Prostorno-vremenski modeli u geostatici

Marušić, Laura

Sveučilište J.J.Strossmayera u Osijeku
Fakultet primijenjene matematike i informatike
Sveučilišni diplomski studij matematike
Smjer: Financijska matematika i statistika

Laura Marušić

Prostorno-vremenski modeli u geostatistici

Diplomski rad

Osijek, 2023.

Sveučilište J.J.Strossmayera u Osijeku
Fakultet primijenjene matematike i informatike
Sveučilišni diplomski studij matematike
Smjer: Financijska matematika i statistika

Laura Marušić

Prostorno-vremenski modeli u geostatistici

Diplomski rad

Mentor: izv. prof. dr. sc. Danijel Grahovac

Osijek, 2023.

Sadržaj

1 Uvod

Znanstvenici u mnogim primijenjenim područjima, kao što su biologija, ekologija i ekonomija, sve više stvaraju velike baze podataka koje su često i geografski i vremenski obilježene. Takvi podaci nazivaju se prostorno-vremenskim podacima jer ne variraju samo u vremenu, nego i u prostoru, po čemu su i dobili ime. Primjeri takvih podataka su širenje bolesti, praćenje onečišćenja zraka, praćenje ekonomskih pokazatelja, npr. cijene kuća, razine siromaštva i tako dalje.

U ovom radu ćemo se upoznati s prostorno-vremenskim podacima, predstaviti problematiku modeliranja prostorno-vremenskih podataka, a posebno tzv. *„point referenced"* podataka, i upoznati se s osnovnim pojmovima pri modeliranju u geostatistici[1]. Modeliranje je zapravo metoda opisivanja ponašanja promatranih podataka u vremenu i prostoru, a to je sastavni dio statističke analize podataka. Za takve skupove podataka interes često leži u otkrivanju i analiziranju prostornih uzoraka i vremenskih trendova, predviđajući varijablu koja je ovisna i u prostoru i u vremenu. Pod predviđanje podrazumijevamo procjenu nepoznatih parametara ili predikciju nepoznate slučajne ovisne varijable. Nekad se pod predviđanje podrazumijeva i predviđanje ovisne varijable u budućem vremenu, ali to nećemo obuhvatiti u ovom radu.

Za početak ćemo se upoznati s prostorno-vremenskim podacima, navest ćemo razlike među tipovima baza prostorno-vremenskih podataka te spomenuti kojim klasama objekata pripadaju kako bi ih mogli modelirati uz pomoć programskog jezika R. Nakon toga slijedi vizualizacija, odnosno različiti načini vizualnih prikazivanja prostorno-vremenskih podataka te modeliranje podataka prostorno-vremenskim procesima.

2 Prostorno-vremenski podaci

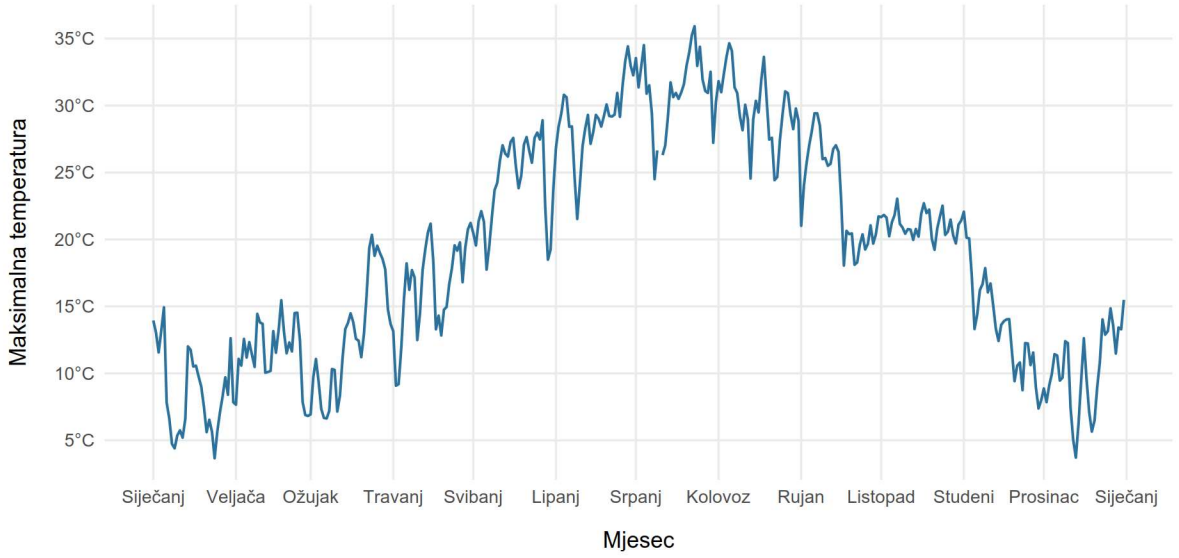
Kao što i sam naziv govori, radi se o opažanjima koja se bilježe na određenim mjestima u prostoru i trenucima u vremenu. Vrlo često su se prostorno-vremenski podaci analizirali na način da se posebno analizira vremenska, a posebno prostorna komponenta podataka. Dok se nisu razvile naprednije tehnike koje obje komponente prostorno-vremenskih podataka mogu analiziriati, vizualizirati i modelirati zajedno, tim podacima se moralo pristupati na način da se posebno proučavaju analizom vremenskih nizova, a potom tzv. prostornom statistikom ili obrnutim redoslijedom ovisno o preferenciji znanstvenika.

Ključno je razumijevanje strukture prostorno-vremenskih podataka. Kako bi došli do njihove strukture, krećemo od strukture vremenskih nizova pa ćemo se osvrnuti na strukturu prostornih podataka te na kraju dobiti prostorno-vremenski niz podataka.

Analitičari vremenskih nizova rade s bazama podataka koje se sastoje od samo jedne komponente, a to je vrijeme (npr. datum, mjesec, godina, sat i sl.) i od jedne ili više opisnih varijabli (svojstava) koje ovise o toj vremenskoj komponenti, odnosno bave se analiziranjem i modeliranjem jednodimenzionalnih ili višedimenzionalnih (multivarijatnih) vremenskih nizova. Jednodimenzionalni niz od n podataka $\{x_t,\, t \in T_0\}$ smatraju realizacijom (ili dijelom realizacije) slučajnog procesa $\{X_t,\, t \in \mathbb{R}\}$, $T_0 \subseteq \mathbb{R}$, a višedimenzionalni niz od n podataka $\{\boldsymbol{x}_t,\, t \in T_0\}$ realizacijom vektorskog slučajnog procesa $\{\boldsymbol{X}_t,\, t \in \mathbb{R}\}$, $T_0 \subseteq \mathbb{R}$ promatranog u određenom vremenskom intervalu, pri čemu se slučajni proces može definirati u neprekidnom ili diskretnom vremenu. Skup $T_0 = \{t_1, t_2, \ldots, t_n\},\ t_1 < t_2 < \cdots < t_n$ zovemo skup

[1] Geostatistika je grana statistike koja se koristi za analizu i procjenu vrijednosti koje imaju prostorne i/ili prostorno-vremenske ovisnosti. Primijenjena je znanost u meteorologiji, oceanografiji, ekologiji, agrikulturi, geografiji, šumarstvu i raznim drugim granama.

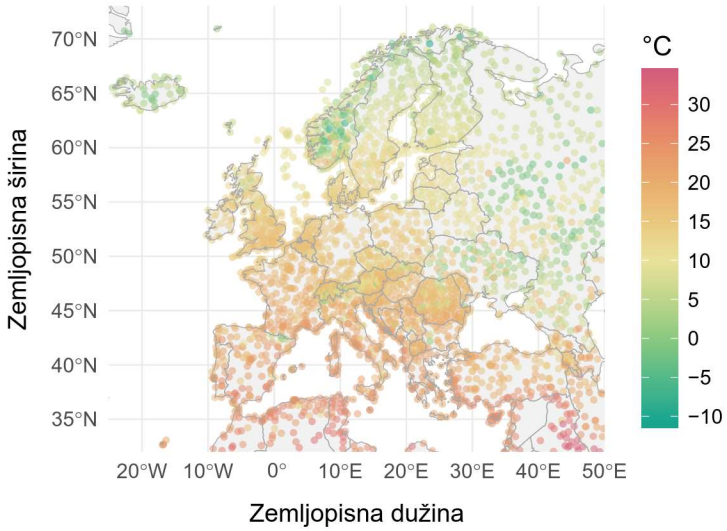
vremenskih trenutaka koji je najčešće konačan i ima ekvidistantno raspoređene vremenske trenutke, odnosno $t_2 - t_1 = t_3 - t_2 = \cdots = t_n - t_{n-1}$ pa je u tom slučaju $T_0 = \{1, 2, \ldots, n\}$ i govorimo o nizu podataka u diskretnom vremenu. Primjer niza podataka u diskretnom vremenu može se vidjeti na slici 2.1, a radi se o uprosiječenim dnevnim maksimalnim temperaturama u Hrvatskoj od siječnja do prosinca 2022. godine iz tzv. baze podataka GSOD, koju ćemo kasnije opisati detaljnije.



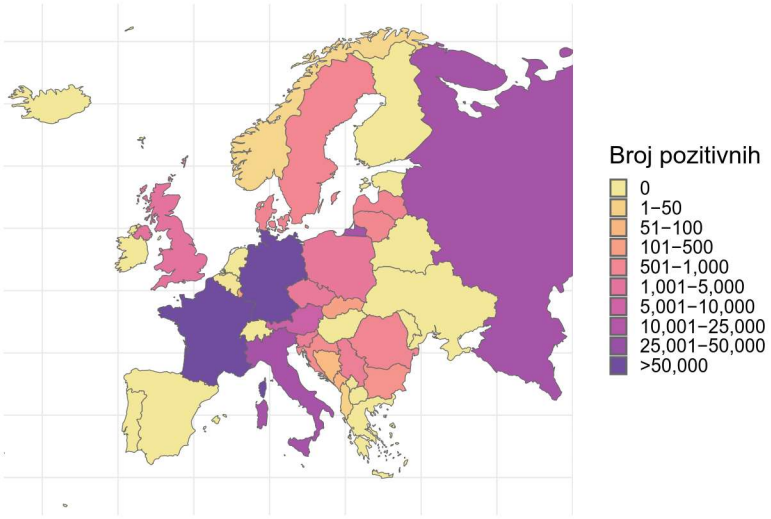
Slika 2.1: Primjer vremenskog niza prosječnih dnevnih maksimalnih temperatura (u °C), iz baze podataka GSOD, u Hrvatskoj od siječnja do prosinca 2022. godine.

Ako niz podataka ovisi i o prostoru u kojemu se podaci bilježe, onda se niz podataka sastoji i od prostorne komponente te ukoliko se vremenska komponenta ne uzima u obzir, tada govorimo o modeliranju prostornih podataka. Statističari koji se bave modeliranjem prostornih podataka mogu višedimenzionalni niz podataka $\{\boldsymbol{x}(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{D}_0\}$ promatrati kao vremensko agregiranje (npr. sumiranje $x^{(i)}(\boldsymbol{s}) = \sum_t x_t^{(i)}(\boldsymbol{s})$, $\boldsymbol{s} \in \mathbb{D}_0 \subseteq \mathbb{D}$, $\forall i = 1, \ldots, m$ po vremenskim oznakama $t \in \mathbb{R}$ tako da nemamo više vremenski niz), a mogu ga promatrati i kao vremenski smrznuta stanja, *„snapshots"*, odnosno opservacije po različitim promatranim prostorima u fiksnom vremenu $t \in \mathbb{R}$, $x^{(i)}(\boldsymbol{s}) = x_t^{(i)}(\boldsymbol{s})$, $\boldsymbol{s} \in \mathbb{D}_0 \subseteq \mathbb{D}$, $\forall i = 1, \ldots, m$. Bio niz podataka promatran kao agregiranje po vremenskoj komponenti ili kao *„snapshots"* u određenom vremenskom trenutku, tako promatran niz podataka statističari smatraju vektorskim slučajnim procesom $\{\boldsymbol{X}(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{D}\}$ u prostoru $\mathbb{D} \subseteq \mathbb{R}^l$, $l \in \mathbb{N}$. Analogno, jednodimenzionalni niz podataka $\{x(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{D}_0\}$ se smatra realizacijom (ili dijelom realizacije) slučajnog procesa $\{X(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{D}\}$ u prostoru $\mathbb{D} \subseteq \mathbb{R}^l$, $l \in \mathbb{N}$. Prostorna komponenta može biti jednodimenzionalna ($l = 1$), npr. grad, zemlja i sl., dvodimenzionalna ($l = 2$), npr. geografska dužina i širina, ili trodimenzionalna ($l = 3$), npr. geografska dužina, širina i visina (prostor je zapravo trodimenzionalan). Prostorni podaci se smatraju slučajnim događajima u prostoru neovisno o tome kako se promatraju. Na slici 2.2 je dan primjer prostornih podataka prosječnih maksimalnih temperatura dobivenih kao agregiranje po lokacijama na kojima su postavljeni termometri u Europi. Slika 2.3 prikazuje primjer prostornih podataka ukupnog broja novih slučajeva COVID-19 [2] kao *„snapshot"* u vremenu (30. rujna 2022.).

