Modeliranje pandemije COVID-a

Šarić, Nikolina

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Nikolina Šarić

MODELIRANJE PANDEMIJE COVID-A

Diplomski rad

Voditelj rada:
doc. dr. sc. Pavle Goldstein

Zagreb, studeni, 2021.

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. _____

Zahvaljujem svima koji su svojim savjetima, prijedlozima i podrškom pridonijeli izradi ovog rada. Zahvaljujem svom mentoru, doc. dr. sc. Pavlu Goldsteinu, na pomoći, strpljenju, trudu i vodstvu tijekom izrade rada. Zahvaljujem svojim prijateljima i kolegama koji su mi olakšali ovaj put i pridonijeli uspješnom završetku studija. Posebna zahvala mojoj majci i sestri koje su mi pružale veliku podršku i ljubav tijekom mog školovanja.

Sadržaj

Poglavlje 1

Uvod

Svojom pojavom na početku 2020. godine koronavirus je u potpunosti promijenio svijet kojeg smo nekad znali. Nitko nije očekivao da se nešto slično moze dogoditi u 21. stoljeću - u vremenu visoke tehnologije i napredne medicine, a neke stvari možda nikad neće biti kao prije. Ono što je iznenadilo svakog pojedinca na zemlji nije pojava virusa, nego njegova nepredvidljivost i duljina trajanja ove nesvakidašnje situacije.

Kako se ovoj situaciji još uvijek ne nazire kraj, o ovom se virusu dosta još ne zna i nije istraženo. Zbog toga prirodno je htjeti saznati više kako bismo naučili nešto novo i pobrinuli se da u budućnosti izbjegnemo ponavljanje ovakve situacije.

Mnoge mjere postavljene su s ciljem suzbijanja zaraze koronavirusom. Nažalost, kako još uvijek ne znamo puno, nečija razmišljanja su da neke mjere nisu urodile plodom nego samo pogoršale svakodnevni život, a posljedica toga je nezadovoljstvo i nepovjerenje javnosti. Prirodno je za čovjeka da se boji nečega što ne razumije niti o čemu se nije potrudio naučiti, a kako svi drugačije razmišljamo, sve skupa rezultiralo je tome da pojedinci ne vjeruju da je cjepivo učinkovito i ne žele se cijepiti.

Cijeli današnji svijet jako je povezan. Poruka poslana s jedne strane svijeta odmah je dostavljena na drugu stranu svijeta, vijest u jednoj državi odmah postane vijest u drugoj državi, tisuće kilometara prođe se za par sati pa se tako brzo i zaraza proširila svijetom.

Cilj ovog rada je pokušaj da se dokaže nešto novo iz podataka koji su dostupni, a to su dnevni brojevi zaraze koronavirusom, i naučiti kako bolje pristupati cijeloj situaciji. Kako su članice Europske unije posebno povezane, kako poslovno, tako i privatno, koristit će se podaci za devet država u Europi. Isto kako se lagano i brzo može putovati iz jedne države u drugu te kako jedna država posluje s drugom, pokušat će se otkriti hoće li rast ili pad novih slučajeva jedne utjecati na rast ili pad novih slučajeva u drugoj državi. Da bismo to dokazali matematički, koristit ćemo brojne statističke testove. Ako se pokaže da utjecaj zaraze postoji, modeliranjem ćemo pokazati da se brojke jedne zemlje mogu aproksimirati brojkama druge zemlje. Dodatno, podatke ćemo detaljnije analizirati da bismo stvorili

sugestije za daljnje analize za one koje zanima i žele sami otkriti više čak i s drugim podacima vezanim uz koronavirus.

Na kraju, pokazat će se da se podaci mogu analizirati na drugi način i da se drukčijim pristupom može pokušati razmatrati i druge povezane probleme.

Poglavlje 2

Teorija

2.1 Matematički pojmovi

Da se opravda korištenje matematičkih formula te statističkih testova i modela za izračunavanje utjecaja toka zaraze jedne države na tok zaraze druge države, prvo navodimo sve definicije i teoreme koji su pozadina ovog istraživanja.

2.1.1 Slučajne varijable

Varijable, korelacije i statistički testovi definirani su na određenim prostorima i imaju određena svojstva koja navodimo u ovom poglavlju (vidi [3]).

Definicija 2.1.1. *Neprazna familija $\mathcal{F}$ podskupova od X je σ-prsten (podskupova od X) ako vrijedi*

(a) $A, B \in \mathcal{F} \implies A \setminus B \in \mathcal{F}$

(b) $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{F} \implies \bigcup_{n=1}^{+\infty} A_n \in \mathcal{F}$

Ako $\mathcal{F}$ sadrži cijeli X, onda se $\mathcal{F}$ zove σ-algebra (podskupova od X).

Primjer 2.1.2. *(a) $\mathcal{P}(X)$ je najveća σ-algebra na X.*

(b) Neka je $\mathcal{E} \subseteq \mathcal{P}(X)$. Definiramo σ-prsten generiran familijom $\mathcal{E}$:

$$\sigma_p(\mathcal{E}) := \bigcap_{\substack{\mathcal{F}:\ \mathcal{E}\subseteq\mathcal{F} \\ \mathcal{F}\ \sigma\text{-prsten}}} \tag{2.1}$$

(c) Za svaki $d \in \mathbb{N}$, $\mathcal{I}^d$ je familija d-intervala:

$$\mathcal{I}^d := \{I_1 \times \ldots \times I_d, I_j \text{ je 1-interval, za } j = 1, \ldots, d\} \tag{2.2}$$

Vidimo da je $\mathcal{I}^d$ poluprsten podskupova od $\mathbb{R}^d$.

(d) $\sigma_p(\mathcal{I}^d)$ se označava s $\mathcal{B}(\mathbb{R}^d)$ i naziva Borelova σ-algebra na $\mathbb{R}^d$ ili σ-algebra Borelovih podskupova od $\mathbb{R}^d$.

Definicija 2.1.3. *Izmjeriv prostor je uređen par $(X, \mathcal{F})$, gdje je X skup, a $\mathcal{F} \subseteq \mathcal{P}(X)$ σ-algebra.*

Definicija 2.1.4. *Neka je $(X, \mathcal{F})$ izmjeriv prostor. Mjera je svaka funkcija $\mu : \mathcal{E} \to [0, +\infty]$, gdje je $\emptyset \in \mathcal{E} \subseteq \mathcal{P}(X)$ koja zadovoljava sljedeća dva uvjeta:*

(i) $\mu(\emptyset) = 0$,

(ii) Ako je $(E_n)_{n\in\mathbb{N}}$ niz međusobno disjunktnih skupova iz $\mathcal{E}$ takav da je $\bigcup\limits_{n\in\mathbb{N}} E_n \in \mathcal{E}$, onda je

$$\mu\Big(\bigcup_{n\in\mathbb{N}} E_n\Big) = \sum_{n\in\mathbb{N}} \mu(E_n) \tag{2.3}$$

Mjera je konačna ako je $\mu(X) < +\infty$.
Mjera je vjerojatnosna ako je $\mu(X) = 1$.
Prostor mjere je uređena trojka $(X, \mathcal{F}, \mu)$.

Sada imamo sve potrebne pojmove da definiramo slučajnu varijablu (vidi [4] i [8]).

Definicija 2.1.5. *Neka je $(\Omega, \mathbb{F})$ bilo koji izmjeriv prostor i neka je $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ izmjeriv prostor sa σ-algebrom Borelovih skupova. Kažemo da je $X : \Omega \to \mathbb{R}^d$ d-dimenzionalna slučajna veličina ako je X izmjerivo preslikavanje u paru σ-algebri $(\mathcal{F}, \mathcal{B}(\mathbb{R}^d))$, tj. ako vrijedi*

$$(\forall B \in \mathcal{B}(\mathbb{R}^d)) \quad \{X \in B\} \in \mathcal{F}. \tag{2.4}$$

Ako je $d = 1$, X zovemo slučajna varijabla, a za $d \geq 2$, X zovemo slučajan vektor.

Definirajmo distribuciju slučajne veličine X.

Definicija 2.1.6. *Neka je $X : \Omega \to \mathbb{R}^d$ d-dimenzionalna slučajna veličina definirana na vjerojatnosnom prostoru $(X, \mathcal{F}, \mu)$. Induciranu vjerojatnost $\mathbb{P}_X$, definiranu na izmjerivom prostoru $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ relacijom*

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}^d), \tag{2.5}$$

zovemo zakon razdiobe od X.

Definicija 2.1.7. *Neka je X d-dimenzionalna slučajna veličina sa zakonom razdiobe $\mathbb{P}_X$. Funkciju $F_X : \mathbb{R}^d \to \mathbb{R}$ definiranu sa*

$$F_X(x) := \mathbb{P}_X(\langle -\infty, x]), \quad x \in \mathbb{R}^d, \tag{2.6}$$

zovemo funkcija razdiobe (ili distribucije) od X.

Neka je $\lambda^d = \lambda^1 \times \cdots \times \lambda^1$ Lebesqueova mjera definirana na izmjerivom prostoru $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (vidi [3]). Definirajmo dvije vrste slučajnih veličina:

Definicija 2.1.8. *Kažemo da je d-dimenzionalna slučajna veličina X apsolutno neprekidna ili neprekidna ako postoji nenegativna Borelova funkcija f_X definirana na $\mathbb{R}^d$ takva da se funkcija razdiobe F_X može prikazati na sljedeći način:*

$$F_X(x) = \int\limits_{\langle -\infty, x]} f_X(y) d\lambda(y), \quad x \in \mathbb{R}^d. \tag{2.7}$$

Funkciju f_X zovemo funkcija gustoće razdiobe od X ili gustoća od X.

Definicija 2.1.9. *Slučajna veličina X dimenzije d je diskretna ako postoji prebrojiv skup $D \subseteq \mathbb{R}^d$ takav da je $\mathbb{P}_X(D) = 1$.*

Primjer 2.1.10. *(a) Slučajna varijabla X ima normalnu distribuciju s parametrima μ i σ^2, $\sigma > 0$, ako joj je funkcija gustoće*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Pišemo $X \sim N(\mu, \sigma^2)$.

(b) Slučajna varijabla X ima (Studentovu) t-distribuciju s n stupnjeva slobode, ako joj je funkcija gustoće

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}}$$

Pišemo $X \sim t(n)$.

(c) Slučajna varijabla X ima binomnu distribuciju s parametrima $n \in \mathbb{N}$ i $0 < p < 1$, ako joj je funkcija gustoće

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, 1, \ldots, n\}.$$

Pišemo $X \sim B(n, p)$.

(d) Slučajna varijabla X ima Poissonovu distribuciju s parametrom λ, ako joj je funkcija gustoće

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \mathbb{N}_0.$$

Pišemo $X \sim P(\lambda)$.

Definiramo korelaciju u teoriji vjerojatnosti.

Definicija 2.1.11. *Neka je X slučajna varijabla definirana na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$. *Definiramo* $X^+ := max\{0, X\}$, $X^- := max\{0, -X\}$, $\mathbb{E}X^+ := \int\limits_\Omega X^+ d\mathbb{P}$ *i* $\mathbb{E}X^- := \int\limits_\Omega X^- d\mathbb{P}$. *Kažemo da slučajna varijabla X ima matematičko očekivanje ako je barem jedan od integrala* $\mathbb{E}X^+$ *i* $\mathbb{E}X^-$ *konačan. U tom slučaju je matematičko očekivanje*

$$\mathbb{E}X := \mathbb{E}X^+ - \mathbb{E}X^-. \tag{2.8}$$

Matematičko očekivanje je konačno ako je $\mathbb{E}|X| = \mathbb{E}X^+ + \mathbb{E}X^- < +\infty$.

Definicija 2.1.12. *Neka su X i Y slučajne varijable takve da je* $\mathbb{E}(X^2) < +\infty$ *i* $\mathbb{E}(Y^2) < +\infty$. *Označimo* $\mu_X = \mathbb{E}X$ *i* $\mu_Y = \mathbb{E}Y$.

(a) Varijanca od X je nenegativni broj

$$VarX := \mathbb{E}[(X - \mu_X)^2].$$

(b) Standardna devijacija od X je nenegativan broj

$$std(X) := \sqrt{VarX}.$$

(c) Kovarijanca od X i Y je broj

$$cov(X, Y) := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

(d) Ako su $\sigma_X = std(X) > 0$ *i* $\sigma_Y = std(Y) > 0$, *tada su dobro definirane standardizirane varijable od X i Y izrazima:*

$$Z_X := \frac{X - \mu_X}{\sigma_X}, \quad Z_Y := \frac{Y - \mu_Y}{\sigma_Y}.$$

U tom slučaju je koeficijent korelacije od X i Y broj

$$corr(X, Y) := cov(Z_X.Z_Y).$$

Na kraju definiramo statističke testove i hipoteze.

Definicija 2.1.13. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i $\mathcal{P}$ množina vjerojatnosnih mjera definiranih na $(\Omega, \mathcal{F})$. Tada uređenu trojku $(\Omega, \mathcal{F}, \mathcal{P})$ zovemo statistička struktura. Množina $\mathcal{P}$ je često parametrizirana konačnodimenzijskim parametrom θ:*

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

Θ je podskup od $\mathbb{R}^d$ ($d \geq 1$) koji zovemo parametarski prostor. Parametarski model je

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

gdje su gustoće $f(\cdot; \theta)$ slučajnih veličina $X_1, X_2, \ldots, X_n$ dimenzije k ($k \geq 1$) parametrizirane parametrom θ dimenzije d ($d \geq 1$).

