji li povezanost razine zagađenosti zraka sa mortalitetom. Model kojeg smo proveli ukazao je na statističku značajnost uz konvergenciju Markovljevog lanca. U županijama gdje je zagađenost zraka bila najveća, ponalazimo istog glavnog zagadivača, PM2.5. Očito, kako te

sitne čestice PM2.5 uzrokuju probleme u dišnom sustavu, uz kombinaciju sa koronavirusom dolazi do više smrtnih slučajeva.

Literatura

[1] Bivand, R., Pebesma, E., Gómez-Rubio, V. (2013) *Applied Spatial Data Analysis with R*, Drugo izdanje, New York, Springer

[2] Bradvica, M.K. (2020) *Monte Carlo Markovljevi lanci.* Diplomski rad. Osijek: Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku. Dostupno na `https://zir.nsk.hr/islandora/object/mathos:493`

[3] Braić, S. (2020) *Uvod u vjerojatnost.* Split: Sveučilište u Splitu, Prirodoslovno-matematički fakultet.

[4] Claytonand, D., Kaldor, J. (1987) *Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping.* Biometrics 43, 671-681.

[5] Fang, Q., *A Brief Introduction to Geweke's Diagnostics.* Dostupno na `https://www.math.arizona.edu/~piegorsch/675/GewekeDiagnostics.pdf`[20.kolovoz 2022]

[6] Geographic information system. Dostupno na `https://en.wikipedia.org/wiki/Geographic_information_system`[20.kolovoz 2022]

[7] Gimenez, O., Bayesova analiza sa skrivenim Markovljevim modelom. Dostupno na: https://oliviergimenez.github.io/banana-book/crashcourse.html [2.rujna 2022]

[8] Huzak M. (2006) *Vjerojatnost i matematička statistika.* Zagreb: Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet.

[9] Inferencijalna statistika, točkovna i intervalne procijene. Dostupno na https://www.grad.unizg.hr/_download/repository/Procjene_parametara.pdf[20.svibanj 2023]

[10] R. J. Marshall (1991) *Mapping disease and mortality rates using Empirical Bayes estimators.* Applied Statistics, 40:283–294.

[11] Programski jezik R. Dostupno na https://www.r-project.org/

[12] Shabir, O., *What is Case Fatality Rate (CFR)?* https://www.news-medical.net/health/What-is-Case-Fatality-Rate-(CFR).aspx[2.rujna 2022]

[13] Shapefile Hrvatske: https://www.diva-gis.org/datadown

[14] Weinberger, K., *Estimating probablilites from data.* Dostupno na https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote04.html [28.kolovoza 2022]

[15] WinBUGS program https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/