[2]Baza podataka sadrži dnevne podatke po zemljama svijeta i može se preuzeti na `https://covid19.who.int/data`



Slika 2.2: Primjer prostornih podataka prosječnih maksimalnih temperatura (u °C), iz baze podataka GSOD, u 2022. godini, po lokacijama na kojima su postavljene meteorološke stanice u Europi.



Slika 2.3: Primjer prostornih podataka prikazan kao *„snapshot"* ukupnog broja novih slučajeva oboljelih od COVID-19 po europskim zemljama dana 30. rujna 2022.

Ukoliko se niz podataka, kojemu opisne varijable ovise i o vremenskoj i o prostornoj komponenti u kojima se podaci bilježe te se pri modeliranju i analizi želi uzeti u obzir obje komponente istovremeno tada se prostorno-vremenski podaci mogu promatrati na dva načina, kao vremenski niz prostorno slučajnih procesa (dinamički pristup) ili kao prostorno slučajni proces vremenskih nizova (deskriptivni pristup). Prostorno-vremenski niz, kao i vremenski niz ili slučajni proces u prostoru, može biti višedimenzionalan (multivarijatan), ovisno o broju opisnih varijabli među podacima koje želimo razmatrati u modeliranju, a da pritom sve varijable ovise i o vremenu i o prostoru.

Neka je $\mathbb{D} \subseteq \mathbb{R}^l$, $l \in \mathbb{N}$ skup prostornih oznaka koji se promatra u modelima, a $t \in \mathbb{R}$ vremenska oznaka. Ako želimo predviđati buduće vrijednosti, onda koristimo dinamički pristup pri modeliranju pa prostorno-vremenske podatke $\{x_t(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{D}_0,\ t \in T_0\}$ opisujemo slučajnim procesom $\{X_t(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{D},\ t \in \mathbb{R}\}$, pri čemu je $T_0 \subseteq \mathbb{R}$ podskup skupa promatranih vremenskih oznaka, a $\mathbb{D}_0 \subseteq \mathbb{D}$ podskup skupa prostornih oznaka. Želimo li odrediti procjenu neke varijable u određenom trenutku u prostoru, modeliranju ćemo pristupiti deskriptivno i u tom slučaju ćemo prostorno-vremenske podatke zapisane u obliku vektora

$$\begin{aligned}\boldsymbol{x} = \big[&x(\boldsymbol{s}_1;t_1),\ x(\boldsymbol{s}_2;t_1),\ \ldots,\ x(\boldsymbol{s}_m;t_1),\\ &x(\boldsymbol{s}_1;t_2),\ x(\boldsymbol{s}_2;t_2),\ \ldots,\ x(\boldsymbol{s}_m;t_2),\\ &\ldots,\\ &x(\boldsymbol{s}_1;t_n),\ x(\boldsymbol{s}_2;t_n),\ \ldots,\ x(\boldsymbol{s}_m;t_n)\big]^\mathsf{T}\end{aligned}$$

opisati slučajnim procesom $\{X(\boldsymbol{s};t),\ \boldsymbol{s} \in \mathbb{D},\ t \in \mathbb{R}\}$. Skup indeksa u ovakvom slučajnom procesu je skup od $m \cdot n$ uređenih parova ili trojki s prostornom i vremenskom oznakom, a značavamo ga s $\mathbb{D} \times \mathbb{R}$. Skup vremenskih oznaka može biti neprekidan, primjerice $[0, \infty\rangle$, ili diskretan, npr. $\mathbb{N}_0$, a skup prostornih oznaka $\mathbb{D}$ je najčešće jednodimenzionalan ili dvodimenzionalan $\mathbb{D} \subseteq \mathbb{R}^l$, $l \in \{1, 2\}$ ovisno kako su podaci definirani u bazi podataka.

Zbog opširnosti teme, u radu ćemo se bazirati na deskriptivni pristup i raditi na podacima u diskretnom vremenu $t \in \{0, 1, 2, ...\} \subseteq \mathbb{N}_0$ s jednodimenzionalnim ili dvodimenzionalnim prostornim oznakama te ćemo razmatrati samo jednu opisnu varijablu u modelima. Stoga ćemo modelirati jednodimenzionalne prostorno-vremenske nizove slučajnim procesima u diskretnom vremenu. Pogledati [4, str. 205] za više o dinamičkom pristupu modeliranja koje su opisali Cressie, Wikle i Mangion te pogledati [3, poglavlje 7.4] za više o multivarijatnim modelima koje su opisali Cressie i Wikle.

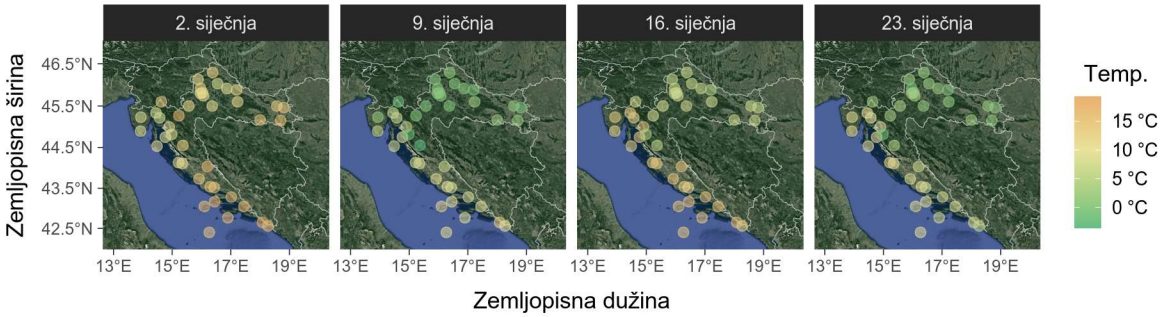
2.1 Tipovi baza prostorno-vremenskih podataka

Pretpostavimo da su dani prostorno-vremenski podaci u diskretnom vremenu, odnosno podaci ovise o dvjema komponentama, vremenskoj i prostornoj, te su dodijeljeni jednaki vremenski i prostorni razmaci među podacima. Pretpostavimo da u podacima ima ukupno T pravilno raspoređenih vremenskih točaka. U tom slučaju proizvoljni podatak ima vremensku oznaku t. Budući da se radi o prostorno-vremenskim podacima, podatak ima i zabilježenu lokaciju koju označavamo sa $\boldsymbol{s}$ u nekom području $\mathbb{D}$. Tada niz podataka (ili opservacija) označavamo s $\{x(\boldsymbol{s};t),\ \boldsymbol{s} \in \mathbb{D}_0,\ t \in \{t_1, t_2, \ldots, t_T\}\}$, $t_1 < t_2 < \cdots < t_T$, $T \in \mathbb{N}$, kao realizaciju slučajnog procesa $\{X(\boldsymbol{s};t),\ \boldsymbol{s} \in \mathbb{D},\ t \in \mathbb{N}\}$, $\mathbb{D}_0 \subseteq \mathbb{D} \subseteq \mathbb{R}^l$, $l \in \{1, 2\}$.

S obzirom na način na koji se u podacima promatraju lokacijske oznake $\boldsymbol{s}$ u području $\mathbb{D}$ nastaju različiti tipovi prostorno-vremenskih podataka, a tri se tipa mogu izdvojiti kao glavni, pri čemu se vremenske oznake promatraju na isti način u svakom tipu:

- *„point referenced" podaci* (geostatistički podaci) - prostorna i vremenska komponenta nisu slučajne varijable u podacima, odnosno znamo gdje i kada će se zabilježiti neka opservacija. Primjerice, znamo na kojoj lokaciji su fiksno postavljeni radari i znamo da radari bilježe određena obilježja (meteorološka, atmosferska, oceanska ...) svakih sat vremena ili jednom dnevno koja se unose u bazu podataka. Obilježja tih podataka pripadaju slučajnim varijablama koje poprimaju slučajne vrijednosti. Lokacije radara su najčešće zabilježene parovima geografske širine i dužine. Za primjer ovog tipa podataka možemo uzeti ovisnu varijablu razine onečišćenja zraka koja se promatra na određenom mjestu gdje je postavljen radar unutar promatranog područja svih radara.

Jedan podatak (opservacija) na lokaciji $\boldsymbol{s}$ varira neprekidno u promatranom području $\mathbb{D}$. Ako promatramo opisnu varijablu $x(\boldsymbol{s}, t)$ na m različitih lokacija koje označavamo sa $\boldsymbol{s}_i$, $i = 1, \ldots, m$ i u T različitih vremenskih točaka označenih s t_j, $j = 1, \ldots, T$. Skup prostornih lokacija mogu biti ili fiksni radari, kao u primjeru onečišćenja zraka, ili mogu varirati s vremenom, na primjer, podaci dobiveni od istraživačkog broda koji mjeri karakteristike oceana dok se kreće u oceanu. Još jedan primjer ovakvog tipa podataka smo spomenuli ranije i prikazali vizualno na slici 2.1, radilo se o prosječnim maksimalnim temperaturama u Hrvatskoj, a na sljedećoj slici 2.4 se može vidjeti i gdje su postavljeni radari za mjerenje temperature po Hrvatskoj i vidimo kako se temperature kreću kroz četiri nedjelje u siječnju 2022. godine.

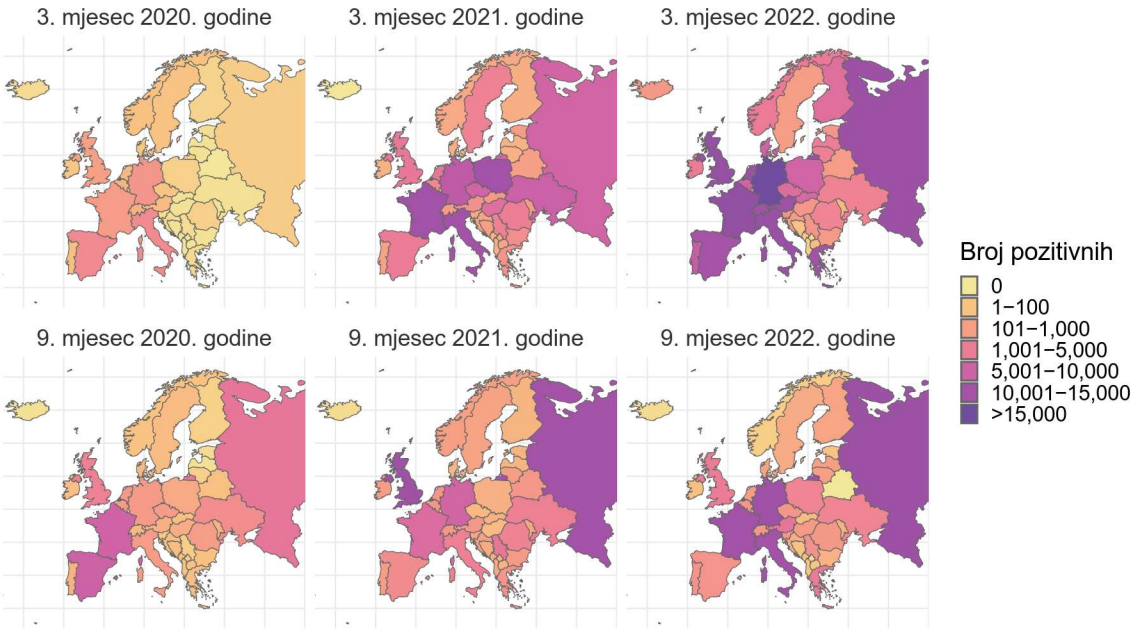


Slika 2.4: Primjer *„point referenced"* prostorno-vremenskih podataka dnevnih maksimalnih temperatura u Hrvatskoj kroz četiri nedjelje u siječnju 2022.

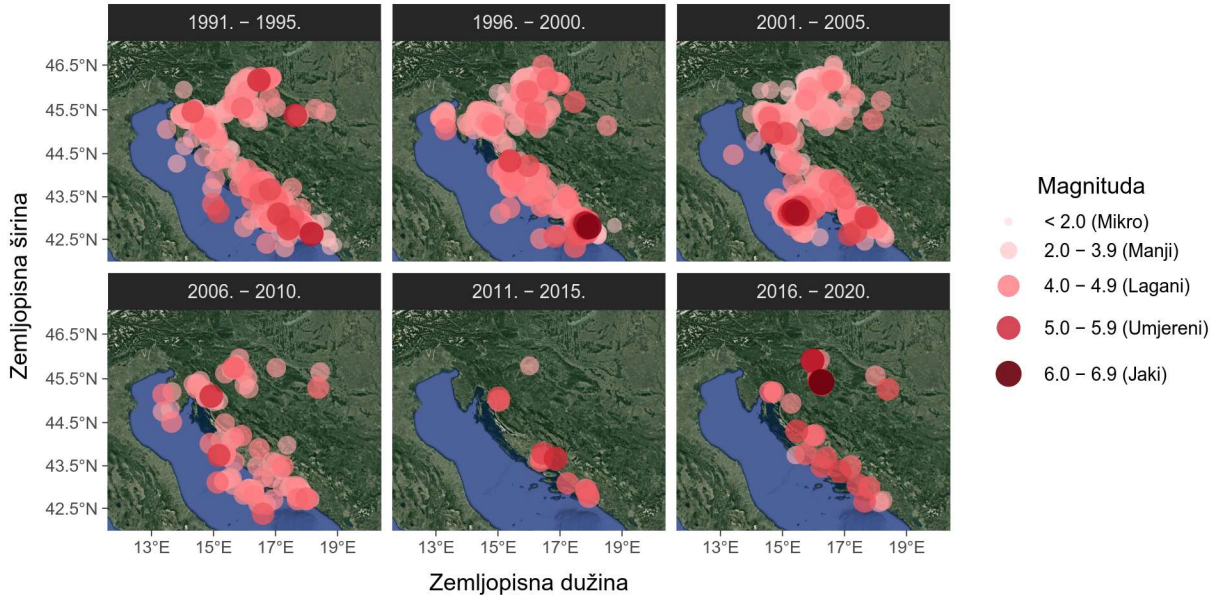
- *„areal unit" podaci* (skup lokacija agregiranih vrijednosti) - često se promatraju u mnogokutima s definiranim granicama ovisno što se želi promatrati. Granice mogu biti proizvoljne (npr. promatrano geografsko područje može biti podijeljeno po jednakim šesterokutima) i administrativne granice (npr. promatrano geografsko područje može biti podijeljeno po granicama kontinenata, zemalja ili županija). Vrijednosti podataka često su agregiranja unutar definiranih geografskih područja. Na primjer, prosječni mjesečni broj potvrđenih slučajeva oboljelih od COVID-19 po zemljama u Europi svakih 6 mjeseci u posljednje 3 godine (slika 2.5).