Definicija 2.1.14. *Statistika na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je svaka slučajna veličina koja je izmjeriva funkcija nekog slučajnog uzorka na toj statističkoj strukturi.*

Neka je $X = (X_1, X_2, \ldots, X_n)$ slučajni uzorak duljine n ($n \in \mathbb{N}$) iz parametarskog modela $\mathcal{P} = \{f(\cdot; \theta) : \theta \in \Theta\}$ zadanog vjerojatnosnim gustoćama i neka je $\{\Theta_0, \Theta_1\}$ jedna particija parametarskog prostora Θ. To znači da su $\Theta_0, \Theta_1 \subset \Theta$ takvi podskupovi da su disjunktni (tj. $\Theta_0 \cap \Theta_1 = \emptyset$) i da je $\Theta_0 \cup \Theta_1 = \Theta$. Pretpostavimo da želimo testirati statističke hipoteze

$$\begin{aligned} H_0 &: \quad \theta \in \Theta_0 \\ H_1 &: \quad \theta \in \Theta_1 \end{aligned}$$

Hipotezu H_0 zovemo *nultom* hipotezom, a H_1 *alternativnom* hipotezom.

Definicija 2.1.15. *Randomizirajući test za testiranje statističkih hipoteza H_0 u odnosu na H_1 je izmjerivo preslikavanje $\tau : \mathbb{R}^{kn} \to [0, 1]$. Testna statistika je statistika korištena u testu. p-vrijednost je vjerojatnost događaja da će testna statistika poprimiti vrijednosti za koje je vjerodostojnost osnovne hipoteze u odnosu na alternativnu hipotezu manja od opažene vrijednosti te statistike ili joj je jednaka, uz uvjet da je osnovna hipoteza istinita.*

2.1.2 Opisna statistika i testiranje statističkih hipoteza

U statistici, korelacija je ovisnost između slučajnih varijabli. Jakost linearne korelacije (linearne ovisnosti) izražava se koeficijentom ρ za koji vrijedi $-1 \leq \rho \leq 1$. ρ zovemo koeficijent korelacije slučajnih varijabli X i Y. Više o korelaciji između dvije slučajne varijable može se naći na [4] i [5].

Definicija 2.1.16. *Procjenitelj od τ je procjena parametra τ. Procjenitelj $T_n = f_n(X_1, \ldots, X_n)$ je nepristrani procjenitelj za parametar τ ako vrijedi $\mathbb{E}T_n = \tau$.*

Postoje dvije vrste koeficijenata korelacije: koeficijent korelacije uzoraka r i koeficijent korelacije populacija ρ. Koeficijent korelacije uzorka je procjena (procjenitelj) koeficijenta korelacije populacije. Navedimo dva koeficijenta korelacije koji se često koriste:

- *Pearsonov koeficijent korelacije* je broj

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\left(\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2\right)}},$$

  gdje je n duljina uzoraka, a $\bar{x}$ i $\bar{y}$ su aritmetičke sredine uzoraka $X = (x_1, \ldots, x_n)$ i $Y = (y_1, \ldots, y_n)$.

- *Spearmanov koeficijent korelacije* je broj

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)},$$

  gdje je n duljina uzoraka, a $d_i = R(x_i) - R(y_i)$ je razlika između dva ranga uzoraka X i Y (posložimo vrijednosti uzorka po veličini od najvećeg prema najmanjem pa najveća vrijednost ima rang 1, druga rang 2 i tako dalje).

Napomena 2.1.17. *Pearsonov koeficijent korelacije je najpoznatiji koeficijent korelacije uzorka i najčešće se koristi. Ako su uzorci X i Y normalno distribuirani, tada je Pearsonov koeficijent korelacije nepristrani procjenitelj koeficijenta korelacije populacije. Kada uzorci X i Y nisu normalno distribuirani, preporučuje se koristiti Spearmanov koeficijent korelacije.*

Da bismo dokazali ili opovrgnuli tvrdnju da je korelacija između dvije varijable X i Y značajna, koristimo *korelacijski test*. Neka su x i y uzorci od X i Y. Hipoteze korelacijskog testa su

$H_0: \quad \rho = 0$
$H_0: \quad \rho \neq 0,$

a testna statistika T korelacijskog testa je

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \sim \quad t(n-2), \tag{2.9}$$

gdje je n duljina uzoraka x i y, r je koeficijent korelacije uzoraka x i y, a $t(n-2)$ označava Studentovu t-distribuciju s $n-2$ stupnjeva slobode. Koeficijent korelacije r može biti Pearsonov koeficijent korelacije, Spearmanov koeficijent korelacije i ostali.

2.1.3 Linearna regresija

Nakon što se pokaže da postoji linearna veza između dva skupa podataka, prirodno je htjeti pronaći model koji aproksimira vrijednosti jednog skupa podataka preko vrijednosti njemu koreliranog skupa podataka. Kako je veza (korelacija) između dva skupa linearna, tada će i model biti linearan. Pravac je jedna vrsta jednostavnog linearnog modela koji aproksimira vrijednosti jednog skupa podataka uzimajući vrijednosti drugog skupa podataka uz uvjet da su skupovi podataka korelirani. Jedna vrsta takvih modela zove se model linearne regresije i često se koristi u statistici. Sada ćemo objasniti što predstavljaju linearni modeli u statistici (vidi [6]).

Neka je

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \epsilon \tag{2.10}$$

jednodimenzionalni linearni model, gdje su $x_1, x_2, \ldots, x_p$ varijable poticaja, ϵ slučajna greška, Y varijabla odaziva i $\beta_0, \beta_1, \ldots, \beta_p$ parametri modela. Kada imamo više opažanja, (2.10) zapisujemo kao

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

gdje pretpostavljamo da su greške $\epsilon_1, \ldots, \epsilon_n$ nezavisne s distribucijom $N(0, \sigma^2)$. U matričnom obliku, to je

$$Y = Xb + \epsilon,$$

gdje je $Y = (Y_1, \ldots, Y_n))^\tau$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n))^\tau \sim N(0, \sigma^2 \mathbb{1})$, $b = (\beta_0, \beta_1, \ldots, \beta_n)$ i

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Želimo li minimizirati $\|\epsilon\|_2 = \|Y - Xb\|_2$ po b, dobivamo da je najbolja procjena za b

$$\hat{b} = (X^\tau X)^{-1} X^\tau Y,$$

uz uvjet da je $X^\tau X$ regularna. Tada su procijenjene vrijednosti

$$\hat{Y} = X\hat{b} = X(X^\tau X)^{-1} X^\tau Y,$$

i ostaci

$$e = Y - \hat{Y} = (\mathbb{1} - H)Y, \quad H := X(X^\tau X)^{-1}X^\tau.$$

Linearni model zadovoljava sljedeća svojstva:

- $\hat{b} \sim N(b, (X^\tau X)^{-1}\sigma^2)$
- $\frac{\hat{b}_i - b_i}{\hat{\sigma}\sqrt{(X\tau X)^{-1}_{ii}}} \sim t(n-p-1)$
- $e \sim N(0, (\mathbb{1} - H)\sigma^2)$
- $\sum_{i=1}^{n} e_i = 0$
- $\hat{\sigma^2} = \frac{e^\tau e}{n-p-1}$ je nepristrani procjenitelj za σ^2 i vrijedi $\frac{(n-p-1)\hat{\sigma^2}}{\sigma^2} \sim \chi^2(n-p-1)$

Možemo procijeniti koliko je linearni model dobar. Neki od načina su:

- $R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$, R^2 zovemo koeficijent determinacije
- $\hat{\sigma^2} = \frac{SSR}{n-p-1}$
- ${R_a}^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2/(n-p-1)}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2/(n-1)}$, ${R_a}^2$ zovemo prilagođeni (adjusted) R^2
- $F = \frac{\frac{SSR_0 - SSR}{p-p_0}}{\frac{SSR}{n-p}} \sim F(p-p_0, n-p)$ test značajnosti modela

Kada koristimo linearni regresijski model, moramo provjeriti zadovoljava li on određene uvjete. Sada navodimo četiri glavne pretpostavke koje bi trebale biti zadovoljene te kako ih možemo ispitati:

- *linearni odnos* između varijabli poticaja (nezavisne varijable) i odaziva (zavisna varijabla)
  - korištenjem $Y-\hat{Y}$ grafa i $\hat{Y}-e$ grafa (residual plot) možemo provjeriti linearnost
  - u $Y-\hat{Y}$ grafu očekujemo simetrično raspršenje podataka oko dijagonale, a u $\hat{Y}-e$ grafu simetrično raspršenje oko apcise
  - nelinearnost možemo ukloniti transformacijom varijable koristeći neku nelinearnu funkciju
- *nezavisnost* grešaka
  - korištenjem `acf`() funkcije u `R`-u možemo provjeriti nezavisnost
  - na grafu očekujemo da je oko 95% stupića unutar 95% pouzdane pruge prikazane na njemu
  - nezavisnost možemo riješiti statistički boljim sakupljanjem podataka ili korištenjem pravilnog ARIMA modela za greške
- *homogenost* grešaka
  - korištenjem $t-e$ grafa (reziduali s obzirom na vrijeme) i $\hat{Y}-e$ grafa (residual plot) možemo provjeriti linearnost
  - u oba grafa očekujemo simetrično raspršenje oko apcise
  - homogenost možemo ukloniti analizom po dijelovima podataka s jednakom varijancom, ispitivanjem pravilnog odabira modela ili korištenjem ARCH modela
- *normalnost* grešaka
  - normalnost možemo provjeriti statističkim testiranjem residuala za ispitivanje normalnosti podataka (Lillieforsova inačica Kolmogorov-Smirnovljevog testa, Shapiro-Wilk test i ostali)

Kako je navedeno iznad, kada su greške linearnog modela zavisne, jedan od načina da rješimo taj problem je modeliranjem grešaka ARIMA modelima.

Definicija 2.1.18. *Neka je S skup. Slučajan proces s diskretnim vremenom i prostorom stanja S je familija $X = (X_n : n \geq 0)$ slučajnih varijabli (ili elemenata) definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u S. Dakle, za svaki $n \geq 0$, $X_n : \Omega \to S$ je slučajna varijabla.*

Definicija 2.1.19. *Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ zove se stacionaran ako za sve $k \geq 0$ i sve $n \geq 0$, slučajni vektori $(X_0, X_1, \ldots, X_k)$ i $(X_n, X_{n+1}, \ldots, X_{n+k})$ imaju istu distribuciju (u odnosu na vjerojatnost $\mathbb{P}$).*

Definicija 2.1.20. *Za niz $(X_t)_{t\in\mathbb{Z}}$ kažemo da je bijeli šum ako je nekoreliran s očekivanjem 0 i konstantnom varijancom σ^2. Pišemo $(X_t) \sim WN(0, \sigma^2)$.*

Definicija 2.1.21. *Slučajni proces $\{X_t : t \in \mathbb{Z}\}$ zove se $ARMA(p, q)$ proces ako je (X_t) stacionaran i za svaki t vrijedi*

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{2.11}$$

gdje je $(Zt) \sim WN(0, \sigma^2)$. Ako je $\phi_1 = \phi_2 = \cdots = \phi_p = 0$, proces zovemo $MA(q)$, a ako je $\theta_1 = \theta_2 = \cdots = \theta_q = 0$ proces zovemo $AR(p)$.

Definicija 2.1.22. *Kažemo da je proces $ARIMA(p, d, q)$ proces, ako ga trebamo diferencirati d puta da bismo dobili $ARMA(p, q)$ proces (postoji d jediničnih korijena). Proces će biti stacionaran $\Leftrightarrow d = 0$.*

Model linearne regresije s ARIMA greškama oblika je

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \eta_i, \quad i = 1, 2, \ldots, n, \tag{2.12}$$

gdje je η_t $ARIMA(p, d, q)$ greška.

U slučaju kada se uspoređuje više ARIMA modela za najbolji odabir aproksimacije grešaka, često se koriste AIC i BIC kriteriji te ANOVA test.

Kada su sve pretpostavke dobrog modela zadovoljene, korisno je provjeriti i usporediti aproksimacije modela sa stvarnim vrijednostima. Također, modeli linearne regresije često se koriste za predviđanja. Više o modelima linearne regresije može se naći na [1] i [6].

2.2 Modeliranje diskretiziranih podataka

Jedan od razloga zbog kojih diskretiziramo podatke je da ih transformiramo u modele koji služe za predviđanje i nova otkrića. Jedan od poznatih modela koji se koriste u bioinformatici, strojnom učenju, u prepoznavanju govora i ostalima je skriven Markovljev model (*engl.* hidden Markov model). Da bismo objasnili teoriju iza skrivenog Markovljevog modela, prvo moramo definirati Markovljeve lance i sva njihova svojstva koja će nam trebati za implementaciju HMM-a. Više o Markovljevim lancima može se naći na [9].

2.2.1 Markovljevi lanci

Definicija 2.2.1. *Neka je S skup. Slučajan proces s diskretnim vremenom i prostorom stanja S je familija $X = (X_n : n \geq 0)$ slučajnih varijabli (ili elemenata) definiranih na nekom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u S. Dakle, za svaki $n \geq 0$, $X_n : \Omega \to S$ je slučajna varijabla.*

Definicija 2.2.2. *Neka je S prebrojiv skup. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s vrijednostima u skupu S je Markovljev lanac ako vrijedi*

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i) \tag{2.13}$$

za svaki $n \geq 0$ i za sve $i_0, \ldots, i_{n-1}, i, j \in S$ za koje su obje uvjetne vjerojatnosti dobro definirane.

Svojstvo u relaciji (2.13) naziva se *Markovljevim svojstvom*.

Definicija 2.2.3. *Matrica $P = (p_{ij} : i, j \in S)$ naziva se stohastičkom matricom ako je $p_{ij} \geq 0$ za sve $i, j \in S$ te*

$$\sum_{j \in S} p_{ij} = 1, \quad \text{za sve } i \in S. \tag{2.14}$$

Ukoliko je broj stanja u S konačan, tada je P "prava" (konačna) matrica. S druge strane, ako je S beskonačan skup, tada će P biti beskonačna matrica.