- *„point pattern" podaci* (uzorak slučajnih lokacija nekih događaja) - tzv. točkasti uzorak podataka se pojavljuje kada promatrani događaj, npr. izbijanje bolesti, dogodi na slučajnim mjestima. Stoga je u ovom tipu podataka lokacija slučajna varijabla i po tome se ovaj tip podataka razlikuje od geostatističkih podataka („point referenced" - referirane točkama). U ovom tipu podataka su također lokacije referirane točkama, parovima zemljopisne širine i dužine. Podaci se modeliraju točkovnim procesima, poput Poissonovog procesa, što su zapravo stohastički modeli, u kojima se procjenjuje vjerojatnost pojave promatranog događaja na nekom mjestu u nekom vremenskom trenutku. Primjeri ovih podataka mogu se naći u šumarstvu, proučavanju potresa i padalina, astronomiji i epidemiologiji. Vizualni prikaz primjera ovakvog tipa podataka možemo vidjeti na slici 2.6. Prikazani su potresi[3] određenih magnituda Richterove ljestvice u Hrvatskoj svakih 5 godina.

[3]Baza podataka sadrži informacije o potresima po zemljama svijeta i može se preuzeti na `https://earthquake.usgs.gov/earthquakes/search/`. Postoji velika mogućnost filtriranja podataka i opcija prikaza potresa prilikom preuzimanja.



Slika 2.5: Primjer *„areal"* prostorno-vremenskih podataka prosječnih mjesečnih novooboljelih od COVID-19 po europskim zemljama svakih 6 mjeseci u posljednje 3 godine.



Slika 2.6: Primjer *„point pattern"* prostorno-vremenskih podataka potresa u Hrvatskoj svakih 5 godina.

Za više o tipovima podataka pogledati [11, poglavlje 1] koje je opisao Sujit K. Sahu i naveo nekoliko primjera.

2.2 Baze podataka

Upoznajmo se ukratno s bazama podataka korištenima u ovom radu. Neke od njih smo već i spominjali.

Dnevni sažetak globalne površine (GSOD)

Baza podataka sadrži globalne podatke dobivene iz klimatološkog centra USAF-a (američkog ratnog zrakoplovstva). Najnoviji dnevni sažeti podaci obično su dostupni 1-2 dana nakon datuma i vremena opažanja korištenih u dnevnim sažetcima. Datoteke s podacima na mreži počinju s 1929. godinom. Obično su dostupni podaci preko 9000 postaja. Baza se sastoji od 47 varijabli od kojih su neke podaci o stanicama (id stanice, geografska širina i dužina, naziv i ISO kod države u kojoj je stanica), podaci o trenutku zapisa podatka (datum, dan, mjesec, godina), datum početka i kraja mjerenja stanice. Opisne varijable koje se nalaze u bazi su:

- Srednja temperatura (Fahrenheit)
- Srednja točka rosišta (Fahrenheit)
- Srednji tlak na razini mora (milibar)
- Srednji tlak na postaji (milibar)
- Srednja vidljivost (milja)
- Srednja brzina vjetra (čvor)
- Maksimalna trajna brzina vjetra (čvor)
- Maksimalni udar vjetra (čvor)
- Maksimalna temperatura (Fahrenheit)
- Minimalna temperatura (Fahrenheit)
- Količina oborine (inč)
- Dubina snijega (inč)

Podaci se mogu direktno preuzeti R paketom ***GSODR*** jednostavnim pozivanjem funkcije `get_GSOD()`, pri čemu su sve mjerne jedinice pretvorene u međunarodni sustav jedinica (SI), npr. Fahrenheiti u Celzijuse i inči u milimetre.

Za naše potrebe u ovom radu smo preuzeli podatke od 2022. godine i filtrirali smo ih prema području Europe i Hrvatske. Uzeli smo u razmatranje varijablu *MAX*, dnevne maksimalne temperature. U tom slučaju imamo 1,049,022 opservacija i 7 varijabli. Ove podatke možemo svrstati u diskretne i nepravilno raspoređene u vremenu (nemamo zabilježene podatke svaki dan za svaku stanicu), a geostatističke i nepravilno raspoređene u prostoru (nemamo stanice raspoređene na jednako udaljenim koordinatama) pa time ovi podaci spadaju u iregularnu prostorno-vremensku podatkovnu strukturu (**STI**[4]). Neke od tih podataka smo mogli vidjeti na slikama 2.1, 2.2 i 2.4.

[4]engl. „*spatio-temporal irregular*" (STI) klasa je jedna od četiri klase prostorno-vremenskih podataka, Pebesma ih je svakako uzeo u obzir pri razvoju R paketa ***spacetime*** [8] jer je kod svake metode u analizi potrebno prilagoditi dane podatke.

E-OBS dnevni meteorološki podaci u Europi (E-OBS)

Skup dnevnih mrežastih opservacija za oborine, temperaturu, tlak na razini mora, globalno zračenje i brzinu vjetra u Europi nazvan E-OBS. Skup podataka je raspoređen na regularnoj mreži po 0.1 stupanj i sadrži opisne varijable srednje, minimalne i maksimalne dnevne temperature, dnevne količine oborina, dnevne prosječne tlakove u razini mora, dnevnu prosječnu relativnu vlažnost, srednje dnevne brzine vjetra i srednje dnevne globalne radijacije.

COVID-19 u Europi

Baza se sastoji od dnevnih slučajeva oboljelih od COVID-19 po državama. Sadrži datum, naziv i ISO kod države, WHO regiju, a od opisnih varijabli sadrži broj novih i broj kumulativnih potvrđenih slučajeva oboljelih, broj novih i broj kumulativnih potvrđenih smrti oboljelih. Podaci preuzeti s `https://covid19.who.int/data`.

Ovaj rad će se bazirati na analizi i modeliranju geostatističkog tipa podataka. Opisivat ćemo ih procesima koji su diskretni u vremenu i geostatistički u prostoru. Ali prije nego što krenemo u samu analizu podataka, reći ćemo nešto ukratko o R paketima koje ćemo koristiti pri analizi, vizualizaciji i modeliranju podataka i kratko se upoznati s bazama podataka korištenima u ovom radu. Neki od R paketa su:

- ***dplyr*** - paket koji pomaže pri tranformaciji i manipuliranju podacima poput filtriranja i agregiranja.
- ***ggforce*** - cilj ovog paketa je nadopuna nedostataka *ggplot2*-a, kako bi detaljnije napravili vizualizaciju podataka, npr. dodavanje anotacija na grafu.
- ***ggmap*** - alat koji omogućuje vizualizaciju kombiniranjem prostornih informacija statičkih karata iz Google Maps-a, OpenStreetMap-a, Stamen Maps-a ili CloudMade Maps-a sa grafičkom implementacijom *ggplot2*. Uz to, paket sadrži nekoliko funkcija koje korisniku omogućuju pristup API-jima, „Google Geocoding[5]", „Distance Matrix" i „Directions".
- ***ggplot2*** - paket koji pruža korisne naredbe za stvaranje složenih grafova iz podataka. Može doista poboljšati kvalitetu i estetiku grafike i učinkovitije je za korištenje pri izradi grafike. S razlogom ga zovu „gramatikom grafike".
- ***gifski*** - alat za izradu animacije u obliku GIF datoteke iz niza PNG datoteka.
- ***gridExtra*** - pruža niz funkcija za rad s „mrežom" više grafova, osobito za raspoređivanje višestrukih grafova odjednom (kao jednu sliku). Radi kompatibilno s *ggplot2*.
- ***gstat*** - sadrži metode za prostorno i prostorno-vremensko geostatističko modeliranje, predviđanje, simulaciju i modeliranje variograma.
- ***lubridate*** - olakšava rad s formatom datuma.
- ***raster*** - paket sadrži funkcije za kreiranje, čitanje, manipuliranje i pisanje rasterskih podataka[6].

[5]Potrebno je imati svoj Google račun i omogućen Billing (naplaćivanje usluge) za Google Cloud. Upute su na sljedećem linku `https://rpubs.com/jcraggy/841199`.

[6]Raster je prostorna struktura podataka koja dijeli regiju na pravokutnike koji se nazivaju „ćelije" (ili „pikseli") koji mogu pohraniti jednu ili više vrijednosti za svaku od tih ćelija.

- ***readxl*** - olakšava učitavanje podataka iz Microsoft Excela u R.
- ***rnaturalearth*** - paket za držanje i olakšavanje interakcije s podacima vektorskih karata Zemlje.
- ***scales*** - pruža metode za automatsko određivanje prijeloma i oznaka za osi i legende radi estetike. Radi kompatibilno s *ggplot2*.
- ***spacetime*** - paket koji sadrži klase podataka koje proširuju one koje se koriste za prostorne podatke iz paketa ***sp*** i podatke vremenskih nizova iz paketa ***xts***. Objekti iz paketa *spacetime* sadrže dodatne informacije, kao što su projekcija karte, vremenske zone i granice nekih regija.
- ***tidyverse*** - kolekcija R paketa dizajnirana za podatkovnu znanost. Pomaže pri uvozu podataka, pospremanju, manipulaciji i vizualizaciji podataka. Paket je besplatno dostupan za korištenje te se stalno mijenja i poboljšava. Sadrži pakete *ggplot2*, *dplyr*, *tidyr*.

3 Analiza podataka

U ovom poglavlju će se razmatrati prvi korak do izgradnje prostorno-vremenskih modela, a to su pristupi analize podataka metodama vizualizacije i određivanjem numeričkih karakteristika procesa. Cilj je koristiti ove analitičke metode za dobivanje nekog prvog dojma odnosa koji se mogu ugraditi u prostorno-vremenske modele u svrhu donošenja zaključaka, odnosno analiza podataka je uvod u modeliranje jer dolazimo do pitanja na koja želimo dobiti odgovore. Nakon što u podacima otkrijemo maksimalne i minimalne vrijednosti, trendove i sezonalnosti, što s njima? Kako možemo znati što je važno među njima? Modeli bi nam u ovom slučaju trebali pomoći dati odgovore na ova pitanja.

3.1 Vizualizacija

Metoda vizualizacije podataka je najvažnija i najnužnija komponenta analize podataka, uključuje korištenje karata, boja i animacija, što je vrlo moćan način za pružanje uvida u analizu podataka.

Pri analizi prostorno-vremenskih podataka dolazimo do raznih izazova zbog činjenice da se mora uzeti u obzir više od jedne komponente istovremeno (npr. jedna ili dvije prostorne i jedna vremenska komponenta), ali zato postoje alati koji mogu pomoći u vizualizaciji takvih podataka.

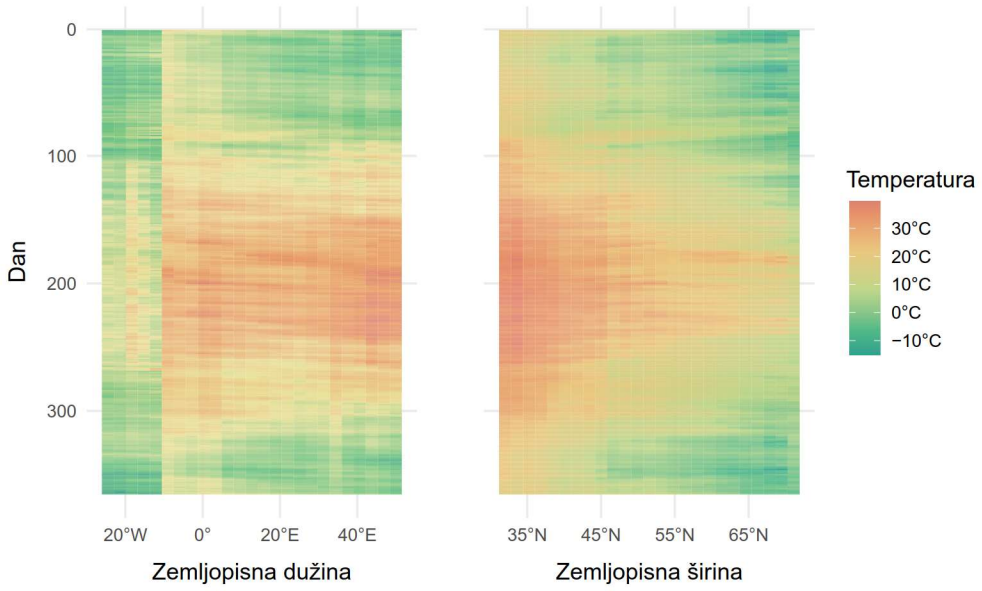
Neki od načina vizualizacije su višepanelni dijagrami, Hovmöllerov dijagrami, interaktivni grafovi i animacije.