U ovom radu zanimaju nas *homogeni* Markovljevi lanci.

Definicija 2.2.4. *Neka je $\lambda = (\lambda_i : i \in S)$ vjerojatnosna distribucija na S, te neka je $P = (p_{ij} : i, j \in S)$ stohastička matrica. Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ s prostorom stanja S je homogen Markovljev lanac s početnom distribucijom λ i prijelaznom matricom P ako vrijedi*

(i) $\mathbb{P}(X_0 = i) = \lambda_i$, *za sve $i \in S$, te*

(ii) $\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = p_{ij}$, *za svaki $n \geq 0$ i za sve $i_0, \ldots, i_{n-1}, i, j \in S$.*

Ponekad ćemo Markovljev lanac iz definicije 2.2.4 nazivati (λ, P)-Markovljevim lancom.

Sada definiramo posebnu vrstu Markovljevih lanaca.

Definicija 2.2.5. *Slučajni proces $X = (X_n : n \geq 0)$ definiran na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ zove se stacionaran ako za sve $k \geq 0$ i sve $n \geq 0$, slučajni vektori $(X_0, X_1, \ldots, X_k)$ i $(X_n, X_{n+1}, \ldots, X_{n+k})$ imaju istu distribuciju (u odnosu na vjerojatnost $\mathbb{P}$).*

Definicija 2.2.6. *Neka je $X = (X_n : n \geq 0)$ Markovljev lanac s prebrojivim skupom stanja S i prijelaznom matricom P. Vjerojatnosna distribucija $\pi = (\pi_i : i \in S)$ na S je stacionarna distribucija (ili invarijantna distribucija) Markovljevog lanca X (odnosno prijelazne matrice P) ako vrijedi*

$$\pi = \pi P, \tag{2.15}$$

odnosno po komponentama

$$\pi_j = \sum_{k \in S} \pi_k p_{kj}, \quad \text{za sve } j \in S. \tag{2.16}$$

Teorem 2.2.7. *Neka je $X = (X_n : n \geq 0)$ (π, P)-Markovljev lanac gdje je π stacionarna distribucija za P. Tada je X stacionaran proces. Preciznije, X je stacionaran uz vjerojatnost $\mathbb{P}_\pi = \sum_{i \in S} \pi_i \mathbb{P}_i$. Nadalje, za svaki $m \geq 0$ je $(X_{m+n} : n \geq 0)$ ponovno (π, P)-Markovljev lanac.*

Dokaz. Uočimo prvo da iz $\pi = \pi P$ slijedi $\pi = (\pi P)P = \pi P^2$, te indukcijom $\pi = \pi P^n$ za sve $n \geq 1$. Po komponentama,

$$\pi_j = \sum_{k \in S} \pi_k {p_{kj}}^n. \tag{2.17}$$

Neka je sada $k \geq 0$, $n \geq 0$, te neka su $i_0, i_1, \ldots, i_k \in S$. Vrijedi:

$$\mathbb{P}_\pi(X_n = i_0, X_{n+1} = i_1, \ldots, X_{n+k} = i_k) = \sum_{i \in S} \pi_i {p_{ii_0}}^{(n)} p_{i_0 i_1} \ldots p_{i_{k-1} i_k}$$

$$= \pi_{i_0} p_{i_0 i_1} \ldots p_{i_{k-1} i_k} = \mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_k = i_k),$$

gdje drugi redak slijedi zbog (2.17).

Nadalje, otprije znamo da je $(X_{m+n} : n \geq 0)$ Markovljev lanac s prijelaznom matricom P. Početna distribucija tog lanca jednaka je $(\mathbb{P}_\pi(X_m = i) : i \in S)$. Budući da je $\mathbb{P}_\pi(X_m = i) = \pi_i$, tvrdnja slijedi. □

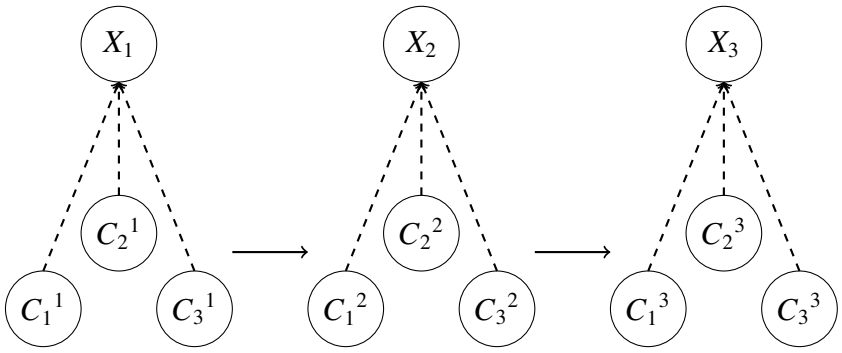
2.2.2 Skriveni Markovljev model (HMM)

Definicija 2.2.8. *Skriveni Markovljev model (engl. hidden Markov model (**HMM**)) $\{X_t : t \in \mathbb{N}\}$ je posebna vrsta zavisne strukture. Neka $\boldsymbol{X}^{(t)}$ i $\boldsymbol{C}^{(t)}$ predstavljaju povijest od vremena 1 do vremena t. Najjednostavniji skriveni Markovljev model zadovoljava svojstva*

$$\mathbb{P}(C_t | \boldsymbol{C}^{(t-1)}) = \mathbb{P}(C_t | C_{t-1}), \quad t = 2, 3, \ldots \tag{2.18}$$

$$\mathbb{P}(C_t | \boldsymbol{X}^{(t-1)}, \boldsymbol{C}^{(t)}) = \mathbb{P}(X_t | C_t), \quad t \in \mathbb{N}. \tag{2.19}$$

Model se sastoji od dva dijela: prvi je *parametarski proces* $\{C_t : t = 1, 2, \dots\}$ koji zadovoljava Markovljev uvjet (2.13), a drugi je *proces opservacija* - proces koji ovisi o stanjima $\{X_t : t = 1, 2, \dots\}$, takav da kad je C_t poznat, distribucija od X_t ovisi samo o trenutnom stanju C_t, a ne o prethodnim stanjima ili opservacijama. Ako Markovljev lanac $\{C_t\}$ ima m stanja, tada $\{X_t\}$ zovemo skriveni Markovljev model s m stanja. Više na [10].



Slika 2.1: Shematski prikaz jednostavnog HMM-a.

Sljedeći parametri opisuju tipičnu implementaciju HMM-a (vidi [7]):

- Skup stanja $S = \{S_1, \dots, S_N\}$ takav da je q_t stanje u vremenu t,
- Skup vjerojatnosti $B = \{b_1(o), \dots, b_N(o)\}$, gdje su $b_j(o_t) = \mathbb{P}(o_t | q_t = S_j)$, za $1 \leq j \leq N$, emisijske vjerojatnosti stanja, a o_t je opservacija u vremenu t iz niza opservacija $O = \{o_1, \dots, o_T\}$ ($b_j(o_t)$ je vjerojatnost da stanje j emitira vrijednost opservacije o_t),
- Prijelazna matrica stanja $A = \{a_{ij}\}$, za $1 \leq i, j \leq N$, gdje je $a_{ij} = \mathbb{P}(q_{t+1} = S_j | q_t = S_i)$,
- Vektor početne distribucije stanja $\Pi = \{\pi_1, \dots, \pi_N\}$.

Viterbijev algoritam

U HMM-u poznat nam je niz opservacija, no ne znamo koji je točno niz stanja emitirao te opservacije, ali možemo saznati koji ih je niz stanja najvjerojatnije emitirao. Da bismo to saznali, koristimo Viterbijev algoritam koji je ujedno jedan od poznatijih algoritama koji se koriste u HMM-u. Na kraju navodimo korake algoritma (vidi [7]):

- Na početku, $\delta_1(i) = \pi_i b_i(o_1)$, $\psi_1(i) = 0$, za $1 \leq i \leq N$.

- $\delta_t(j) = \max\limits_{1 \leq i \leq N}[\delta_{t-1} a_{ij}] b_j(o_t)$, $\Psi_t(j) = \arg\max\limits_{1 \leq i \leq N}[\delta_{t-1}(i) a_{ij}]$, za $t = 2, \ldots, T$ i $1 \leq j \leq N$
- Konačno, $q_T{}^* = \arg\max\limits_{1 \leq i \leq N}[\delta_T(i)]$, $q_t{}^* = \Psi_{t+1}(q_{t+1}{}^*)$, za $t = T-1, \ldots, 1$, s optimalnim putem $Q^* = \{q_1{}^*, \ldots, q_T{}^*\}$.

Poglavlje 3

Opis podataka

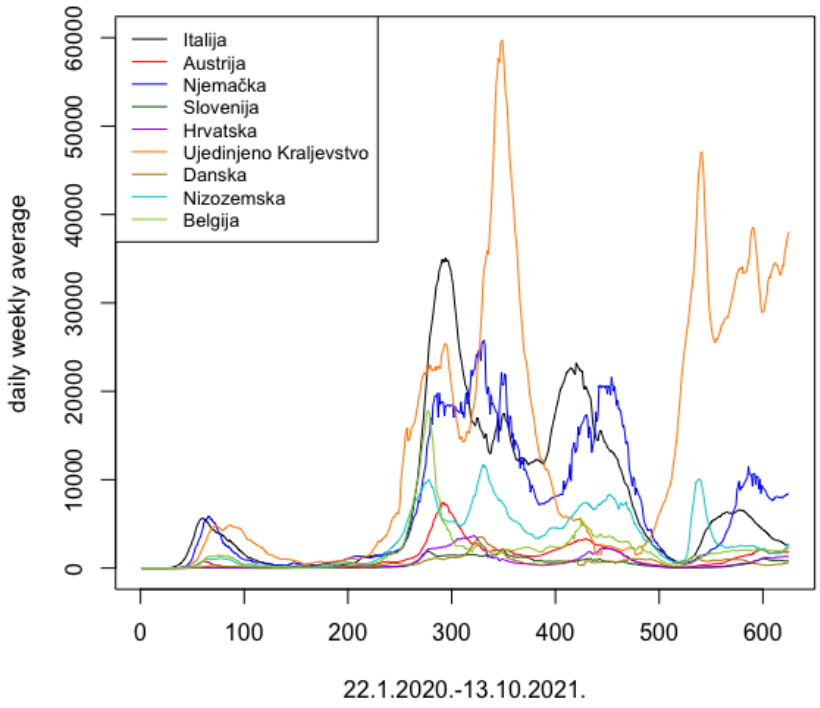
Postoji dosta javnih informacija o koronavirusu i svemu vezano uz tu temu. Podaci koji su nam dostupni i koje smo odlučili koristiti su o dnevnim kumulativnim brojevima slučajeva svake države koji se nalaze na GitHub repozitoriju Johns Hopkins University - Center for Systems Science and Engineering [2]. Naravno, ovo su samo jedni od javno dostupnih podataka te smo mogli izabrati i druge kao što su smrtnost od koronavirusa, hospitalizacija i ostali.

Kako znamo da su u današnjem svijetu mnoge zemlje povezane, a pogotovo države koje se nalaze u Europskoj uniji, za obrađivanje smo odlučili preuzeti podatke za devet država: Italiju, Austriju, Njemačku, Sloveniju, Hrvatsku, Ujedinjeno Kraljevstvo, Dansku, Belgiju i Nizozemsku. Podaci su dani za svaki dan u periodu od 22.1.2020. do 13.10.2021.

Za potrebe daljnjih analiza, najprije smo transformirali kumulativne podatke u dnevne brojeve novih slučajeva, a da bismo dobili željene podatke, bilo je potrebno provesti transformacije dnevnih brojeva novih slučajeva.

3.1 Podaci o dnevnom tjednom prosjeku zaraženih

Jasno je da dan u tjednu ovisi o broju novih dnevnih slučajeva - nedjeljom i ponedjeljkom ima manje pozitivnih na COVID-19 u odnosu na ostale dane u tjednu jer se manje ljudi testira vikendom. Zbog toga, službeni broj pozitivnih u nedjelju i ponedjeljak nikad nije pravi broj. Kako u ovom radu pratimo tok zaraze (rast ili pad), ne želimo analizirati krive brojke. Da to izbjegnemo, transformirali smo podatke o dnevnom broju novih slučajeva u novu varijablu dnevni tjedni prosjek zaraženih (daily weekly average) na sljedeći način: prvo smo uzeli prvu vrijednost (dan) i sljedećih šest te ih uprosječili pa smo uzeli drugu vrijednost i sljedećih šest te ih uprosječili i tako dalje.



Slika 3.1: Vremenski niz dnevnog tjednog prosjeka zaraženih koronavirusom

Iz grafičkog prikaza 3.1 možemo primijetiti da podaci zaraženih za sve države prate skoro isti trend, ali se razlikuju u vremenu kada počinje uzlazni i silazni trend te u visini vrhunca svakog vala. Također vidimo da su te razlike u vremenu od par tjedana (npr. razlika između Italije i Ujedinjenog Kraljevstva od početka drugog i četvrtog vala je pedeset dana, tj. sedam tjedana).

3.2 Diskretizacija podataka - stopa zaraze

Svaki pojedinac ima može razviti neke od mogućih simptoma koronavirusa. Netko nije svjestan da je pozitivan na COVID-19 jer nema nikakve simptome, dok netko drugi pretpostavlja da se zarazio, ali se odbija testirati. Dolaskom delta soja, nove varijante koronavirusa, zaraza je brža i veća, ali cijepljeni uglavnom imaju simptome obične gripe pa nisu sigurni jesu li se zarazili. Ponekad se ni ne registrira kada netko bude pozitivan nakon

što se testira kod kuće. Sve to su razlozi zbog kojih službene brojke o dnevnim novim slučajevima nekad nisu točne. Štoviše, često je pravi broj puno veći.