Višepanelni dijagram[7] smo već vidjeli na primjeru *„point referenced"* podataka na slici 2.4, *„areal"* podataka na slici 2.5 i *„point pattern"* podataka na slici 2.6. To je slika koja sadrži više grafova, a svaki graf prikazuje agregirane podatke u nekom vremenskom trenutku, sastoji se od iste x i y osi (obje prostorne komponente, zempljopisna širina i dužina) te zajedničke legende vrijednosti podataka.

Hovmöllerov dijagram[8] je dvodimenzionalna prostorno-vremenska vizualizacija, gdje je prostor agregiran na jednu komponentu (zemljopisnu širinu ili dužinu) i nalazi se na x osi,

[7] engl. *„Multi-panel Plots"*
[8] engl. *„Hovmöller Plots"*

druga komponenta označava vrijeme i nalazi se na y osi. Prikazuju nam vremenski trend po jednoj prostornoj komponenti. Mogu sugerirati modeliranje prostorno-vremenskih podataka dinamčkim procesima. Prikazat ćemo Hovmöllerov dijagram na primjeru GSOD podataka, tj. dnevnih maksimalnih temperatura 2022. godine u Europi. Budući da nemamo vrijednosti temperatura na svakoj koordinati zemljopisne širine (imamo meteorološke radare samo na nekim mjestima), moramo interpolirati te vrijednosti u pravilnu mrežu zempljopisnih širina i vremenskih trenutaka kako bi podatke mogli prikazati na grafu te lakše odrediti vrmenski trend. Prvo napravimo skup uređenih parova, recimo od 25 prostornih točaka jednako udaljenih u rasponu od minimalne do maksimalne vrijednosti zemljopisne širine među danim podacima i 100 vremenskih točaka jednako udaljenih u rasponu od minimalne do maksimalne vrijednosti vremenskih trenutaka među danim podacima. Zatim podacima pridružujemo odgovarajuću zempljopisnu širinu iz napravljene pravilne mreže uređenih parova s najmanjom udaljenošću od stvarne koordinate zempljopisne širine radara do koordinate zempljopisne širine skupa uređenih podataka. Nakon toga temperature agregiramo po novo pridruženim koordinatama i vremenskim trenucima s prosječnim vrijednostima. Ideja je analogna za zemljopisnu dužinu. Sada ćemo na slici 3.1 vidjeti kako maksimalne temperature opadaju prema sjeveru Europe, a od zapada prema istoku su gotovo pa konstantne, što je i očekivano. Na grafu imamo samo 9 125 agregiranih podataka (uređeni parovi 25 koordinata i 365 dana), a u početku smo imali 1 049 022 podataka.



Slika 3.1: Primjer Hovmöllerova dijagrama na podacima GSOD dnevnih maksimalnih temperatura u 2022. godini od zapada prema istoku Europe (lijeva slika) i od juga prema sjeveru (desna slika).

Promjene varijabli u vremenu i prostoru možemo bolje dočarati animacijom nego pojedinim slikama u određenim trenucima s agregiranim vrijednostima. Za dobivanje animacije u R-u, mogu se koristiti razni paketi poput ***animation***, ***gganimate*** ili ***gifski***. Kod kreiranja animacija potrebno nam je imati pravilno raspoređene koordinatne osi prostornih podataka u svakom zabilježenom trenutku. Ako imamo problem s nepostojećim podacima ili nepravilno raspoređenim prostornim podacima, trebamo „nadopuniti" te podatke interpolacijom, i u vremenu i u prostoru kao što smo to napravili kod Hovmöllerovog dijagrama. Primjer pravilno raspoređenih podataka kojima nije potrebna interpolacija je E-OBS baza podataka

i na njoj je moguće napraviti animaciju temperatura. Korištenjem ***gifski*** R paketa kreirali smo animaciju dnevnih maksimalnih temperatura u Europi od 1. siječnja 2021. do 30. lipnja 2022., a na sljedećem linku se može vidjeti dobivena animacija u GIF datoteci „Tmax europe 21_22.gif". Animacija je dobivena sljedećim R kodom.

```
# Kreiranje niza 546 PNG datoteka (broj datoteka jednak broju dana)
files <- str_c("./slike/DAN",1:546, ".png")

# Petlja izrade grafika po svakom danu
for(i in 1:546){
  ggplot(txEurope_21_22_df_cat) +
    geom_sf(data = map, color = NA) +
    geom_tile(aes(lon, lat, fill = txEurope_21_22_df_cat[,i])) +
    coord_sf(xlim = c(-25,50), ylim = c(32,73), expand = FALSE) +
    scale_fill_manual(values = colours, drop = FALSE) +
    guides(fill = guide_colorsteps(barwidth = 20,
                                   barheight = 0.3,
                                   title.position = "right",
                                   title.vjust = 0.5)) +
    theme_light() +
    theme(legend.position = "top",
          legend.justification = .5,
          legend.text = element_text(size = 7),
          legend.title = element_text(size = 10),
          plot.subtitle = element_text(family = "sans", size = 13,
                                       margin = margin(b = 10, t = 5,
                                                       unit = "pt")),
          plot.margin = margin(6, 15, 0, 0)) +
    labels(x = "", y = "", fill = "ºC",
           subtitle = str_c("Datum: ",label[i]))
  # Spremanje grafike u PNG datoteke
  ggsave(files[i], width = 5, height = 4.8)
}

# Kreiranje gifa iz PNG datoteka
gifski(files, "Tmax europe 21_22.gif", width = 450,
       height = 470, loop = TRUE, delay = 0.15)
```

Ovim načinom se otkriva dinamika među prostorno-vremenskim podacima po svim komponentama odjednom što nije vidljivo korištenjem drugih metoda vizualizacije. Ova metoda također može sugerirati dinamički prostorno-vremenski model kao i Hovmöllerov dijagram.

3.2 Numeričke karakteristike prostorno-vremenskih procesa

Nakon što smo se upoznali s metodama vizualizacije, možemo prijeći na istraživanje numeričkih karakteristika poput očekivanja i kovarijance, prostorno-vremenskog semivariograma te prostorno-vremenske korelacije.

Korištenjem metoda vizualizacije pokušat ćemo istražiti prostorno-vremenske podatke kako bi podatke pripremili za modeliranje. Točnije, definirat ćemo, a zatim ćemo na danim podacima proučiti prve momente (očekivanja) po svakoj komponenti (u prostoru i u vremenu), druge momente (kovarijance) zasebno po svakoj komponenti (prostoru i u vremenu) te po objema komponentama istovremeno u obliku matrice. Proučit ćemo postoje li varijabilnosti u podacima po objema komponentama, odvojeno i istovremeno. Kako ćemo prostorno-vremenske podatke opisivati prostorno-vremenskim procesima, navest ćemo općenite definicije funkcije očekivanja i funkcije kovarijanci tih procesa.

Definicija 3.1. Neka je $\{X(\boldsymbol{s};t),\ \boldsymbol{s}\in\mathbb{D},\ t\in\mathbb{N}\}$, $\mathbb{D}\subseteq\mathbb{R}^l$, $l\in\{1,2\}$, slučajni proces koji je diskretan u vremenu i geostatistički u prostoru.

- Funkcija prostornog očekivanja procesa $\{X(\boldsymbol{s};t)\}$ je funkcija $\mu:\mathbb{D}\to\mathbb{R}$ definirana s
$$\mu_{space}(\boldsymbol{s})=\mathbf{E}X(\boldsymbol{s};t),$$
gdje je $\mathbf{E}[\,\cdot\,]$ funkcija očekivanja ovisna o lokaciji $\boldsymbol{s}$, ne i o vremenskom trenutku t.
- Funkcija vremenskog očekivanja procesa $\{X(\boldsymbol{s};t)\}$ je funkcija $\mu:\mathbb{N}\to\mathbb{R}$ definirana s
$$\mu_{time}(t)=\mathbf{E}X(\boldsymbol{s};t),$$
gdje je $\mathbf{E}[\,\cdot\,]$ funkcija očekivanja ovisna samo o trenutku t, ne i o lokaciji u prostoru $\boldsymbol{s}$.
- Funkcija kovarijanci procesa $\{X(\boldsymbol{s};t)\}$ je funkcija definirana s
$$\begin{aligned}C_X\left(\boldsymbol{s},\tilde{\boldsymbol{s}};t,\tilde{t}\right)&=Cov\left(X(\boldsymbol{s};t),X(\tilde{\boldsymbol{s}};\tilde{t})\right)=\\&=\mathbf{E}\left[(X(\boldsymbol{s};t)-\mathbf{E}X(\boldsymbol{s};t))(X(\tilde{\boldsymbol{s}};\tilde{t})-\mathbf{E}X(\tilde{\boldsymbol{s}};\tilde{t}))\right],\ \forall(\boldsymbol{s};t),(\tilde{\boldsymbol{s}};\tilde{t})\in\mathbb{D}\times\mathbb{N}.\end{aligned}$$

Napomena 3.1. Funkciju kovarijanci procesa $\{X(\boldsymbol{s};t)\}$ možemo kraće označiti s $\boldsymbol{C}_X$.

Definicija 3.2. Slučajni proces $\{X(\boldsymbol{s};t),\ \boldsymbol{s}\in\mathbb{D},\ t\in\mathbb{N}\}$, $\mathbb{D}\subseteq\mathbb{R}^l$, $l\in\{1,2\}$ je prostorno i vremenski stacionaran (u širem smislu, stacionaran drugog reda) ako vrijedi:

1) Postoje prostorni i vremenski drugi momenti procesa, odnosno
$$\mathbf{E}(X(\boldsymbol{s};t))^2<\infty,\ \forall(\boldsymbol{s};t)\in\mathbb{D}\times\mathbb{N}.$$

2) Prostorno-vremenski prvi momenti procesa su konstantni, tj.
$$\mathbf{E}X(\boldsymbol{s};t)=\mathbf{E}X(\tilde{\boldsymbol{s}};\tilde{t}),\ \forall(\boldsymbol{s};t),(\tilde{\boldsymbol{s}};\tilde{t})\in\mathbb{D}\times\mathbb{N}.\tag{3.1}$$

3) Funkcija kovarijanci procesa ovisi samo o pomaku među prostornim lokacijama $\tilde{\boldsymbol{s}}$ i $\boldsymbol{s}$ te razlici među vremenskim trenucima $\tilde{t}$ i t, odnosno
$$\begin{aligned}C_X\left(\boldsymbol{s},\tilde{\boldsymbol{s}};t,\tilde{t}\right)&=Cov\left(X(\boldsymbol{s};t),X(\tilde{\boldsymbol{s}};\tilde{t})\right)=\\&=C_X\left(\tilde{\boldsymbol{s}}-\boldsymbol{s};\tilde{t}-t\right),\ \forall(\boldsymbol{s};t),(\tilde{\boldsymbol{s}};\tilde{t})\in\mathbb{D}\times\mathbb{N}.\end{aligned}\tag{3.2}$$

Napomena 3.2. Uočimo sljedeće:

- Ako pomak u prostoru $\tilde{\boldsymbol{s}}-\boldsymbol{s}$ i razliku u vremenu $\tilde{t}-t$ označimo s $\boldsymbol{h}$ i τ, tim redom, onda funkciju kovarijanci stacionarnog procesa $\{X(\boldsymbol{s};t)\}$ možemo promatrati kao funkciju dvije varijable, pri čemu je udaljenost pomaka u prostoru $\|\boldsymbol{h}\|=\|\tilde{\boldsymbol{s}}-\boldsymbol{s}\|$ euklidska udaljenost[9]. Pišemo:
$$\begin{aligned}&C_X(\boldsymbol{s},\boldsymbol{s}+\boldsymbol{h};t,t+\tau)=Cov\left(X(\boldsymbol{s};t),X(\boldsymbol{s}+\boldsymbol{h};t+\tau)\right)=\\&=\mathbf{E}\left[(X(\boldsymbol{s};t)-\mathbf{E}X(\boldsymbol{s};t))(X(\boldsymbol{s}+\boldsymbol{h};t+\tau)-\mathbf{E}X(\boldsymbol{s}+\boldsymbol{h};t+\tau))\right]=\\&=C_X(\boldsymbol{h};\tau)\ \ \forall\boldsymbol{s},\boldsymbol{h}\in\mathbb{D},\forall t,\tau\in\mathbb{N}.\end{aligned}$$

[9]Neka su $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$ i $\boldsymbol{y}=(y_1,y_2,\ldots,y_n)$ n-dimenzionalni realni vektori. Euklidska udaljenost u n-dimenzionalnom realnom prostoru je preslikavanje $d:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ zadano formulom
$$d(\boldsymbol{x},\boldsymbol{y}):=\|\boldsymbol{x}-\boldsymbol{y}\|=\sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\intercal}(\boldsymbol{x}-\boldsymbol{y})}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2},$$
pri čemu je $\|\cdot\|$ euklidska norma na $\mathbb{R}^n$. Za više vidjeti [6].

- Razlika u vremenu (engl. *„time lag"*) τ je skalar, a pomak u prostoru (engl. *„spatial lag"*) $\boldsymbol{h}$ je l-dimenzionalan vektor (što odgovara pomaku između lokacija u l-dimenzionalnom prostoru $\mathbb{D}$).