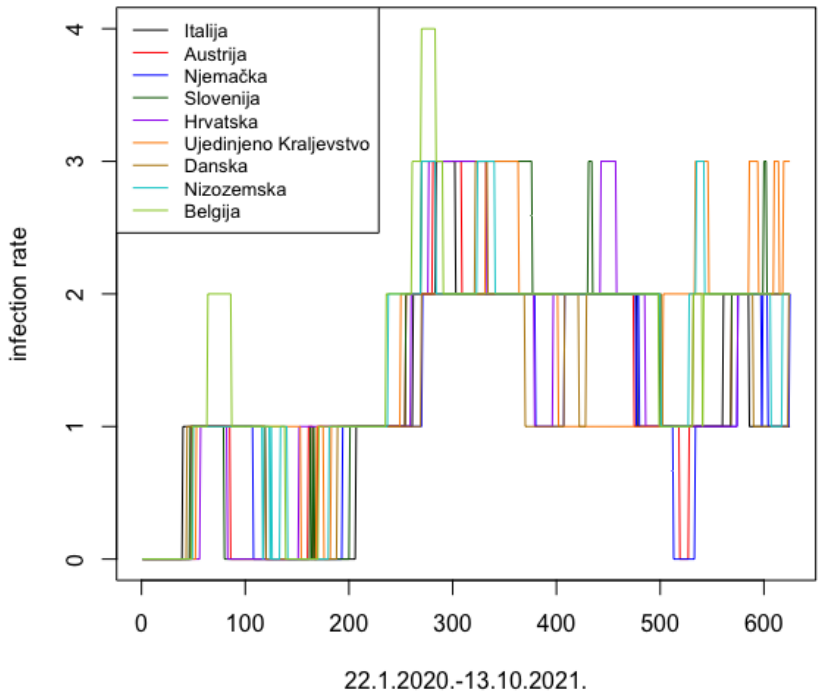
Da bismo "popravili" krive podatke, varijablu dnevnih tjednih prosjeka zaraženih transformirali smo u novu diskretnu varijablu stope zaraze (infection rate). Varijablu smo diskretizirali tako da smo prosjeke stavili u kategorije pa nam time pravi broj zaraženih nije trebao. Vrijednosti nove varijable stope zaraze jedne su od sljedećih:

- 0 : daily weekly average / populacija države $\leq 10/1000000$
- 1 : $10/1000000 <$ daily weekly average / populacija države $\leq 50/1000000$
- 2 : $50/1000000 <$ daily weekly average / populacija države $\leq 100/1000000$
- 3 : $100/1000000 <$ daily weekly average / populacija države $\leq 500/1000000$
- 4 : daily weekly average / populacija države $> 500/1000000$

Broj stanovnika[1] svake države je sljedeći:

	broj stanovnika
Italija	60 342 758
Austrija	9 076 194
Njemačka	84 146 021
Slovenija	2 079 322
Hrvatska	4 071 669
Ujedinjeno Kraljevstvo	68 369 926
Danska	5 819 787
Nizozemska	17 186 015
Belgija	11 657 948

[1] podaci preuzeti s `https://www.worldometers.info/population/europe/`



Slika 3.2: Vremenski niz stope zaraze koronavirusom

Na slici 3.2 možemo uočiti da je tok stope zaraze na početku pandemije koronavirusa gotovo isti za sve navedene države s razlikom u vremenu od par tjedana, dok ipak možemo primijetiti razlike u toku pandemije za različite države pri kraju obrađenih podataka. To nam daje naslutiti kako je bilo povezanosti između država s obzirom na stopu zaraze, no da su se dogodile promjene nakon godinu dana.

Grafičkim prikazom dviju varijabli vidjeli smo da je tok zaraženih koronavirusom za sve države međusobno vrlo sličan, ali da bi moglo biti razlike između država godinu i pol dana nakon početka pandemije. Zbog toga naslućujemo da bi ovisnost količine zaraženih između zemalja zaista mogla postojati, barem u početku pandemije.

Poglavlje 4

Primjena teorije na podatke

Članice Europske unije podosta su povezane zbog današnje mobilnosti te poslovnog i privatnog života svojih stanovnika. Zbog toga mislimo da bi i tok zaraze jedne države mogao utjecati na tok zaraze druge države u Europskoj uniji. Između ostalog, u poglavlju 3 grafički prikazi podataka dali su nam naslutiti da bi linearne ovisnosti broja slučajeva zaraze virusom i stopa zaraze između navedenih zemalja mogle postojati. Da bismo dokazali ili opovrgnuli te tvrdnje, koristit ćemo definicije i teoreme iz poglavlja 2.

Dodatno, modelirati ćemo diskretiziranu varijablu stope zaraze u skriven Markovljev model i pokazati kako je to samo još jedan uspješan način kako možemo analizirati različite vrste podataka vezane uz pandemiju koronavirusa.

4.1 Korelacije

Na početku zabilježenih podataka o zarazi koronavirusom u grafičkom prikazu u poglavlju 3 primijetili smo da je tok gotovo isti za sve analizirane države. Međutim, oko petstotog dana mogli smo vidjeti razlike u tokovima različitih zemalja. Vrijednosti za taj period predstavljaju podatke za lipanj 2021. godine, a to je vrijeme kada su pojedinci koji su se odlučili cijepiti većinom dobili drugu dozu cjepiva. Kako nas zanima utječe li povezanost zemalja na tok zaraze i kako smo uočili razlike u odnosu na stanje na početku pandemije i stanje u novije vrijeme, podatke ćemo podijeliti na dva dijela i posebno analizirati podatke za svaki period. Prvo dio će sadržati vrijednosti za period od 1.5.2020. do 30.6.2021., a drugi za period od 1.7.2021. do 13.10.2021. Podatke prije 1.5.2020. nećemo analizirati jer ne predstavljaju reprezentativan uzorak za taj period pandemije.

Prije nego što izračunamo i testiramo koeficijente korelacije, moramo testirati imaju li varijable normalnu distribuciju da bismo znali koje koeficijente korelacije možemo koristiti. Kako je stopa zaraze diskretna varijabla, ona očito nije normalno distribuirana pa ćemo

samo testirati varijablu dnevnog tjednog prosjeka, ali posebno za svaki period. Postavljamo hipoteze:

H_0: podaci su normalno distribuirani
H_1: ne H_0

Koristeći programski jezik R dobili smo jako male p-vrijednosti za svaku državu pa odbacujemo hipotezu H_0 na svim razumnim razinama značajnosti. Dakle, pretpostavljamo da podaci nisu normalno distribuirani ni za jednu državu.

4.1.1 Korelacijski koeficijenti i testovi

Podaci za period 1.5.2020. - 30.6.2021.

Prethodno smo testirali jesu li varijable dnevnog tjednog prosjeka zaraze normalno distribuirane. Kako smo odbacili hipotezu o normalnosti za svaku državu, koristimo Spearmanov koeficijent korelacije definiran u 2.1.2 za procjenitelja koeficijenta korelacije populacije. Dobivamo sljedeće rezultate:

	Italija	Austrija	Njemačka	Slovenija	Hrvatska
Italija	1	0.961944	0.897087	0.906498	0.89087
Austrija	0.9619442	1	0.913416	0.920212	0.945113
Njemačka	0.897087	0.913416	1	0.911793	0.949522
Slovenija	0.906498	0.920212	0.911793	1	0.911468
Hrvatska	0.890871	0.945113	0.949523	0.911468	1
Ujedinjeno Kraljevstvo	0.740426	0.658401	0.682142	0.806388	0.606273
Danska	0.836223	0.807321	0.902447	0.868901	0.85912
Nizozemska	0.852443	0.865148	0.924945	0.868436	0.900772
Belgija	0.861008	0.881635	0.812776	0.776790	0.84901

	Ujedinjeno Kraljevstvo	Danska	Nizozemska	Belgija
Italija	0.740426	0.836223	0.852443	0.861008
Austrija	0.658401	0.807321	0.865148	0.881635
Njemačka	0.682142	0.902447	0.924945	0.812776
Slovenija	0.806388	0.868901	0.868436	0.776790
Hrvatska	0.606273	0.859115	0.900772	0.849014
Ujedinjeno Kraljevstvo	1	0.732496	0.693446	0.587262
Danska	0.732496	1	0.847694	0.705556
Nizozemska	0.693446	0.847694	1	0.878720
Belgija	0.587262	0.705556	0.878720	1

Tablica 4.1: Spearmanovi koeficijenti korelacije za dnevni tjedni prosjek zaraze

Primijetimo da su Spearmanovi koeficijenti korelacije vrlo visoki za sve testirane države osim Ujedinjenog Kraljevstva koji ima nešto slabije koeficijente korelacije. Također, brojevi zaraze Ujedinjenog Kraljevstva imaju najveću korelaciju sa Slovenijom, a najmanju s Belgijom. Nedavnim izlaskom Ujedinjenog Kraljevstva iz Europske unije te drukčijim pristupom suzbijanju pandemije koronavirusa od ostalih zemalja Europske unije, sukladno je za očekivati da će koeficijenti korelacije biti manji nego za ostale države.

Koeficijenti korelacije u 4.1 značajni su za sve parove država, a time možemo naslutiti da za ovaj period postoje jake korelacije između podataka, tj. postoji jaka povezanost između toka zaraze između svih navedenih zemalja prije druge doze cjepiva. Da bismo dokazali ili opovrgnuli ovu tvrdnju, koristit ćemo korelacijske testove. Postavljamo sljedeće hipoteze:

H_0: $\rho = 0$ (ρ je koeficijent korelacije varijabli daily weekly average)
H_1: ne H_0

Nulta hipoteza tvrdi da korelacije između podataka ne postoji, dok alternativna hipoteza tvrdi suprotno. Za svaki par država dobivamo sljedeći rezultat:

```
1
2         Spearman's rank correlation rho
3
4 data:  data1 and data2
5 S = ..., p-value < 2.2e-16
6 alternative hypothesis: true rho is not equal to 0
7 sample estimates:
8       rho
9       ...
```

Kako je p-vrijednost jako mala, u svakom testu odbacujemo hipotezu H_0 na svim razumnim razinama značajnosti pa pretpostavljamo da su dnevni tjedni prosjeci zaraze svih država

međusobno korelirani. Dakle, pokazali smo da kako su države u Europskoj uniji povezane, tako su i tokovi epidemije svake države bili povezani do datuma 30.6.2021.

Analogan postupak koristimo za varijablu stope zaraze. Spearmanovi koeficijenti korelacije su sljedeći:

	Italija	Austrija	Njemačka	Slovenija	Hrvatska
Italija	1	0.909355	0.864927	0.8831378	0.829858
Austrija	0.909355	1	0.838934	0.879037	0.872946
Njemačka	0.864927	0.838934	1	0.873549	0.858847
Slovenija	0.883138	0.879037	0.873549	1	0.855290
Hrvatska	0.829858	0.872946	0.858847	0.855290	1
Ujedinjeno Kraljevstvo	0.601822	0.598622	0.487149	0.710085	0.460433
Danska	0.764638	0.700028	0.857752	0.828557	0.787902
Nizozemska	0.819053	0.796642	0.802868	0.8675412	0.755435
Belgija	0.827418	0.819898	0.757868	0.833270	0.751789

	Ujedinjeno Kraljevstvo	Danska	Nizozemska	Belgija
Italija	0.601822	0.764638	0.819053	0.827418
Austrija	0.598622	0.700028	0.796642	0.819898
Njemačka	0.487149	0.857752	0.802868	0.757868
Slovenija	0.710085	0.828557	0.867542	0.833270
Hrvatska	0.460433	0.787902	0.755435	0.751789
Ujedinjeno Kraljevstvo	1	0.441381	0.582788	0.553633
Danska	0.441381	1	0.803595	0.718503
Nizozemska	0.582788	0.803595	1	0.921099
Belgija	0.553633	0.718503	0.921099	1

Tablica 4.2: Spearmanovi koeficijenti korelacije za stopu zaraze

Kod varijable stopa zaraze uočavamo velike korelacijske koeficijente kod svih parova država osim kod parova s Ujedinjenim Kraljevstvom. Kako smo dobili identične rezultate kao i kod varijable daily weekly average, ovime samo potvrđujemo da smo ispravno transformirali i diskretizirali varijablu stope zaraze.

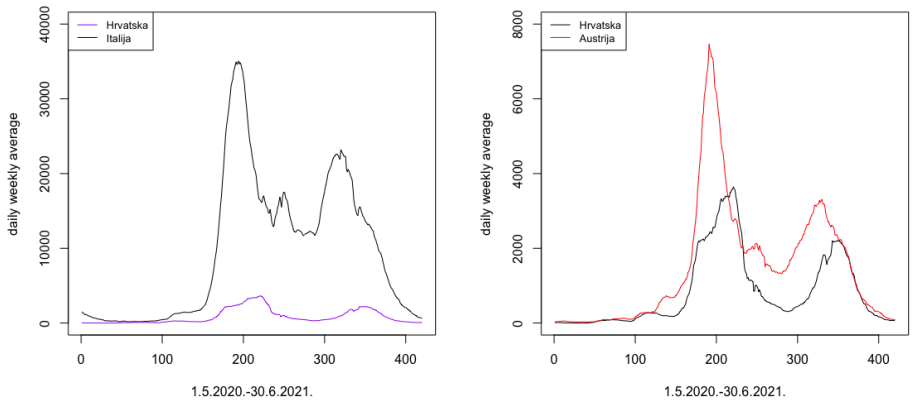
Da bismo dokazali da velike korelacije zaista postoje, ponovno postavljamo hipoteze:

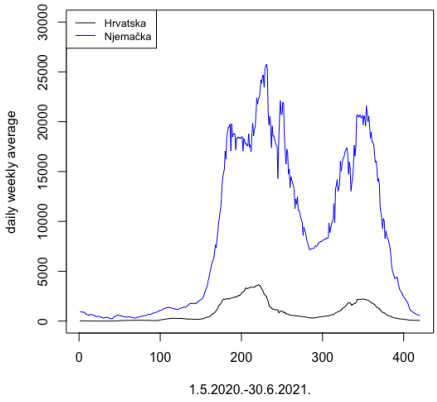
H_0: $\rho = 0$
H_1: ne H_0

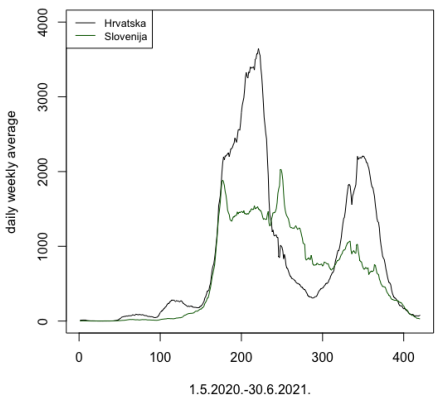
Isto kao i kod varijable dnevnog tjednog prosjeka zaraze, za svaki par država u korelacijskom testu dobivamo p-vrijednost manju od $2.2 \cdot 10^{-16}$ pa na svim razumnim razinama značajnosti odbacujemo hipotezu H_0. Dakle, pretpostavljamo da su podaci korelirani, a kako vidimo iz 4.2, te korelacije su značajne.

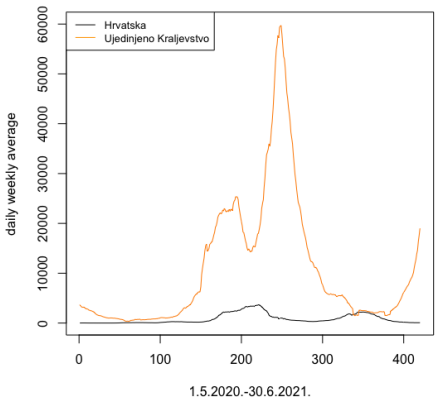
Analizom i statističkim testovima ustvrdili smo da su u periodu od 1.5.2020. do 30.6. 2021. postojale značajne korelacije u toku zaraze koronavirusom između odabranih zemalja u Europi. Zbog različitih vrsta povezanosti zemalja u Europi, očekivali smo da će rast/pad novih slučajeva u jednoj državi utjecati na rast/pad u drugoj, no zanimljiva je činjenica da su korelacije velike, tj. da je tok zaraze virusom jedne države značajno utjecao na tok zaraze virusom druge države u tom periodu.

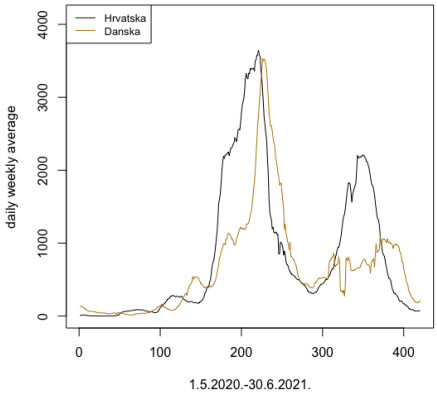
Posebno nas zanima situacija u Hrvatskoj u usporedbi s ostalim državama. Točnije, kako smo pokazali da korelacija Hrvatske s bilo kojom drugom državom postoji i značajna je, želimo vidjeti kolika je vremenska razlika toka zaraze u Hrvatskoj u odnosu na ostale države.

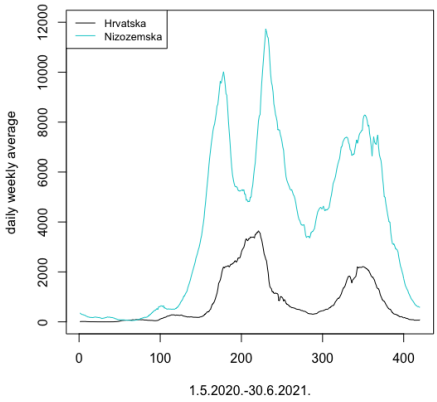


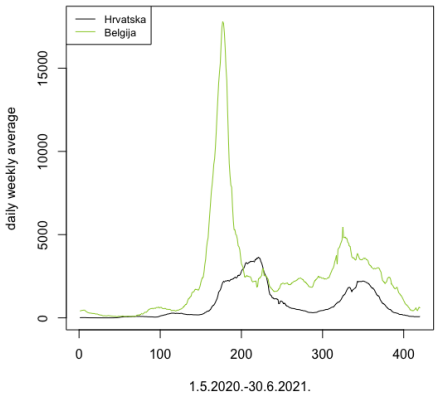

Hrvatska
Njemačka
daily weekly average
0
5000
10000
15000
20000
25000
30000
0
100
200
300
400
1.5.2020.-30.6.2021.


Hrvatska
Slovenija
daily weekly average
0
1000
2000
3000
4000
0
100
200
300
400
1.5.2020.-30.6.2021.


Hrvatska
Ujedinjeno Kraljevstvo
daily weekly average
0
10000
20000
30000
40000
50000
60000
0
100
200
300
400
1.5.2020.-30.6.2021.


Hrvatska
Danska
daily weekly average
0
1000
2000
3000
4000
0
100
200
300
400
1.5.2020.-30.6.2021.


Hrvatska
Nizozemska
daily weekly average
0
2000
4000
6000
8000
10000
12000
0
100
200
300
400
1.5.2020.-30.6.2021.


Hrvatska
Belgija
daily weekly average
0
5000
10000
15000
0
100
200
300
400
1.5.2020.-30.6.2021.

Iz grafova se može vidjeti da je u periodu 1.5.2020.-30.6.2021. vremenska razlika u toku zaraze između Hrvatske i ostalih zemalja sljedeća:

- Hrvatska - Italija: oko 14 dana
- Hrvatska - Austrija: oko 20 dana
- Hrvatska - Njemačka: oko 20 dana
- Hrvatska - Slovenija: par dana
- Hrvatska - Ujedinjeno Kraljevstvo: oko 100 dana
- Hrvatska - Danska: oko 40 dana
- Hrvatska - Nizozemska: 40-50 dana
- Hrvatska - Belgija: oko 30 dana

Podaci za period 1.7.2021. - 13.10.2021.

Za podatke u drugom periodu koristimo identičan postupak kao i za prvi period. Kako pretpostavljamo da podaci o dnevnom tjednom prosjeku zaraze nisu normalno dustribuirani, iznova izračunavamo Spearmanove koeficijente korelacije.

	Italija	Austrija	Njemačka	Slovenija	Hrvatska
Italija	1	0.248352	0.519745	0.235953	0.184025
Austrija	0.248352	1	0.865164	0.971937	0.955495
Njemačka	0.519745	0.865164	1	0.827495	0.810763
Slovenija	0.235953	0.971937	0.827495	1	0.947419
Hrvatska	0.184025	0.955495	0.810763	0.947419	1
Ujedinjeno Kraljevstvo	-0.114182	0.275793	0.347959	0.214825	0.317479
Danska	0.606639	-0.348507	-0.103625	-0.407284	-0.339794
Nizozemska	0.144848	-0.463933	-0.257439	-0.508667	-0.491509
Belgija	0.484742	0.832936	0.883339	0.829011	0.821836

	Ujedinjeno Kraljevstvo	Danska	Nizozemska	Belgija
Italija	-0.114182	0.606639	0.144848	0.484742
Austrija	0.275793	-0.348507	-0.463933	0.832936
Njemačka	0.347959	-0.103625	-0.257439	0.883339
Slovenija	0.214825	-0.407284	-0.508667	0.829012
Hrvatska	0.317479	-0.339794	-0.491509	0.821836
Ujedinjeno Kraljevstvo	1	0.043584	0.28390	0.236235
Danska	0.043584	1	0.614715	-0.07664
Nizozemska	0.283900	0.614715	1	-0.271224
Belgija	0.236235	-0.07664	-0.271224	1

Tablica 4.3: Spearmanovi koeficijenti korelacije za dnevni tjedni prosjek zaraze

Iz tablice 4.3 možemo vidjeti da su se u ovom periodu dogodile mnoge promjene. Čini se kako dnevni tjedni prosjek zaraze Ujedinjenog Kraljevstva više nije koreliran niti s jednim prosjekom zaraze druge države, a da prosjek zaraze Italije ostaje koreliran jedino s prosjekom Njemačke i Danske. Također možemo uočiti da Austrija, Njemačka, Slovenija, Hrvatska i Belgija ostaju međusobno korelirane, ali niti s jednom drugom državom (osim dodatno Njemačka i Italija). Danska ostaje korelirana s Nizozemskom i Italijom, a Nizozemska jedino s Danskom.

Da bismo dokazali ili opovrgnuli naše tvrdnje, koristimo Spearmanov korelacijski test. Postavljamo hipoteze:

H_0: $\rho = 0$
H_1: ne H_0

Prikazujemo korelacijske testove u kojima nismo mogli odbaciti nultu hipotezu:

```
1
2         Spearman's rank correlation rho
3
4 data:  dwa_it2 and dwa_hr2
5 S = 157418, p-value = 0.06022
6 alternative hypothesis: true rho is not equal to 0
7 sample estimates:
8       rho
9 0.1840245
```

```
1
2         Spearman's rank correlation rho
3
4 data:  dwa_it2 and dwa_uk2
```

```
S = 214948, p-value = 0.2461
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.114182
```

```

        Spearman's rank correlation rho

data:  dwa_it2 and dwa_ne2
S = 164976, p-value = 0.1404
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1448476
```

```

        Spearman's rank correlation rho

data:  dwa_dm2 and dwa_de2
S = 212911, p-value = 0.2928
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1036251
```

```

        Spearman's rank correlation rho

data:  dwa_uk2 and dwa_dm2
S = 184512, p-value = 0.6589
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.04358351
```

```

        Spearman's rank correlation rho

data:  dwa_dm2 and dwa_be2
S = 207705, p-value = 0.4371
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.07663973
```

Iz rezultata korelacijskih testova zaključujemo da na razini značajnosti 5% ne odbacujemo nultu hipotezu za parove država Italija i Hrvatska, Italija i Ujedinjeno Kraljevstvo, Italija

i Nizozemska, Danska i Njemačka, Ujedinjeno Kraljevstvo i Danska te Danska i Belgija. Dakle, u periodu od 1.7.2021. do 13.10.2021. pretpostavljamo da korelacija između svih navedenih država ne postoji. Ostale korelacije za koje smo spomenuli da ih možda više nema ipak postoje, ali nisu značajne.

Provjerimo sada korelacije za varijablu stope smrtnosti ponovno koristeći Spearmanove koeficijente korelacije:

	Italija	Austrija	Njemačka	Slovenija	Hrvatska
Italija	1	0.17372	0.258765	0.130022	-0.051131
Austrija	0.17372	1	0.682727	0.961522	0.871165
Njemačka	0.258765	0.682727	1	0.651025	0.667311
Slovenija	0.130022	0.961522	0.651025	1	0.884267
Hrvatska	-0.051131	0.871165	0.667311	0.884267	1
Ujedinjeno Kraljevstvo	-0.386844	0.136077	0.144707	0.078133	0.182671
Danska	0.484123	-0.268119	0.131006	-0.388951	-0.46757
Nizozemska	0.122289	-0.075770	0.269478	-0.192043	-0.232186
Belgija	0.200805	0.523257	0.581876	0.385297	0.34909

	Ujedinjeno Kraljevstvo	Danska	Nizozemska	Belgija
Italija	-0.386844	0.484123	0.122289	0.200805
Austrija	0.136077	-0.268119	-0.075770	0.523257
Njemačka	0.144707	0.131006	0.269478	0.581876
Slovenija	0.078133	-0.388951	-0.192043	0.385297
Hrvatska	0.182671	-0.46757	-0.232186	0.34909
Ujedinjeno Kraljevstvo	1	-0.017626	0.245327	0.184604
Danska	-0.017626	1	0.545614	0.354292
Nizozemska	0.245327	0.545614	1	0.309081
Belgija	0.184604	0.354292	0.309081	1

Tablica 4.4: Spearmanovi koeficijenti korelacije za stopu zaraze

Pogledamo li tablicu 4.4, možemo primijetiti da Italija i Ujedinjeno Kraljevstvo nemaju značajan koeficijent korelacije stope zaraze niti s jednom drugom državom, dok Danska i Nizozemska jedino imaju značajnije koeficijente jedna s drugom. Podaci za Belgiju imaju značajnije koeficijente jedino s podacima Austrije i Njemačke, a Austrija, Njemačka, Slovenija i Hrvatska imaju međusobno velike koeficijente korelacije.