- Prostornu kovarijancu stacionarnog procesa (koja ovisi samo o pomaku u prostoru) označavamo s $C_X(\boldsymbol{h}; 0)$ ili $C_X^{(\boldsymbol{s})}(\boldsymbol{h})$. Analogno, vremensku kovarijancu stacionarnog procesa označavamo s $C_X(\boldsymbol{0}; \tau)$ ili $C_X^{(t)}(\tau)$.

Budući da ne znamo stvarno očekivanje i stvarnu kovarijancu procesa, odredit ćemo njihovu procjenu na temelju uzorka među podacima.

3.2.1 Empirijsko prostorno i vremensko očekivanje

Pretpostavimo da imamo konačan niz podataka $\{x(\boldsymbol{s}; t)\}$ slučajnog procesa $\{X(\boldsymbol{s}; t),\ \boldsymbol{s} \in \mathbb{D},\ t \in \mathbb{N}\}$, $\mathbb{D} \subseteq \mathbb{R}^l$, $l \in \{1, 2\}$, na lokacijama $\{\boldsymbol{s}_i : i = 1, \ldots, m\}$ i u vremenima $\{t_j : j = 1, \ldots, T\}$. Empirijsko prostorno očekivanje na lokaciji $\boldsymbol{s}_i$ je srednja vrijednost u podacima kroz T vremenskih trenutaka, označavamo ga s $\widehat{\mu}_{space}$ i dano je formulom:

$$\widehat{\mu}_{space}(\boldsymbol{s}_i) = \frac{1}{T} \sum_{j=1}^{T} x(\boldsymbol{s}_i; t_j). \tag{3.3}$$

Tada je prostorno očekivanje m-dimenzionalni vektor, $\widehat{\boldsymbol{\mu}}_{space}$, pri čemu su $\widehat{\mu}(\boldsymbol{s}_1), \ldots, \widehat{\mu}(\boldsymbol{s}_m)$ empirijska prostorna očekivanja na lokacijama $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$ u T vremenskih trenutaka. Pišemo:

$$\widehat{\boldsymbol{\mu}}_{space} = \begin{bmatrix} \widehat{\mu}_{space}(\boldsymbol{s}_1) \\ \vdots \\ \widehat{\mu}_{space}(\boldsymbol{s}_m) \end{bmatrix} = \begin{bmatrix} \frac{1}{T}\sum_{j=1}^{T} x(\boldsymbol{s}_1; t_j) \\ \vdots \\ \frac{1}{T}\sum_{j=1}^{T} x(\boldsymbol{s}_m; t_j) \end{bmatrix} = \frac{1}{T}\sum_{j=1}^{T} \boldsymbol{x}_{t_j}, \tag{3.4}$$

pri čemu je $\boldsymbol{x}_{t_j} = [x(\boldsymbol{s}_1, t_j), \ldots, x(\boldsymbol{s}_m, t_j)]^\intercal$, $\forall\, j = 1, \ldots, T$, m-dimenzionalan vektor podataka na svakoj lokaciji.

Analogno definiramo empirijsko vremensko očekivanje u trenutku t_j, $\widehat{\mu}_{time}$, koje dobijemo kao prosjek u podacima kroz m lokacija, a dano je formulom:

$$\widehat{\mu}_{time}(t_j) = \frac{1}{m} \sum_{i=1}^{m} x(\boldsymbol{s}_i; t_j). \tag{3.5}$$

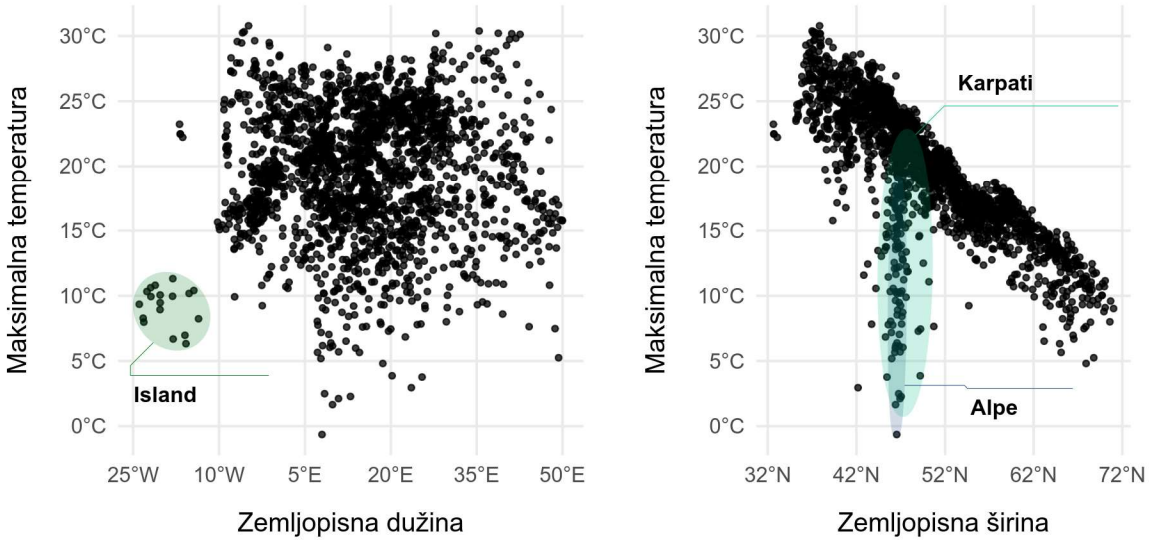
Procjena empirijskog očekivanja je brza i laka uz pomoć fukcija iz R paketa ***dplyr***. Na primjer, kako bi pronašli prostorno očekivanje, prvo grupiramo podatke po geografskoj dužini i širini funkcijom `group_by()`, a zatim računamo prosjek po svakoj lokacijskoj oznaci funkcijom `summarise()`. Analogno računamo vremensko očekivanje. Pogledajmo R kod.

```
spat_mean <- w_2022_Eu_3_8 %>%
  group_by(LATITUDE, LONGITUDE) %>%
  summarise(mu_emp = mean(MAX))

temp_mean <- w_2022_Eu_3_8 %>%
  group_by(YEARMODA) %>%
  summarise(mu_emp = mean(MAX))
```

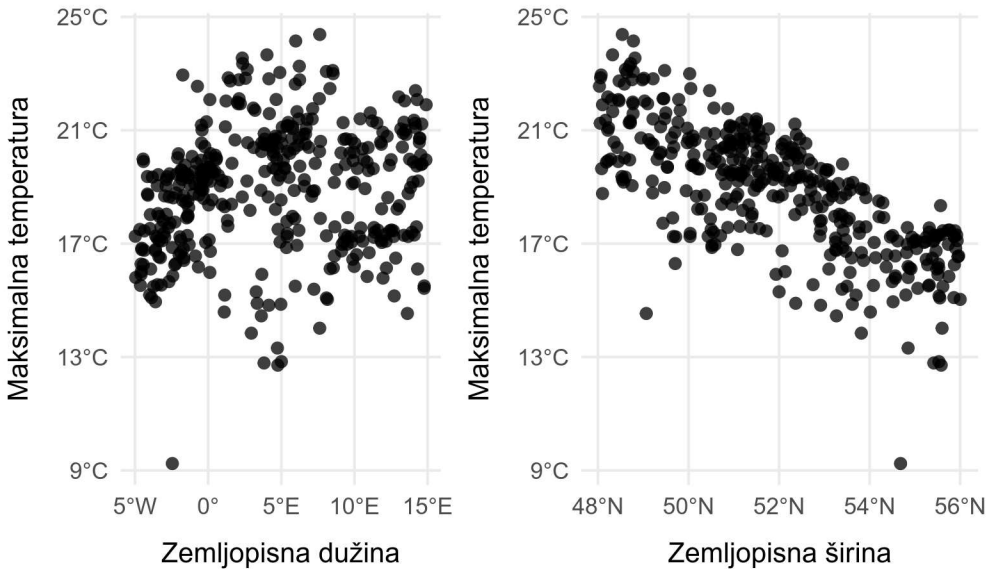
Prostorno očekivanje možemo prikazati grafički na dva načina. Jedan način smo već vidjeli na primjeru GSOD podataka dnevnih maksimalnih temperatura (slika 2.4), vide se vrijednosti temperatura u obliku točaka na zemljopisnoj karti kroz nekoliko tjedana. Ovakav prikaz nam može dati informaciju o promjenama vrijednosti svih triju komponenata istovremeno. Svake nedjelje se može vidjeti da je na jugu Hrvatske toplije, a na sjeveru hladnije, bez obzira koji dan bio. Također vidimo da se temperature od zapada prema istoku ne mijenjaju.

Drugi prikaz prostornog očekivanja je prikazan na slici 3.2, a podaci su od ožujka do kolovoza 2022. godine (temperature su u rastućem trendu kroz vrijeme) u cijeloj Europi. Na ovakav način vidimo očekivanje po svakoj prostornoj komponenti zasebno, a ne istovremeno (lijevi graf prikazuje očekivanje po zemljopisnoj dužini, a desni po širini). Iz ovih grafova se mogu uočiti nekakve anomalije koje vizualno odstupaju od ostatka podataka. Ono što odstupa su planine Karpati i Alpe te zemlja Island. Ta područja imaju znatno nižu temperaturu cijele godine u odnosu na cijelu Europu. Stoga, da bi nam bilo lakše kasnije modelirati podatke, uzet ćemo podskup podataka koji će isključiti ova anomalična područja.



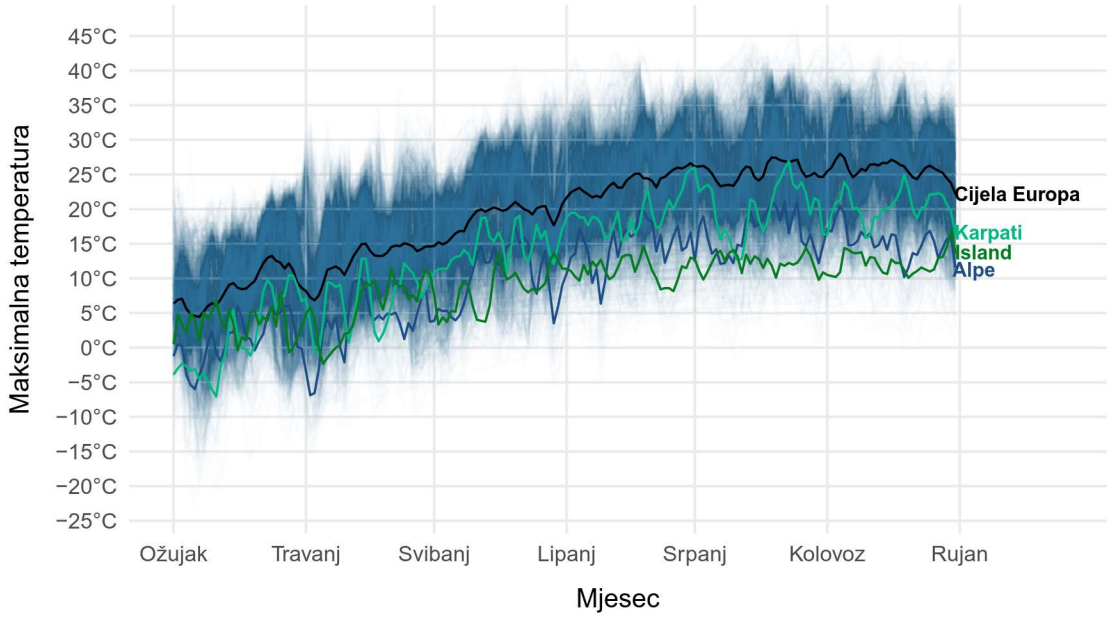
Slika 3.2: Primjer prostornog očekivanja na podacima GSOD dnevnih maksimalnih temperatura u 2022. godini od ožujka do kolovoza.

To možemo vidjeti na slici 3.3. Ograničili smo područje cijele Europe na „centar" (od 5° zapadne zemljopisne dužine do 15° istočne zemljopisne dužine i od 48° sjeverne zemljopisne širine do 56° sjeverne zemljopisne širine). Iz ovih grafova se jasno vidi trend empirijskog prostornog očekivanja maksimalnih temperatura po geografskoj širini, temperature opadaju prema sjeveru Europe, ali ne vidimo nikakav trend po geografskoj dužini.