Da provjerimo je li zaista toliko korelacija nestalo, koristimo Spearmanov korelacijski test. Kako su u prvom periodu svi podaci međusobno bili korelirani i u svakom slučaju smo odbacili hipotezu H_0, ovdje ćemo ponovno prikazati statističke testove u kojima nismo mogli odbaciti nultu hipotezu. Postavljamo hipoteze:

H_0: $\rho = 0$
H_1: ne H_0

```
        Spearman's rank correlation rho

data:  inf_rate_it2 and inf_rate_at2
S = 159406, p-value = 0.07635
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1737198
```

```
        Spearman's rank correlation rho

data:  inf_rate_it2 and inf_rate_slo2
S = 167836, p-value = 0.1862
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1300216
```

```
        Spearman's rank correlation rho

data:  inf_rate_it2 and inf_rate_hr2
S = 202784, p-value = 0.6045
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.051131
```

```
        Spearman's rank correlation rho

data:  inf_rate_it2 and inf_rate_ne2
S = 169328, p-value = 0.214
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.122289
```

```
        Spearman's rank correlation rho

data:  inf_rate_uk2 and inf_rate_at2
S = 166668, p-value = 0.1663
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1360772
```

```
        Spearman's rank correlation rho

data:  inf_rate_ne2 and inf_rate_at2
S = 207538, p-value = 0.4424
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.07577036
```

```
        Spearman's rank correlation rho

data:  inf_rate_uk2 and inf_rate_de2
S = 165003, p-value = 0.1408
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1447067
```

```
        Spearman's rank correlation rho

data:  inf_rate_dm2 and inf_rate_de2
S = 167646, p-value = 0.1828
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1310056
```

```
        Spearman's rank correlation rho

data:  inf_rate_uk2 and inf_rate_slo2
S = 177847, p-value = 0.4282
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
```

```
0.07813283
```

```

        Spearman's rank correlation rho

data:  inf_rate_uk2 and inf_rate_dm2
S = 196320, p-value = 0.8584
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01762633
```

```

        Spearman's rank correlation rho

data:  inf_rate_uk2 and inf_rate_be2
S = 157306, p-value = 0.0594
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1846043
```

```

        Spearman's rank correlation rho

data:  inf_rate_hr2 and inf_rate_uk2
S = 157679, p-value = 0.06216
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1826713
```

Na razini značajnosti 5% ne odbacujemo nultu hipotezu za sljedeće parove: Italija i Austrija, Italija i Slovenija, Italija i Hrvatska, Italija i Nizozemska, Ujedinjeno Kraljevstvo i Austrija, Nizozemska i Austrija, Ujedinjeno Kraljevstvo i Njemačka, Danska i Njemačka, Ujedinjeno Kraljevstvo i Slovenija, Ujedinjeno Kraljevstvo i Danska, Ujedinjeno Kraljevstvo i Belgija te Ujedinjeno Kraljevstvo i Hrvatska. Dakle, pretpostavljamo da u periodu od 1.7.2021. do 13.10.2021. ne postoje korelacije stope zaraze za navedene države. Ostale slabije korelacije koje smo spomenuli postoje između država, ali nisu značajne. Primijetimo da Ujedinjeno Kraljevstvo gubi korelaciju s čak šest država, Italija s četiri, Austrija s tri, Hrvatska, Slovenija, Njemačka i Nizozemska s dvije i Belgija s jednom.

Važno je još napomenuti da Italija gubi korelaciju s Hrvatskom, Austrijom i Slovenijom koje sve imaju nešto važniju ulogu za Hrvatsku zbog jače povezanosti i lokacije, ali Austrija, Slovenija i Hrvatska međusobno i dalje imaju jake korelacije.

Uočimo kako za obje varijable u ovom periodu gubimo korelacije između Italije i Hrvatske, Italije i Nizozemske, Danske i Njemačke te Ujedinjenog Kraljevstva i Danske. Posebno, za varijablu dnevni tjedni prosjek zaraze gubi se korelacija između Italije i Ujedinjenog Kraljevstva te Danske i Belgije, a za varijablu stopa zaraze Italija i Austrija, Italija i Slovenija, Ujedinjeno Kraljevstvo i Austrija, Nizozemska i Austrija, Ujedinjeno Kraljevstvo i Njemačka, Ujedinjeno Kraljevsvo i Slovenija, Ujedinjeno Kraljevstvo i Belgija te Ujedinjeno Kraljevstvo i Hrvatska.

Spomenuli smo da je razlog zbog kojeg smo razdvojili podatke na prije i poslije 1.7.2021. je taj da je to period u kojem su cijepljeni dobili drugu dozu. Stoga, pokažimo još procijepljenost[1] u svakoj navedenoj državi:

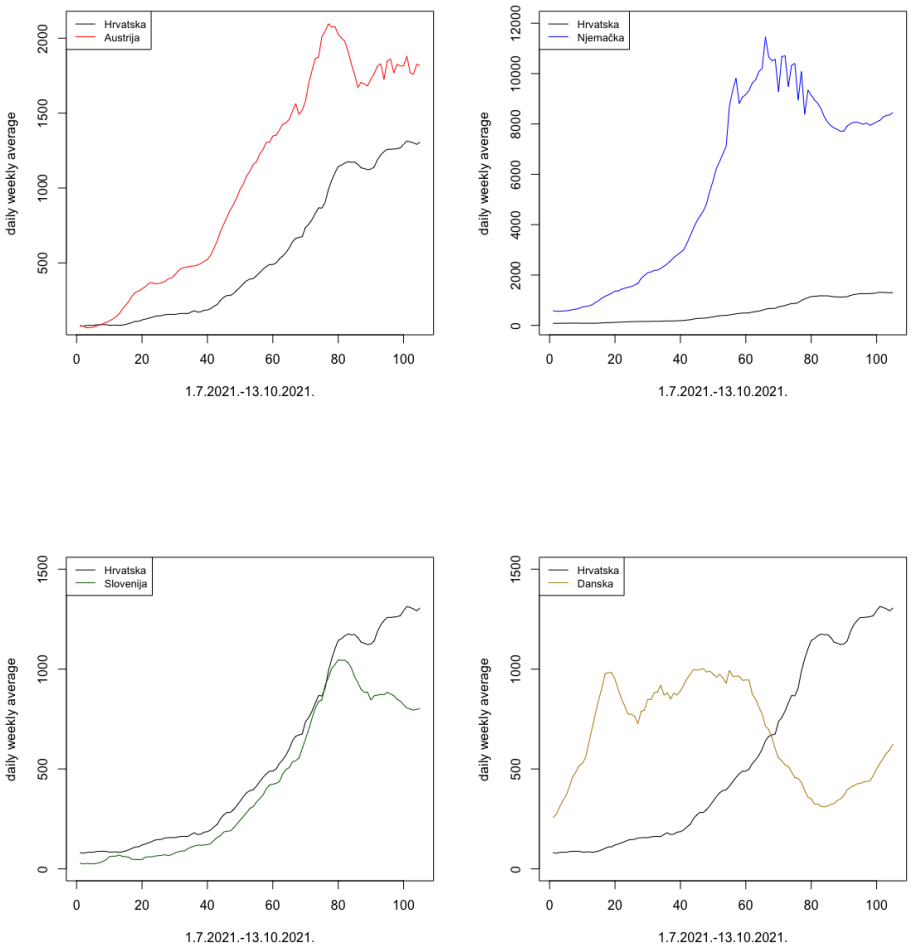
	procijepljenost
Italija	70.08%
Austrija	62.31%
Njemačka	65.75%
Slovenija	50.99%
Hrvatska	43.14%
Ujedinjeno Kraljevstvo	68.24%
Danska	75.75%
Nizozemska	67.99%
Belgija	74.38%

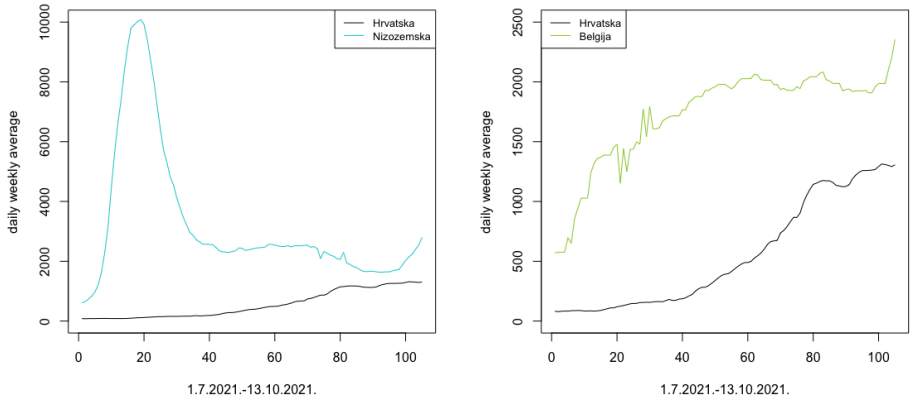
Ujedinjeno Kraljevstvo koje gubi korelacije s najviše ostalih država ima visoku procijepljenost, ali ne najveću, no znamo da ima najmanju povezanost s drugim državama (možda je i izlazak iz Europske unije utjecao na to). Nemamo konkretnih dokaza da veća procijepljenost utječe na promjenu, ali činjenica je da se korelacije gube nakon druge doze cjepiva. Također, oni su imali nešto drukčiji pristup prema suzbijanju zaraze i drukčije mjere nego u ostatku Europe pa su i to mogući čimbenici koji su pridonijeli tome.

Italija ima jedan od najvećih postotaka procijepljenosti pa je zanimljivo da korelacija njenih podataka nestaje sa susjednim zemljama - Austrijom, Slovenijom i Hrvatskom, a korelacije između Austrije, Slovenije i Hrvatske ne nestaju. Pogledamo li tablicu s procijepljenostima svake zemlje, uočit ćemo da Italija jedina ima procijepljenost iznad 70%, dok Austrija, Slovenija i Hrvatska imaju procijepljenost ispod 70%. U ovom slučaju ponovno nemamo konkretnih dokaza da velika procijepljenost i različit pristup suzbijanju pandemije imaju utjecaj na prestanak korelacije podataka Italije i ostalih država, no opet moramo biti svjestni da se promjene događaju nakon perioda kada su cijepljeni dobili drugu dozu cjepiva.

[1] podaci se odnose na procijepljenost cijelog stanovništva jedne države do datuma 13.10.2021. i preuzeti su s `https://coronavirus.jhu.edu/region`

Zanimljivo je istaknuti vremensku razliku u toku zaraze Hrvatske s ostalim državama s kojima i dalje ima korelaciju. Prikazat ćemo podatke varijable dnevni tjedni broj zaraženih pa kako Hrvatska gubi korelaciju samo s Italijom, vremenska razlika između Hrvatske i Italije neće biti opisana.





Iz grafova primijećujemo da su vremenske razlike u toku zaraze između Hrvatske i njoj koreliranih zemalja sljedeće:

- Hrvatska - Austrija: oko 20 dana
- Hrvatska - Njemačka: oko 30 dana
- Hrvatska - Slovenija: par dana
- Hrvatska - Danska: oko 60 dana
- Hrvatska - Nizozemska: oko 60 dana
- Hrvatska - Belgija: oko 60 dana

4.1.2 Model linearne regresije

Nakon što smo testirali postojanje korelacije za obje varijable, korelirane podatke možemo transformirati u model linearne regresije. Izračunat ćemo četiri modela linearne regresije za varijablu stope zaraze - dvije za prvi period i dvije za drugi period.

Kao što smo naveli u poglavlju 2.1.3, za dobar model linearne regresije trebaju nam četiri glavne pretpostavke:

- linearan odnos između zavisne i nezavisnih varijabli
- nezavisnost grešaka

- homogenost grešaka
- normalna distribuiranost grešaka

Kako su za prvi period svi podaci međusobno korelirani, u linearnom modelu možemo imati samo jednu nezavisnu i jednu zavisnu varijablu. Koristit ćemo podatke za Hrvatsku i Italiju te Hrvatsku i Austriju. Italiju i Austriju uzet ćemo za nezavisne varijable, a Hrvatsku za zavisnu varijablu. Dobivamo sljedeće rezultate:

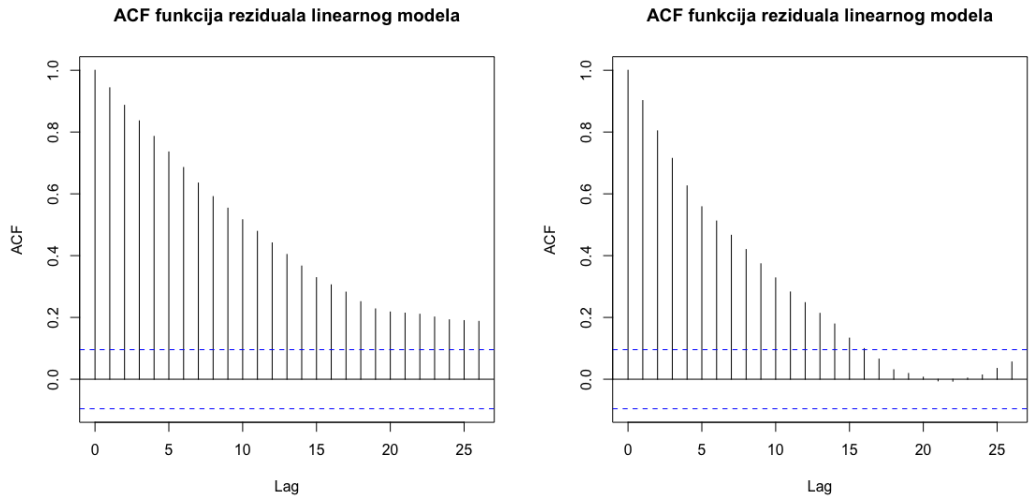
```
> summary(lm(inf_rate_hr1 ~ inf_rate_it1))

Call:
lm(formula = inf_rate_hr1 ~ inf_rate_it1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24678 -0.24678 -0.08699  0.59343  0.91301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.40657    0.05105   7.964 1.59e-14 ***
inf_rate_it1  0.84021    0.03189  26.343  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5596 on 418 degrees of freedom
Multiple R-squared:  0.6241, Adjusted R-squared:  0.6232
F-statistic:   694 on 1 and 418 DF,  p-value: < 2.2e-16
```

```
> summary(lm(inf_rate_hr1 ~ inf_rate_at1))

Call:
lm(formula = inf_rate_hr1 ~ inf_rate_at1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08046 -0.13175 -0.08046 -0.08046  0.91954

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.18304    0.04412   4.149 4.04e-05 ***
inf_rate_at1  0.94871    0.02661  35.647  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.4541 on 418 degrees of freedom
Multiple R-squared:  0.7525, Adjusted R-squared:  0.7519
F-statistic:  1271 on 1 and 418 DF,  p-value: < 2.2e-16
```

Prvi model linearne regresije ima vrijednost R^2 jednaku 0.6241, a drugi 0.7525. Vrijednost R^2 opisuje varijabilnost modela, što znači da su oba modela imaju veliku varijabilnost. Ipak, moramo provjeriti zadovoljavaju li naši modeli pretpostavke dobrog modela.