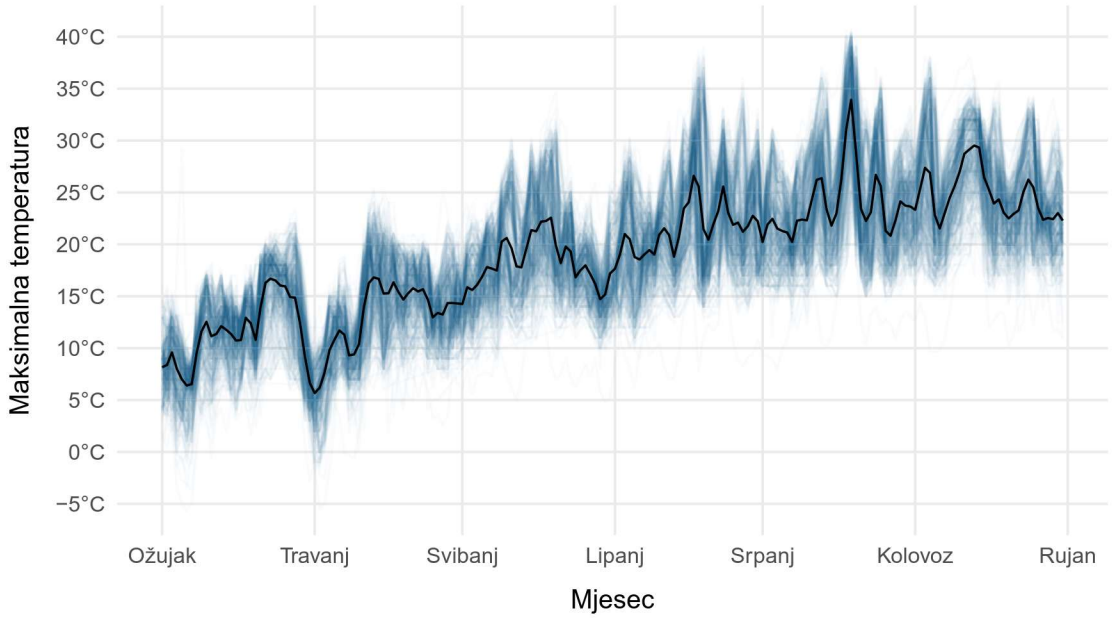
Slika 3.3: Primjer prostornog očekivanja na podacima GSOD dnevnih maksimalnih temperatura u 2022. godini od ožujka do kolovoza bez područja koja predstavljaju anomaliju.

Na slici 3.4 vidimo primjer empirijskog vremenskog očekivanja na istim podacima iz baze GSOD od ožujka do rujna. Na istom grafu, posebno smo istaknuli prosječno kretanje maksimalnih temperatura kroz vrijeme na području Karpata, Islanda i Alpa.



Slika 3.4: Primjer vremenskog očekivanja na podacima GSOD dnevnih maksimalnih temperatura u 2022. godini od ožujka do rujna.

Sada se definitivno vidi ta anomalija nižih temperatura od prosjeka cijele Europe. Vidi se i varijacija temperatura jer je uzeta cijela Europa u obzir. Stoga pogledajmo na slici 3.5 vremensko očekivanje bez anomalija na umanjenom području Europe (od 5° zapadne zemljopisne dužine do 15° istočne zemljopisne dužine i od 48° sjeverne zemljopisne širine do 56° sjeverne zemljopisne širine).



Slika 3.5: Primjer vremenskog očekivanja na podacima GSOD dnevnih maksimalnih temperatura u 2022. godini od ožujka do rujna u Europi bez područja koja predstavljaju anomaliju.

Već se tu može naslutiti da se ne radi o stacionarnom procesu, a kako bi se niz mogao modelirati, potrebno je imati neki oblik konstantnosti u vremenu i prostoru. Nestacionarnost se može ukloniti raznim transformacijama. U vremenskom očekivanju se pokazao polinomijalni trend jer od kolovoza do prosinca temperature opadaju, a u prostornom očekivanju se pokazao linearni trend.

3.2.2 Empirijska prostorna kovarijanca

Često je korisno razmotriti empirijsku prostornu kovarijancu u skupu prostorno-vremenskih podataka. Ovakva kovarijanca može se koristiti za određivanje kako opservacije u skupu podataka kovariraju (ponašaju se slično) kao funkcija dviju prostornih lokacija i funkcija koja je ovisna o razlici među vremenskim trenucima. Na primjer, zanima nas ponašaju li se vrijednosti temperatura na dvije lokacije u razmaku od 7 dana slično, početni dan ($t = 0$) i 7 dana nakon početnog dana ($t + \tau = 7$). Dok nam kovarijanca može samo reći nešto o tipu veze među dvije lokacije, korelacija će nam reći o jačini tih veza. Pozitivna kovarijanca će nam sugerirati da se opservacije na različitim lokacijama ponašaju slično, odnosno povećanjem vrijednosti opservacija na jednoj lokaciji dovest će do povećanja vrijednosti na nekoj drugoj lokaciji, ali ne znamo koliko. Negativna kovarijanca će sugerirati da će povećanje vrijednosti na jednoj lokaciji umanjiti vrijednosti na nekoj drugoj lokaciji, odnosno imat ćemo suprotnu vezu od pozitivne kovarijance. Ako gledamo kovarijancu između dvije iste lokacije, govorit ćemo o varijanci. Budući da ne znamo stvarnu prostornu kovarijancu $C_X(\boldsymbol{s}_i, \boldsymbol{s}_k; t, t+\tau)$ procesa $X(\boldsymbol{s}; t)$, naći ćemo njenu procjenu. Definirajmo empirijsku kovarijancu između dviju prostornih lokacija.

Definicija 3.3. Neka je $\{x(\boldsymbol{s}; t)\}$ konačan niz podataka slučajnog procesa $\{X(\boldsymbol{s}; t), \boldsymbol{s} \in \mathbb{D}, t \in \mathbb{N}\}$, $\mathbb{D} \subseteq \mathbb{R}^l$, $l \in \{1, 2\}$, koji je diskretan u vremenu i geostatistički u prostoru. Neka je $\tau = 0, 1, \ldots, T-1$ razlika u vremenu dva podatka tog procesa.

- Empirijska lag-τ prostorna kovarijanca je kovarijanca među prostornim lokacijama $\boldsymbol{s}_i$

i $\boldsymbol{s}_k$ ovisna o razlici u vremenu τ definirana je s

$$\widehat{C}_X^{(\tau)}(\boldsymbol{s}_i, \boldsymbol{s}_k) = \frac{1}{T-\tau} \sum_{j=1}^{T-\tau} \left(x(\boldsymbol{s}_i; t_{j+\tau}) - \widehat{\mu}_{space}(\boldsymbol{s}_i)\right)\left(x(\boldsymbol{s}_k; t_j) - \widehat{\mu}_{space}(\boldsymbol{s}_k)\right), \tag{3.6}$$

pri čemu je $\widehat{\mu}_{space}(\cdot) = \frac{1}{T-\tau}\sum_{j=1}^{T-\tau} x(\cdot; t_j)$ empirijsko prostorno očekivanje.

Raspisom prethodne sumacije (3.6), možemo doći do ekvivalentnog izraza sume

$$\widehat{C}_X^{(\tau)}(\boldsymbol{s}_i, \boldsymbol{s}_k) = \frac{1}{T-\tau} \sum_{j=\tau+1}^{T} \left(x(\boldsymbol{s}_i; t_j) - \widehat{\mu}_{space}(\boldsymbol{s}_i)\right)\left(x(\boldsymbol{s}_k; t_{j-\tau}) - \widehat{\mu}_{space}(\boldsymbol{s}_k)\right). \tag{3.7}$$

Uočimo da je ovo vremenski prosjek vektorskog umnoška centriranih opservacija na dvije lokacije $\boldsymbol{s}_i$ i $\boldsymbol{s}_k$, tj. (3.7) je ukupna kovarijacija među podacima.

- Izračunamo li empirijsku lag-τ prostornu kovarijancu za svaki par lokacija, (i, k), dobit ćemo empirijsku $m \times m$ lag-τ matricu prostornih kovarijanci[10] koja je dana izrazom

$$\widehat{\boldsymbol{C}}_X^{(\tau)} = \frac{1}{T-\tau} \sum_{j=\tau+1}^{T} \left(\boldsymbol{x}_{t_j} - \widehat{\boldsymbol{\mu}}_{space}\right)\left(\boldsymbol{x}_{t_j-\tau} - \widehat{\boldsymbol{\mu}}_{space}\right)^{\intercal}, \tag{3.8}$$

gdje su $\boldsymbol{x}_{t_j} = [x(\boldsymbol{s}_1, t_j), \ldots, x(\boldsymbol{s}_m, t_j)]^{\intercal}$, $\boldsymbol{x}_{t_j-\tau} = [x(\boldsymbol{s}_1, t_j - \tau), \ldots, x(\boldsymbol{s}_m, t_j - \tau)]^{\intercal}$, $\forall\, j = 1, \ldots, T$, i $\widehat{\boldsymbol{\mu}}_{space} = [\widehat{\mu}_{space}(\boldsymbol{s}_1) \ldots \widehat{\mu}_{space}(\boldsymbol{s}_m)]^{\intercal}$ m-dimenzionalni vektori.

Primjetimo da smo u (3.7) odabrali podijeliti s $(T-\tau)$, a ne s T. Ako podijelimo s T, empirijska lag-τ matrica prostornih kovarijanci (3.8) će biti nenegativno definitna[11]. Koristimo djelitelj $(T-\tau)$ jer tako dobijemo nepristranog procjenitelja kovarijance ako je pretpostavljena stacionarnost procesa. (vidi [3, poglavlje 5.1.3.])

Analogno možemo izračunati empirijsku lag-τ matricu kroskovarijanci, odnosno empirijsku lag-τ matricu kovarijanci dvaju prostorno-vremenskih procesa, $\{X(\boldsymbol{s}; t)\}$ i $\{Y(\boldsymbol{s}; t)\}$ (pretpostavlja se da vrijednosti procesa odgovaraju istim vremenskim trenucima).

- Empirijska lag-τ kroskovarijanca, $\widehat{\boldsymbol{C}}_{X,Y}^{(\tau)}$, između dva prostorno-vremenska skupa podataka, $\{x(\boldsymbol{s}; t)\}$ i $\{y(\boldsymbol{s}; t)\}$, pri čemu se $\{y(\boldsymbol{s}; t)\}$ sastoji od podataka na n različitih

[10] **Matrica kovarijanci** između slučajnih vektora $\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}$ i $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix}$, za koje vrijedi $\mathbf{E}X_i^2 < \infty$ i $\mathbf{E}Y_j^2 < \infty$, $\forall i = 1, \ldots, k$, $j = 1, \ldots, m$, je matrica

$$Cov(\boldsymbol{X}, \boldsymbol{Y}) = \mathbf{E}\left[(\boldsymbol{X} - \mathbf{E}\boldsymbol{X})(\boldsymbol{Y} - \mathbf{E}\boldsymbol{Y})^{\intercal}\right] = \begin{bmatrix} Cov(X_1, Y_1) & \cdots & Cov(X_1, Y_m) \\ \vdots & \ddots & \vdots \\ Cov(X_k, Y_1) & \cdots & Cov(X_k, Y_m) \end{bmatrix}.$$

[11] Pozitivno semidefinitna (nenegativno definitna) matrica je realna simetrična matrica $\mathbf{A}$ za koju vrijedi

$$\mathbf{x}^{\mathrm{T}} \cdot \mathbf{A} \cdot \mathbf{x} \geq 0, \, \forall\, \mathbf{x} \neq \mathbf{0}.$$

Vidi [2].

lokacija, ali u T istih vremenskih trenutaka kao $\{x(\boldsymbol{s};t)\}$. $\widehat{\boldsymbol{C}}_{X,Y}^{(\tau)}$ je $m \times n$ matrica dana s

$$\widehat{\boldsymbol{C}}_{X,Y}^{(\tau)} = \frac{1}{T-\tau} \sum_{j=\tau+1}^{T} \left(\boldsymbol{x}_{t_j} - \widehat{\boldsymbol{\mu}}_{X,space}\right) \left(\boldsymbol{y}_{t_j-\tau} - \widehat{\boldsymbol{\mu}}_{Y,space}\right)^{\intercal},$$

za svaki $\tau = 0, 1, \ldots, T-1$, gdje su $\widehat{\boldsymbol{\mu}}_{X,space}$ i $\widehat{\boldsymbol{\mu}}_{Y,space}$, empirijski prostorni vektori očekivanja podataka $\{x(\boldsymbol{s};t)\}$ i $\{y(\boldsymbol{s};t)\}$, tim redom. Kroskovarijance mogu biti korisne u opisivanju odnosa prostorno-vremenske ovisnosti između dviju različitih opisnih varijabli, na primjer maksimalne temperature i padaline. Koristimo ju pri modeliranju multivarijatnih geostatističkih procesa.

- Nadalje, definirajmo empirijsku $m \times m$ lag-τ matricu prostornih korelacija kao

$$\widehat{\boldsymbol{R}}_X^{(\tau)} = \widehat{\boldsymbol{D}}_X^{-\frac{1}{2}} \widehat{\boldsymbol{C}}_X^{(\tau)} \widehat{\boldsymbol{D}}_X^{-\frac{1}{2}},$$

gdje je $\widehat{\boldsymbol{D}}_X = \text{diag}\left(\widehat{\boldsymbol{C}}_X^{(0)}\right)$ dijagonalna matrica s empirijskim prostornim varijancama (s vremenskim prosjecima) na glavnoj dijagonali.

Korelacija nam je zapravo kvantiteta kovarijance, odnosno pokazuje nam koliko je jaka veza vrijednosti podataka između dvije lokacije (normalizirane vrijednosti kovarijanci). Korelacija može biti broj između -1 i 1. 1 označava jaku i direktnu vezu, -1 jaku suprotnu vezu, a 0 da su vrijednosti na dvije lokacije nezavisne jedna od druge.

Prije određivanja empirijskih lag-τ prostornih kovarijanci, važno je ukloniti trend. Jedan od načina je prvo napraviti procjenu podataka linearnim modelom (koji ima prostorne i/ili vremenske zavisnosti) jednostavnim pozivanjem funkcije `lm()` u R. Na primjeru GSOD baze podataka, u linearni model od nezavisnih varijabli uključili smo geografsku širinu (jer smo primijetili linearni trend), vremenski trenutak i kvadrat vremenskog trenutka (jer je pretpostavljen polinomijalni trend uočen na slici 2.1).