Da bismo provjerili nezavisnost grešaka oba modela, koristimo `acf()` funkciju u `R` - u:



Slika 4.1: Reziduali linearnih modela Hrvatske i Italije (lijevo) te Hrvatske i Austrije (desno).

Vidimo da greške niti jednog modela nisu nezavisne pa zbog toga linearni modeli neće biti dobri.

Kako bi se izbjegla zavisnost grešaka, koristi se linearna regresija s ARIMA grešakama. Funkcija `auto.arima()` u `R`-u sama preko AIC kriterija (model s najmanjom vrijednosti AIC kriterija izabire se kao najbolji model) određuje koji je ARIMA model najbolji za aproksimaciju greške.

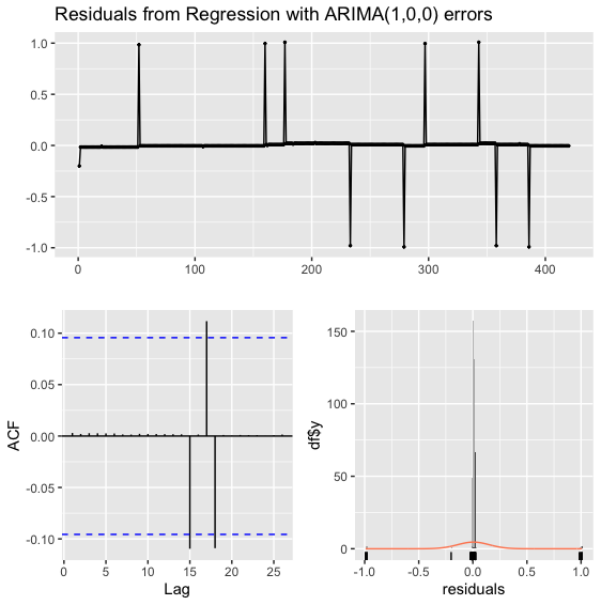
```

> fit <- auto.arima(inf_rate_hr1, xreg = inf_rate_it1)
> summary(fit)
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept    xreg
      0.9878     1.2818  0.0076
```

```
s.e.  0.0071     0.5086  0.0603

sigma^2 estimated as 0.0215:  log likelihood=210
AIC=-412   AICc=-411.9   BIC=-395.84

Training set error measures:
                     ME      RMSE       MAE  MPE MAPE     MASE
Training set 0.00496472 0.1461141 0.0310433 -Inf  Inf 1.445238
                    ACF1
Training set 0.002975647
```

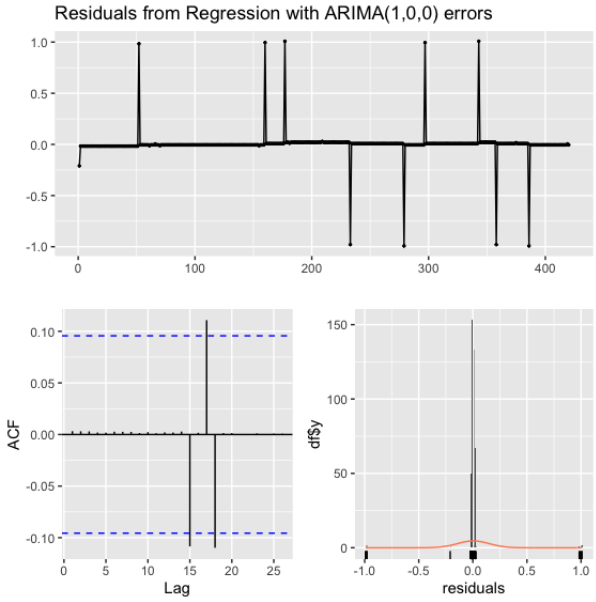
```

> fit <- auto.arima(inf_rate_hr1, xreg = inf_rate_at1)
> summary(fit)
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept    xreg
      0.9875     1.3307  0.0090
s.e.  0.0071     0.4910  0.0524

sigma^2 estimated as 0.0215:  log likelihood=209.99
AIC=-411.98   AICc=-411.89   BIC=-395.82

Training set error measures:
                      ME      RMSE        MAE  MPE MAPE     MASE
Training set 0.004376987 0.1461208 0.03137849 -Inf  Inf 1.460843
                    ACF1
Training set 0.003222598
```

Funkcija je izabrala $ARIMA(1,0,0)$, tj. $AR(1)$ za najbolji model za obje linearne regresije. σ^2 vrijednost jako je mala (blizu nuli), a iz definicije kriterija za dobar model u 2.1.3 vidimo da bi tako trebalo i biti.

Zbog definicije, ovakav model sadrži dvije vrste grešaka - greške linearne regresije i greške ARIMA modela. Ako je model dobar, greške ARIMA modela trebale bi biti bijeli šum 2.1.20. Provjerimo zadovoljavaju li greške pretpostavke dobrog modela:



Slika 4.2: Reziduali linearne regresije Hrvatske i Italije.



Slika 4.3: Reziduali linearne regresije Hrvatske i Austrije.

```
> checkresiduals(fit)
```

```
2
3         Ljung-Box test
4
5 data:  Residuals from Regression with ARIMA(1,0,0) errors
6 Q* = 0.022934, df = 7, p-value = 1
7
8 Model df: 3.   Total lags used: 10
```

```
1 > checkresiduals(fit)
2
3       Ljung-Box test
4
5 data:  Residuals from Regression with ARIMA(1,0,0) errors
6 Q* = 0.02706, df = 7, p-value = 1
7
8 Model df: 3.   Total lags used: 10
```

Iz grafičkog prikaza ACF funkcije reziduala zaključujemo da su reziduali nezavisni te da zadovoljavaju uvjete dobrog modela. Hipoteze gore nevedenog Ljung-Box testa su sljedeće:

H_0 : procjena modela je zadovoljavajuća
H_1 : ne H_0

p-vrijednosti oba modela veće su od 0.05 pa na razini značajnosti 5% ne odbacujemo hipotezu H_0 ni za jedan model. Pretpostavljamo da su procjene modela zadovoljavajuće.

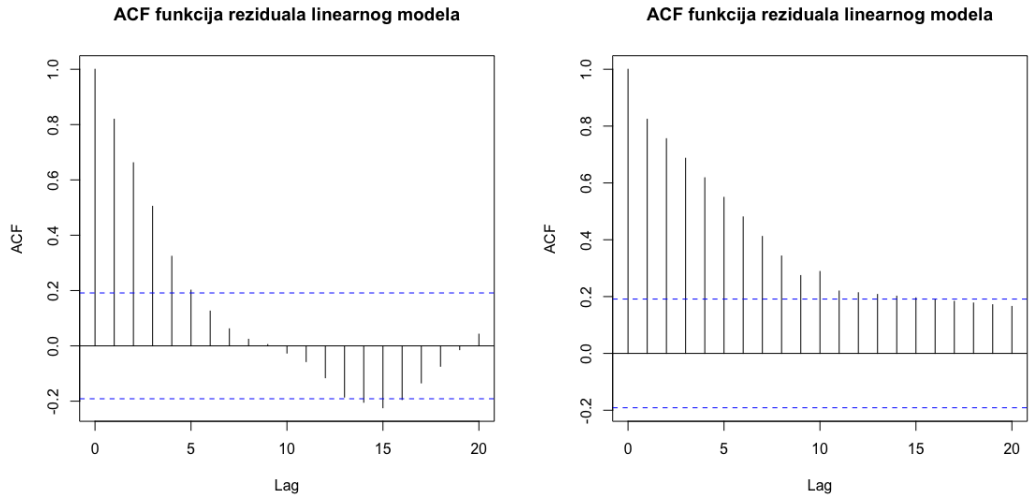
Za drugi period uzimamo Ujedinjeno Kraljevstvo i Nizozemsku te Belgiju i Dansku za modele linearne regresije. Nizozemsku i Dansku postavljamo za nezavisne varijable, a Ujedinjeno Kraljevstvo i Belgiju za zavisne varijable.

```
1 > summary(lm(inf_rate_uk2 ~ inf_rate_ne2))
2
3 Call:
4 lm(formula = inf_rate_uk2 ~ inf_rate_ne2)
5
6 Residuals:
7     Min      1Q  Median      3Q     Max
8 -0.3475 -0.3475 -0.3475  0.4034  0.9015
9
10 Coefficients:
11             Estimate Std. Error t value Pr(>|t|)
12 (Intercept)   1.84943    0.17925  10.318  < 2e-16 ***
13 inf_rate_ne2  0.24905    0.09115   2.732  0.00741 **
14 ---
15 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
16
```

```
Residual standard error: 0.4562 on 103 degrees of freedom
Multiple R-squared:  0.06758,  Adjusted R-squared:  0.05852
F-statistic: 7.465 on 1 and 103 DF,  p-value: 0.007405
```

```
> summary(lm(inf_rate_be2 ~ inf_rate_dm2))

Call:
lm(formula = inf_rate_be2 ~ inf_rate_dm2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98333  0.01667  0.01667  0.24444  0.24444

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.52778    0.09760  15.654  < 2e-16 ***
inf_rate_dm2  0.22778    0.05924   3.845 0.000209 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3004 on 103 degrees of freedom
Multiple R-squared:  0.1255, Adjusted R-squared:  0.117
F-statistic: 14.78 on 1 and 103 DF,  p-value: 0.0002089
```

Pogledamo li vrijednosti R^2 oba modela, vidimo da oba modela imaju jako malu varijabilnost te već time naslućujemo da modeli nisu dobri. Provjerimo još nezavisnost grešaka:



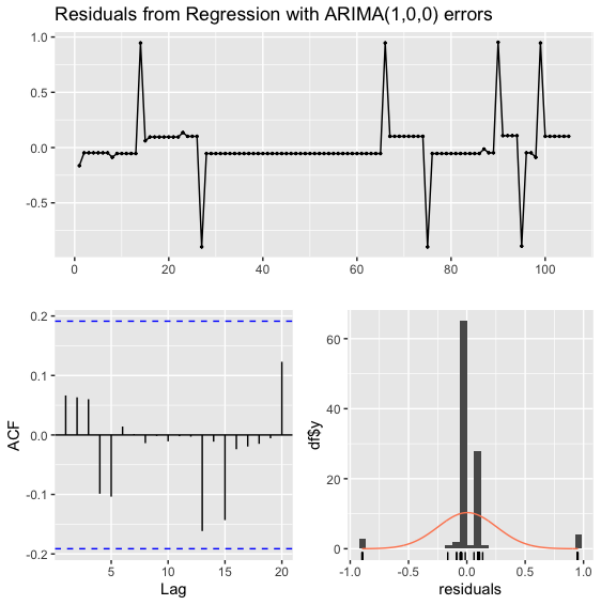
Slika 4.4: Reziduali linearnih modela Ujedinjenog Kraljevstva i Nizozemske (lijevo) te Belgije i Danske (desno).

Možemo uočiti da postoji velika zavisnost među rezidualima. Da se izbjegne zavisnost grešaka, ponovno koristimo modele linearne regresije s ARIMA greškama.

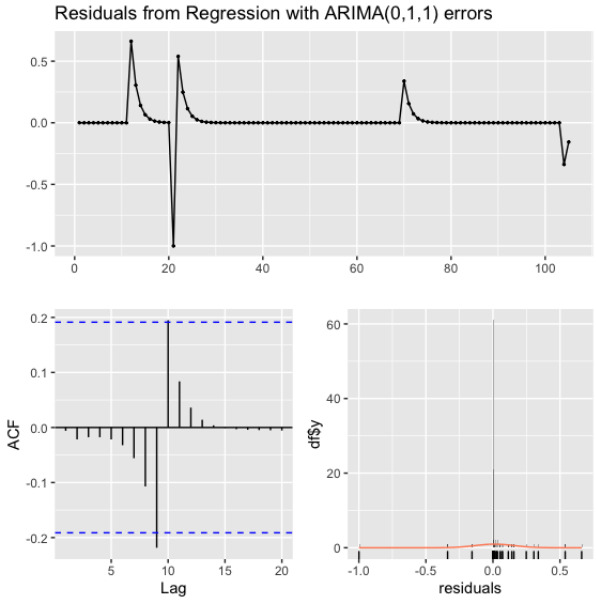
```
> fit <- auto.arima(inf_rate_uk2, xreg = inf_rate_ne2)
> summary(fit)
Series: inf_rate_uk2
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept    xreg
      0.8446     2.2654  0.0404
s.e.  0.0526     0.2616  0.1144

sigma^2 estimated as 0.06404:  log likelihood=-3.81
AIC=15.62   AICc=16.02   BIC=26.24

Training set error measures:
                      ME      RMSE       MAE        MPE     MAPE
Training set 0.003749325 0.2494253 0.1260604 -0.8942817 5.221273
                  MASE    ACF1
Training set 1.872898 0.06639
```

```
> fit <- auto.arima(inf_rate_be2, xreg = inf_rate_dm2)
> summary(fit)
Regression with ARIMA(0,1,1) errors

Coefficients:
          ma1    xreg
      -0.4613  0.3383
s.e.   0.0919  0.0753

sigma^2 estimated as 0.02165:  log likelihood=52.63
AIC=-99.25   AICc=-99.01   BIC=-91.32

Training set error measures:
                     ME      RMSE        MAE       MPE     MAPE
Training set 0.01297791 0.1450182 0.04142999 0.1733184 2.547707
                MASE         ACF1
Training set 1.43624 -0.006010829
```

AIC kriterijem izabrani su $ARIMA(1,0,0)$ ($AR(1)$) i $ARIMA(0,1,1)$ modeli za greške linearne regresije. Vrijednosti σ^2 jako su male, a to je jedan indikator da bi modeli mogli biti zadovoljavajući. Provjerimo još pretpostavke za greške modela:



Slika 4.5: Reziduali linearne regresije Ujedinjenog Kraljevstva i Nizozemske.



Slika 4.6: Reziduali linearne regresije Belgije i Danske.

```
> checkresiduals(fit)
```

```
2
3       Ljung-Box test
4
5  data:  Residuals from Regression with ARIMA(1,0,0) errors
6  Q* = 3.6605, df = 7, p-value = 0.8179
7
8  Model df: 3.   Total lags used: 10
```

```
1  > checkresiduals(fit)
2
3       Ljung-Box test
4
5  data:  Residuals from Regression with ARIMA(0,1,1) errors
6  Q* = 12.056, df = 8, p-value = 0.1487
7
8  Model df: 2.   Total lags used: 10
```

ACF funkcije reziduala pokazuju da su reziduali nezavisni te za zadovoljavaju uvjete dobrog modela. Kako su p-vrijednosti Ljung-Box testa veće od 0.05, na razini značajnosti 5% ne odbacujemo hipotezu H_0. Dakle, pretpostavljamo da su modeli zadovoljavajući.

4.2 Diskretno modeliranje

Jedan od razloga zbog kojeg smo stvorili diskretiziranu varijablu stope zaraze je taj da ju možemo modelirati na više načina. Jedan od tih načina je da ju pretvorimo u skriven Markovljev model, a time potencijalno otkrivamo i dokazujemo nova zanimljiva otkrića vezana uz koronavirus. U skriven Markovljev modelirat ćemo podatke stope zaraze Italije, Austrije, Njemačke, Slovenije i Hrvatske.

Iz definicije 2.2.8 vidimo da je potrebno definirati stanja, emisijske vjerojatnosti svakog stanja, tranzicijsku matricu stanja te početnu distribuciju. Kako je model stacionaran, iz teorema 2.2.7 dovoljno je izračunati stacionarnu distribuciju jer je ona tada i početna distribucija.

Definiramo pet stanja: Ništa, Malo, Srednje, Puno i Extra, a vrijednosti stope zaraze kao niz opservacija. Postavljamo matricu emisijskih vjerojatnosti

$$\begin{pmatrix} 0.9 & 0.05 & 0.03 & 0.01 & 0.01 \\ 0.05 & 0.9 & 0.03 & 0.01 & 0.01 \\ 0.03 & 0.05 & 0.9 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.05 & 0.9 & 0.03 \\ 0.01 & 0.01 & 0.03 & 0.05 & 0.9 \end{pmatrix}$$

te tranzicijsku matricu

$$\begin{pmatrix} 0.70 & 0.10 & 0.1 & 0.05 & 0.05 \\ 0.10 & 0.70 & 0.1 & 0.05 & 0.05 \\ 0.05 & 0.10 & 0.7 & 0.10 & 0.05 \\ 0.05 & 0.05 & 0.1 & 0.70 & 0.10 \\ 0.05 & 0.05 & 0.1 & 0.10 & 0.70 \end{pmatrix}$$

Pomoću jednakosti (2.15) iz definicije 2.2.6 dobivamo stacionarnu distribuciju:
$\Pi = \left(\frac{11}{64}, \frac{13}{64}, \frac{16}{64}, \frac{13}{64}, \frac{11}{64}\right)$.

Kako bismo saznali koji je niz stanja najvjerojatnije emitirao pet nizova opservacija, koristimo Viterbijev algoritam 2.2.2. Rezultati za Hrvatsku su sljedeći:

```
> stanja_viterbi_hr
  [1] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
  [7] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [13] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [19] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [25] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [31] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [37] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [43] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [49] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [55] "Nista"   "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [61] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [67] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [73] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [79] "Malo"    "Malo"    "Malo"    "Nista"   "Nista"   "Nista"
 [85] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [91] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
 [97] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[103] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[109] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[115] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[121] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[127] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[133] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[139] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[145] "Nista"   "Nista"   "Nista"   "Nista"   "Nista"   "Nista"
[151] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[157] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[163] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[169] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[175] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[181] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
```

```
[187] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[193] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[199] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[205] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[211] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[217] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[223] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[229] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[235] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[241] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[247] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[253] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[259] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[265] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[271] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Puno"
[277] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[283] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[289] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[295] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[301] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[307] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[313] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[319] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[325] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[331] "Puno"    "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[337] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[343] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[349] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[355] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[361] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[367] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[373] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Malo"
[379] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[385] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
[391] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Srednje"
[397] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[403] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[409] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[415] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[421] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[427] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[433] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[439] "Srednje" "Srednje" "Srednje" "Puno"    "Puno"    "Puno"
[445] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[451] "Puno"    "Puno"    "Puno"    "Puno"    "Puno"    "Puno"
[457] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
[463] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
```

```
 [469] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [475] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [481] "Srednje" "Srednje" "Srednje" "Srednje" "Malo"    "Malo"
 [487] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [493] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [499] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [505] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [511] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [517] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [523] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [529] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [535] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [541] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [547] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [553] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [559] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [565] "Malo"    "Malo"    "Malo"    "Malo"    "Malo"    "Malo"
 [571] "Malo"    "Malo"    "Malo"    "Srednje" "Srednje" "Srednje"
 [577] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [583] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [589] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [595] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [601] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [607] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [613] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [619] "Srednje" "Srednje" "Srednje" "Srednje" "Srednje" "Srednje"
 [625] "Srednje"
```

Pogledajmo sada niz vrijednosti stope zaraze za Hrvatsku:

```
> inf_rate_hr
  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [32] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1
 [63] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 [94] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[125] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
[156] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[187] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[218] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[249] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3
[280] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[311] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2
[342] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[373] 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
[404] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[435] 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2
[466] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1
[497] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
[528] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[559] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[590] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[621] 2 2 2 2 2
```

Možemo primijetiti da svako stanje emitira vrijednost koju ima najveću vjerojatnost emitirati, a to je za očekivati jer Viterbijev algoritam traži put koji je najvjerojatniji da se dogodi. Analogne rezultate dobivamo za sve druge države.

Ovime smo pokazali da smo diskretnu varijablu stope zaraze uspješno modelirali u skriven Markovljev model te da bi se moglo još detaljnije istraživati u ovome smjeru dok se ne pronađe nešto neočekivano. Skriven Markovljev model samo je jedan od diskretnih vrsta modeliranja te bi se pojavom novih informacija i podataka o koronavirusu mogle koristiti druge metode diskretnih modeliranja.

Poglavlje 5

Zaključak

Detaljnom analizom podataka ustvrđeno je da u periodu 1.5.2020.-30.6.2021. korelacije varijabli toka zaraze postoje između svih država i značajne su. Posebno, za zemlje koje su važnije za Hrvatsku - Italija, Austrija, Njemačka i Slovenija, korelacije između njih su vrlo značajne i očekivane. Ostale države također imaju velike korelacije s Hrvatskom, ali Ujedinjeno Kraljevstvo ima najmanji koeficijent korelacije. Sjeverozapadne države međusobno imaju jaki utjecaj jedna na drugu, ali i na države na jugoistoku Europe. Ujedinjeno Kraljevstvo ima najmanje koeficijente korelacije od ostalih država, a mogući razlozi za to su nedavan izlazak iz Europske unije te drukčiji pristup suzbijanju pandemije.

Primijetili smo vremenske razlike toka zaraze između Hrvatske i ostalih zemalja od kojih su najmanje s državama geografski i ekonomski najbliže Hrvatskoj, a najveće s državama na sjeverozapadu Europe.

U periodu 1.7.2021.-13.10.2021. događaju se promjene. Za obje varijable nestaju korelacije Italije i Hrvatske, Italije i Nizozemske, Danske i Njemačke, Danske i Ujedinjenog Kraljevstva. Dodatno, za varijablu dnevnog tjednog prosjeka zaraze nestaju korelacije između Italije i Ujedinjenog Kraljevstva te Danske i Belgije, dok za varijablu stope zaraze nestaju korelacije između Italije i Austrije, Italije i Slovenije, Ujedinjenog Kraljevstva i Austrije, Nizozemske i Austrije, Ujedinjenog Kraljevstva i Njemačke, Belgije te Hrvatske. Primijetimo da je Ujedinjeno Kraljevstvo koje je imalo najmanje koeficijente korelacije u prvom periodu izgubilo korelacije s čak šest država, Italija s četiri, Austrija s tri, Hrvatska, Slovenija, Njemačka, Danska i Nizozemska s dvije, a Belgija s jednom državom.

Također je vrijedno spomena da Hrvatska i dalje ima korelaciju s Austrijom, Njemačkom i Slovenijom koje do tada nisu imale procijepljenost veću od 70%, dok se korelacija s Italijom gubi, a ona je do tada imala procijepljenost veću od 70%.

Nemamo još dovoljno podataka da bismo znali koji je pravi razlog zbog kojeg se ovakve značajne promjene događaju nakon 30.6.2021., no kombinacija veće procijepljenosti u nekim državama, korištenje COVID potvrda te ispravno provođenje mjera mogli su utjecati

na prestanak utjecaja nekih država na druge.

Dodatak A

Dijelovi koda korišteni za analizu podataka

Viterbijev algoritam implementiran u R-u

```
Viterbi <- function(pie, trans_matrix, emission_matrix, data) {
  l <- length(data)
  delta <- numeric(5)

  for(i in 1:5)
    delta[i] <- pie[i] * emission_matrix[i, data[1] + 1]

  logdelta <- matrix(numeric(5*l), l, 5)
  logdelta[1, ] <- log(delta)

  psi <- matrix(numeric(5*l), l, 5)
  psi[1, ] <- c(0, 0, 0, 0, 0)

  for(t in 2:l) {
    for(j in 1:5) {
      pomocni <- numeric(5)

      for(i in 1:5) {
        pomocni[i] <- log(trans_matrix[i, j]) + logdelta[(t - 1), i]
      }
      logdelta[t, j] <- max(pomocni) + log(emission_matrix[j, (data[t]
 + 1)])
      psi[t, j] <- which.max(pomocni)
    }
  }

  Q <- numeric(l)
  Q[l] <- which.max(logdelta[l, ])
```

```
28
29    for(t in 1:(l - 1)) {
30      pom <- numeric(5)
31
32      for(i in 1:5) {
33        pom[i] <- logdelta[(l - t + 1), i] + log(trans_matrix[i, Q[l - t
    + 1]])
34      }
35      Q[l - t] <- which.max(pom)
36    }
37    return(Q)
38  }
```

Bibliografija

[1] C. Chapman, E. McDonnell Feit, *R for Marketing Research and Analytics*, Springer, 2015.

[2] Johns Hopkins University Center for Systems Science i Engineering, *Time series covid-19 - confirmed global*, 2021, https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv.

[3] H. Šikić, *Mjera i integral*, 2012, https://web.math.pmf.unizg.hr/nastava/mii/files/mii-predavanja-sikic.pdf.

[4] M. Huzak, *Matematička statistika*, https://web.math.pmf.unizg.hr/nastava/ms/index.php?sadrzaj=predavanja.php.

[5] A. Mimica, M. Ninčević, *Statistika*, 2010, https://web.math.pmf.unizg.hr/nastava/stat/files/vjezbe_novo.pdf.

[6] P. Lazić, *Statistički praktikum 2*, https://web.math.pmf.unizg.hr/nastava/statpr2/materijali.html.

[7] A. Churbanov, S. Winters-Hilt, *Implementing EM and Viterbi algorithms for Hidden Markov Model in linear memory*, (2008), https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-224#.

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

[9] Z. Vondraček, *Markovljevi lanci*, 2012/13, https://web.math.pmf.unizg.hr/~vondra/ml12-predavanja.html.

[10] I.L. MacDonald, W. Zucchini, *Hidden Markov Models for Time Series: An Introduction Using R*, Chapman & Hall/CRC, 2009.

Sažetak

S obzirom na trenutnu situaciju, koronavirus trenutno je jedna od najčešćih tema novih istraživanja. U ovom radu ispitujemo i uspoređujemo dnevne stope zaraze susjednih zemalja i regija. Naime, razumno je očekivati da će stope zaraze zemalja ili regija koje su vrlo povezane utjecati jedna na drugu.

Divergencija između stopa zaraze raznih država opažena je nakon 30.6.2021. vjerojatno zbog različitih postotaka procijepljenosti. Zbog toga podaci su podijeljeni na dva dijela: prije i poslije 30.6.2021.

U prvom periodu dokazane su značajne korelacije između svih odabranih država. U drugom periodu, značajna korelacija je nestala između nekih država.

S koreliranim državama napravljeni su modeli linearne regresije te je utvrđena njihova zadovoljavajuća procjena.

Da bi se isprobao drugačiji pristup podacima, diskretizirani podaci su na kraju pretvoreni u skrivljen Markovljev model.

Summary

Considering the current crisis, COVID-19 related topics are at the forefront of research. In this work, we are concerned with daily infection rates in neighbouring countries and regions. Namely, it stands to reason that infection rates in highly connected countries or regions will be related.

However, divergence has been observed for the period after 30.6.2021., probably due to varying vaccination rates. For that reason, data was separated into two parts: before and after 30.6.2021.

In the first period, significant correlation between all the countries was obtained. In the second period, significant correlation did not exist for some countries.

Using the correlated countries, linear regression models were made and tested.

To try a different approach to data analysis, discrete data was modeled into a hidden Markov model.

Životopis

Zovem se Nikolina Šarić i rođena sam 6. srpnja 1995. godine u Zagrebu. Oduvijek me zanimala matematika i bila sam sposobna riješiti razne matematičke probleme. Završila sam VII. gimnaziju u Zagrebu 2014. godine te iste upisala Prirodoslovno-matematički fakultet u Zagrebu. Pet godina nakon, 2019. godine stekla sam titulu prvostupnice matematike (*bacc. math.*) nakon koje sam upisala diplomski sveučilišni studij Matematička statistika.