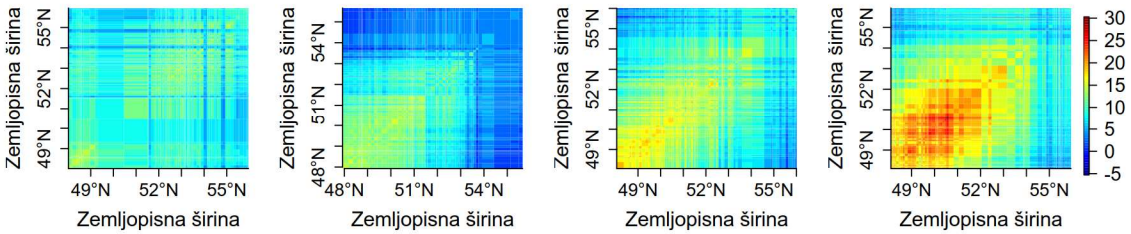
```
lmEu <- lm(MAX ~ LATITUDE + t + I(t^2), data = w_2022_Eu_3_8_bez_planina)
```

Dobivene reziduale[12] spremimo u matricu **rez_Eu** čiji retci označavaju vremenske trenutke, a stupci uređene parove prostornih oznaka. Zatim radimo procjenu kovarijanci reziduala dobivenog modela kao vektorski produkt reziduala za svaku prostornu lokaciju i za svaki vremenski trenutak. Najjednostavniji način za izračunavanje empirijske lag-τ matrice prostornih kovarijanci (3.8) je korištenje funkcije `cov()` u R-u. Za empirijsku lag-1 matricu kovarijanci određujemo kovarijancu između reziduala iz **rez_Eu** isključujući prvu vremensku točku i **rez_Eu** isključujući posljednju vremensku točku. Za empirijsku lag-7 matricu kovarijanci određujemo kovarijancu između reziduala iz **rez_Eu** isključujući prvih 7 vremenskih točaka i **rez_Eu** isključujući posljednjih 7 vremenskih točaka.

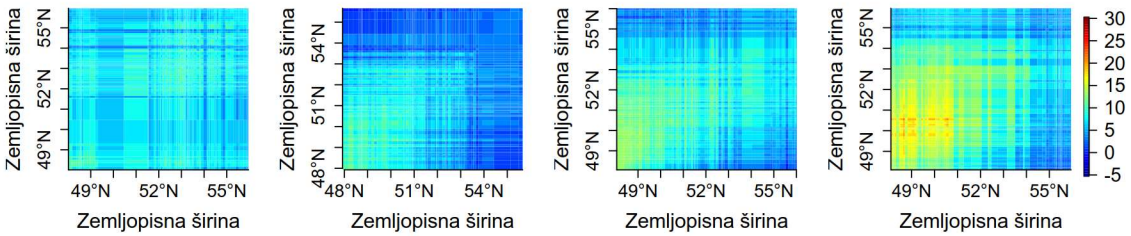
```
# empirijska lag-0 matrica kovarijanci
Lag0_covEu <- cov(rez_Eu, use = 'complete.obs')
# empirijska lag-1 matrica kovarijanci
Lag1_covEu <- cov(rez_Eu[-1, ], rez_Eu[-nrow(rez_Eu), ], use = 'complete.obs')
# empirijska lag-7 matrica kovarijanci
Lag7_covEu <- cov(rez_Eu[-(1:7), ], rez_Eu[-(nrow(rez_Eu):(nrow(rez_Eu)+1-7)), ],
                  use = 'complete.obs')
```

[12]Rezidual je odstupanje teorijske od procjenjene vrijednosti dobivene modelom među danim podacima.

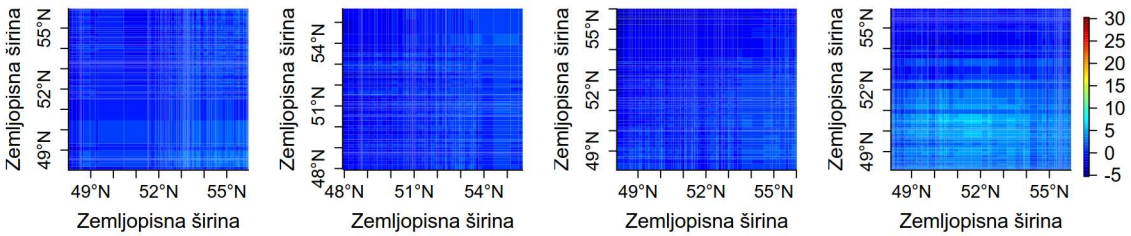
Kako bi nam bilo lakše predočiti matrice kovarijanci grafički jer lokacije u dvodimenzionalnom prostoru nemaju prirodni poredak, rastavit ćemo domenu geografskih dužina u četiri tzv. „dužinske trake" pa onda poredati podatke prema geografskoj širini. Pogledajmo kako to izgleda na slici 3.6, na primjeru GSOD podataka dnevnih maksimalnih temperatura u Europi. Nije iznenađujuće da ove empirijske matrice prostorne kovarijance otkrivaju prisutnost prostorne ovisnosti u rezidualima. Čini se da su grafovi 3.6a kvalitativno slični, što sugerira da ne postoji jaka korelacijska ovisnost o zemljopisnoj dužini, ali da postoji korelacijska ovisnost o zemljopisnoj širini, pri čemu prostorna varijanca po zemljopisnoj širini (glavna dijagonala matrica) opada prema sjeveru Europe, što znači da su na sjeveru Europe podaci manje raspršeni oko svog očekivanja. Na jugu Europe je jača veza među temperaturama na različitim lokacijama. Na 3.6b se vidi da su kovarijance nešto slabija nego lag-0 kovarijance, tj. slabija je veza u temperaturama od jučer na kojoj god lokaciji se nalazili. Na 3.6c se vidi da su kovarijance gotovo 0, odnosno ne uočava se veza u temperaturama otprije tjedan dana na kojoj god lokaciji se nalazili. Ova ovisnost je vrsta prostorne nestacionarnosti, a takvi grafovi se mogu koristiti za donošenje odluke jesu li nestacionarni prostorno-vremenski modeli potrebni ili ne ([4, str. 71]).



(a) Empirijska lag-0 matrica kovarijanci



(b) Empirijska lag-1 matrica kovarijanci



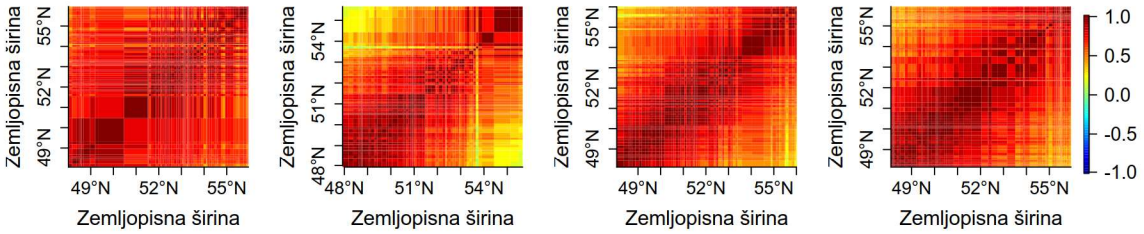
(c) Empirijska lag-7 matrica kovarijanci

Slika 3.6: Empirijske matrice kovarijanci rastavljene na četiri trake po geografskoj dužini na primjeru dnevnih maksimalnih temperatura u Europi iz baze GSOD.

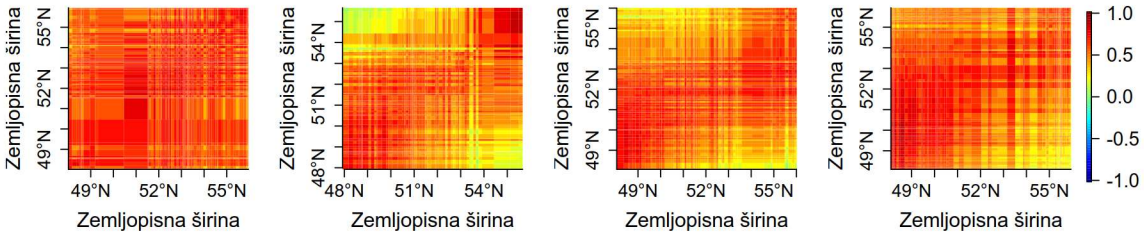
Sličnim postupkom dolazimo do empirijske lag-τ matrice prostornih korelacija.

```
# empirijska lag-0 matrica korelacija
Lag0_corEu <- cor(rez_Eu, use = 'complete.obs')
# empirijska lag-1 matrica korelacija
Lag1_corEu <- cor(rez_Eu[-1, ], rez_Eu[-nrow(rez_Eu), ], use = 'complete.obs')
# empirijska lag-7 matrica korelacija
Lag7_corEu <- cor(rez_Eu[-(1:7), ], rez_Eu[-(nrow(rez_Eu):(nrow(rez_Eu)+1-7)), ],
                  use = 'complete.obs')
```

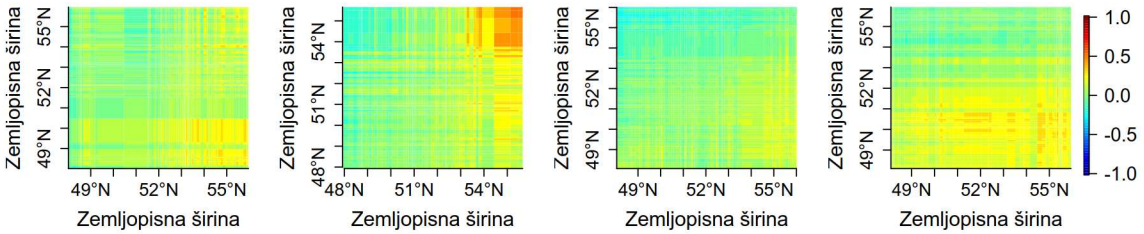
Pogledajmo na primjeru GSOD podataka matrice korelacija na slici 3.7. Sada se još bolje može uočiti sličnost među trakama na vremenskoj razlici 0 (lag-0) pa ne postoji jaka korelacijska ovisnost o zemljopisnoj dužini. Također se vidi na 3.7a da korelacija opada sa smanjenjem razlika između geografskih širina i da je na bližim lokacijama veza pozitivno jaka i direktna. Temperature jučer i danas imaju nešto manju korelaciju (3.7b), ali još uvijek pozitivnu, a temperature od prije tjedan dana i danas gotovo da nemaju korelacije (3.7c).



(a) Empirijska lag-0 matrica korelacija



(b) Empirijska lag-1 matrica korelacija



(c) Empirijska lag-7 matrica korelacija

Slika 3.7: Empirijske matrice korelacija rastavljene na četiri trake po geografskoj dužini na primjeru dnevnih maksimalnih temperatura u Europi iz baze GSOD.

3.2.3 Prostorno-vremenski variogram

Nakon što smo promotrili empirijsko prostorno i vremensko očekivanje i empirijske prostorne kovarijance, preostaje nam definirati i promotriti strukturu prostorno-vremenske ovisnosti jer se ovdje ipak radi o prostorno-vremenskim podacima koje želimo modelirati prostorno-vremenskim procesom. Mjeru zajedničke prostorne i vremenske ovisnosti zovemo variogram.

Definicija 3.4. Neka je $\{X(\boldsymbol{s};t),\ \boldsymbol{s}\in\mathbb{D},\ t\in\mathbb{N}\}$, $\mathbb{D}\subseteq\mathbb{R}^l$, $l\in\{1,2\}$, slučajni proces koji je diskretan u vremenu i geostatistički u prostoru. Variogram je definiran kao

$$2\gamma_X\left(\boldsymbol{s},\tilde{\boldsymbol{s}};t,\tilde{t}\right)=Var\left(X(\boldsymbol{s};t)-X(\tilde{\boldsymbol{s}};\tilde{t})\right)\ \forall\,(\boldsymbol{s};t),\ (\tilde{\boldsymbol{s}};\tilde{t})\in\mathbb{D}\times\mathbb{N},$$

pri čemu γ_X zovemo semivariogram.

U slučaju da kovarijanca procesa ovisi samo o pomaku među prostornim lokacijama $\tilde{\boldsymbol{s}}$ i $\boldsymbol{s}$ te razlici među vremenskim trenucima $\tilde{t}$ i t dolazimo do sljedećeg raspisa semivariograma:

$$\begin{aligned}
\gamma_X(\boldsymbol{h};\tau) &= \gamma_X(\boldsymbol{s},\boldsymbol{s}+\boldsymbol{h};t,t+\tau)\ = \\
&= \frac{1}{2}Var\big(X(\boldsymbol{s}+\boldsymbol{h};t+\tau)-X(\boldsymbol{s};t)\big)= \\
&= \frac{1}{2}\Big[\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau)-X(\boldsymbol{s};t)\right]^2-\big(\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau)-X(\boldsymbol{s};t)\right]\big)^2\Big]= \\
&= \frac{1}{2}\Big[\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau)\right]^2-2\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau)X(\boldsymbol{s};t)\right]+\mathbf{E}\left[X(\boldsymbol{s};t)\right]^2\Big]+ \\
&\quad-\frac{1}{2}\Big[\big(\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau)\right]\big)^2-2\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau)\right]\mathbf{E}\left[X(\boldsymbol{s};t)\right]+\big(\mathbf{E}\left[X(\boldsymbol{s};t)\right]\big)^2\Big]= \\
&= \frac{1}{2}\Big(\underbrace{Var\big(X(\boldsymbol{s}+\boldsymbol{h};t+\tau)\big)}_{(\star)}+\underbrace{Var\big(X(\boldsymbol{s};t)\big)}_{(\star\star)}-2Cov\big(X(\boldsymbol{s}+\boldsymbol{h};t+\tau),X(\boldsymbol{s};t)\big)\Big), \\
&\quad\forall\,\boldsymbol{s},\boldsymbol{h}\in\mathbb{D},\ \forall\,t,\tau\in\mathbb{N}.
\end{aligned}$$

Iskoristimo sada činjenicu da je varijanca slučajne varijable zapravo kovarijanca sama sa sobom, a potom definiciju kovarijance stacionarnog prostorno-vremenskog procesa (3.2):

$$(\star\star)\ =\ Var\big(X(\boldsymbol{s};t)\big)=\ Cov\big(X(\boldsymbol{s};t),X(\boldsymbol{s};t)\big)=\ C_X(\boldsymbol{s}-\boldsymbol{s};t-t)\ =\ C_X(\mathbf{0};0)$$

i analogno,

$$(\star)\ =\ Var\big(X(\boldsymbol{s}+\boldsymbol{h};t+\tau)\big)=\ C_X(\mathbf{0};0)\,.$$

Uvrstimo li $(\star)$ i $(\star\star)$ u definiciju semivariograma, dobivamo:

$$\begin{aligned}
\gamma_X(\boldsymbol{h};\tau)\ &=\ \frac{1}{2}\Big(C_X(\mathbf{0};0)+C_X(\mathbf{0};0)-2C_X(\boldsymbol{h};\tau)\Big)= \\
&=\ C_X(\mathbf{0};0)-C_X(\boldsymbol{h};\tau)\ \ \forall\,\boldsymbol{s},\boldsymbol{h}\in\mathbb{D},\ \forall\,t,\tau\in\mathbb{N},
\end{aligned}\tag{3.9}$$

gdje je $C_X(\boldsymbol{h};\tau)$ funkcija kovarijanci stacionarnog procesa. Variogram je ovim raspisom funkcija ovisna o pomaku u prostoru i vremenskoj razlici. Znači da su semivariogram i funkcija kovarijanci inverzno proporcionalni [13]. Odsada će se riječ variogram koristiti kao sinonim za semivariogram.

Kako ne znamo točno odrediti variogram, definirat ćemo empirijski variogram prostorno-vremenskog procesa kao procjenu variograma.

[13] Gaurav Arora je detaljnije opisao variogram, kovarijancu i korelaciju na sljedećem videu: `https://youtu.be/YR1YFpLwvPs?t=391`

Definicija 3.5. Neka je $\{x(\boldsymbol{s};t)\}$ konačan niz podataka slučajnog procesa $\{X(\boldsymbol{s};t),\ \boldsymbol{s} \in \mathbb{D},\ t \in \mathbb{N}\}$, $\mathbb{D} \subseteq \mathbb{R}^l$, $l \in \{1,2\}$, koji je diskretan u vremenu i geostatistički u prostoru. Pretpostavimo da prvi momenti ovise samo o prostoru, ne i o vremenu, i pretpostavimo da drugi momenti ovise samo o razlikama u vremenu i pomacima u prostoru. Tada empirijski prostorno-vremenski variogram za pomak u prostoru $\boldsymbol{h}$ i razliku u vremenu τ definiramo kao

$$\widehat{\gamma}_X(\boldsymbol{h};\tau) = \widehat{C}_X(\boldsymbol{0};0) - \widehat{C}_X(\boldsymbol{h};\tau) \ \ \forall \boldsymbol{h} \in \mathbb{D},\ \forall \tau \in \mathbb{N}.$$

Ako osim pretpostavki iz definicije 3.5 još pretpostavimo da je prostorno očekivanje μ_{space} konstantno (3.1), u tom slučaju slučajni proces je stacionaran u užem smislu pa variogram kao u 3.9 možemo raspisati na drugi način:

$$\begin{aligned}
\gamma_X(\boldsymbol{h};\tau) &= \gamma_X(\boldsymbol{s},\boldsymbol{s}+\boldsymbol{h};t,t+\tau) = \\
&= \frac{1}{2}Var\big(X(\boldsymbol{s}+\boldsymbol{h};t+\tau) - X(\boldsymbol{s};t)\big) = \\
&= \frac{1}{2}\Big[\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau) - X(\boldsymbol{s};t)\right]^2 - \big(\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau) - X(\boldsymbol{s};t)\right]\big)^2\Big] = \\
&= \frac{1}{2}\Big[\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau) - X(\boldsymbol{s};t)\right]^2 - \big(\mathbf{E}X(\boldsymbol{s}+\boldsymbol{h};t+\tau) - \mathbf{E}X(\boldsymbol{s};t)\big)^2\Big] = \\
&= \frac{1}{2}\Big[\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau) - X(\boldsymbol{s};t)\right]^2 - \big(\mathbf{E}X(\boldsymbol{s};t) - \mathbf{E}X(\boldsymbol{s};t)\big)^2\Big] = \\
&= \frac{1}{2}\mathbf{E}\left[X(\boldsymbol{s}+\boldsymbol{h};t+\tau) - X(\boldsymbol{s};t)\right]^2 \ \ \forall \boldsymbol{s},\boldsymbol{h} \in \mathbb{D},\ \forall t,\tau \in \mathbb{N}
\end{aligned}$$

i time dolazimo do alternativne procjene variograma

$$\widehat{\gamma}_X(\boldsymbol{h};\tau) = \frac{1}{2}\frac{1}{|N_{\boldsymbol{s}}(\boldsymbol{h})|}\frac{1}{|N_t(\tau)|}\sum_{\boldsymbol{s}_i,\boldsymbol{s}_k \in N_{\boldsymbol{s}}(\boldsymbol{h})}\ \sum_{t_j,t_\ell \in N_t(\tau)} \left(x(\boldsymbol{s}_i;t_j) - x(\boldsymbol{s}_k;t_\ell)\right)^2, \tag{3.10}$$

pri čemu je $N_{\boldsymbol{s}}(\boldsymbol{h})$ skup koji sadrži parove lokacija s jednakom međusobnom prostornom udaljenošću $\boldsymbol{h}$, $N_t(\tau)$ skup koji sadrži parove vremenskih trenutaka s jednakom međusobnom vremenskom razlikom τ, a $|N(\cdot)|$ označava broj elemenata u skupu $N(\cdot)$.

Prikazat ćemo semivariogram na primjeru maksimalnih temperatura u Europi iz baze GSOD, ali ćemo odsada raditi na manjem skupu podataka kako bi olakšali računanje u R-u pa ćemo uzeti podatke iz srpnja 2022. Računanje empirijskog semivariograma je puno brže korištenjem klase podataka STFDF[14] nego korištenjem STIDF[15], kojeg većinom već imamo u našim podacima.

Za izračun empirijskog semivariograma na našim podacima, iskoristit ćemo funkciju `variogramST()` [16] u R-u koja uzima za argumente formulu kojom odredimo ovisnu varijablu *MAX* i neovisnu varijablu *LATITUDE*, STFDF objekt *STFDF_EuJuly*, pomake u prostoru od 0, 20, 60, 100, ..., 500 km i razlike u vremenu od 0, 1, 2, ..., 12 dana.

[14] Kombinacija **sp** objekta (vektor lokacijskih oznaka, npr. uređenih parova geografskih širina i dužina) i **xts** objekta (vektor svih vremenskih oznaka, npr. datuma) za predstavljanje svih mogućih lokacija na prostorno-vremenskoj rešetki.

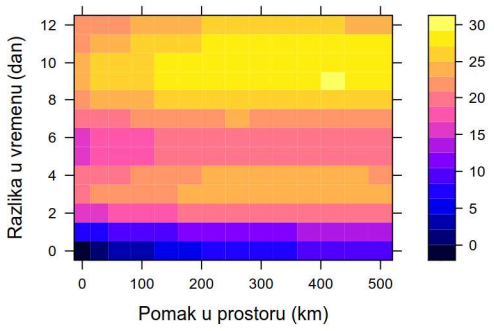
[15] Nepravilna prostorno-vremenska struktura podataka, gdje je svakoj točki dodijeljena prostorna koordinata i vremenska oznaka.

[16] Može i `variogram()`, ali ćemo dobiti isto jer `variogram()` prepoznaje radi li se o prostornim ili prostorno-vremenskim podacima.

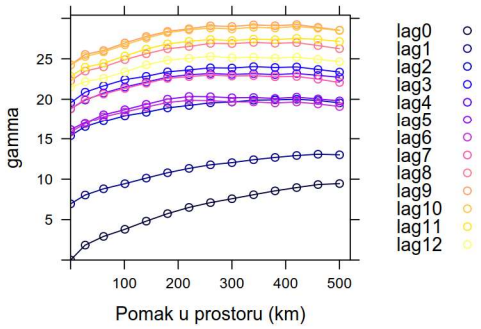
```
var <- variogramST(MAX ~ 1 + LATITUDE,
                   data = STFDF_EuJuly,
                   width = 40,
                   tlags = 0.01:12.01)
```

Variogram se može prikazati grafički na tri načina [10], 2.5D graf na slici 3.8a, paralelni prostorni variogrami, jedan po razlici u vremenu τ na slici 3.8b i 3D „wireframe[17]" graf na slici 3.8c. Ovi grafovi nam preporučuju da postoji prostorno-vremenska ovisnost u podacima pa je opravdano modelirati prostorno-vremenske reziduale, a dobijemo ih sljedećim naredbama:

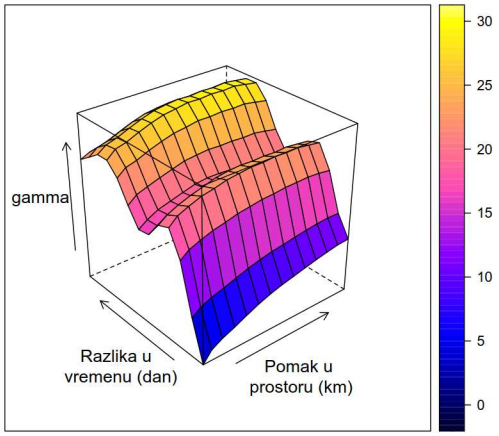
```
plot(var, map = T, xlab = "Pomak u prostoru (km)",
     ylab = "Razlika u vremenu (dan)") # 2.5D graf
plot(var, map = F, xlab = "Pomak u prostoru (km)",
     ylab = "Razlika u vremenu (dan)") # Paralelni prostorni variogrami
plot(var, wireframe = TRUE, xlab = "\n Pomak u \n prostoru (km)",
     ylab = "\nRazlika u \n vremenu (dan)",
     par.settings = list(fontsize = list(text= 14))) # 3D "wireframe" graf
```



(a) 2.5D graf



(b) Paralelni prostorni variogrami, jedan po razlici u vremenu τ



(c) 3D „wireframe" graf

Slika 3.8: Tri načina za prikaz semivariograma na primjeru maksimalnih temperatura u Europi u srpnju 2022.

[17]Tzv. žičani okvir, uzima mrežu vrijednosti i projicira je na navedenu trodimenzionalnu površinu, a rezultirajuće trodimenzionalne oblike može učiniti prilično lakima za vizualizaciju.

4 Modeli u geostatistici

Nakon što smo se upoznali s prostorno-vremenskim podacima u geostatistici i analizirali ih na primjeru GSOD podataka, tj. dnevnih maksimalnih temperatura 2022. godine u Europi, preostaje nam još opisati te podatke modelom. Model je matematički objekt kojim se pokušava opisati kako neke varijable funkcioniraju u stvarnom svijetu pomoću jednadžbi. Koriste se za predviđanje ili procjenu što bi se dogodilo da napravimo promjenu neke varijable. Prostorno-vremenska geostatistika bavi se statističkim modeliranjem varijabli koje variraju i u prostoru i u vremenu. Analogon je uobičajenoj geostatistici, koja razmatra samo prostorne varijabilnosti varijabli. U ovom poglavlju razmatrat ćemo prostorno-vremenske modele korištenjem deskriptivnog pristupa, što je otkrivanje obrazaca među podacima. Ovaj pristup sadrži razne tehnike:

- **inverzna ponderirana udaljenost** (engl. *„Inverse Distance Weighting"* ili skraćeno IDW) je jedna od najjednostavnijih deterministčkih prostorno-vremenskih interpolacijskih metoda. Metodom se određuju težinski prosjeci na nekoj lokaciji u nekom vremenu među danim podacima, pri čemu se više težine dodjeljuje vremenski i prostorno „bližim" opservacijama. Vidi [4, poglavlje 3.1].

- **regresijski modeli** (linearni model, generalizirani linearni model, generalizirani aditivni model) se koriste za dobivanje predviđanja prostorno-vremenskih podataka uz pretpostavku da se sva prostorno-vremenska zavisnost može objasniti u terminima „trenda" kao prediktor. U prostorno-vremenskim podacima se ova metoda pokazuje problematična zbog narušavanja pretpostavki modela (prostorne, vremenske i prostorno-vremenske zavisnosti među opservacijama). Vidi [4, poglavlje 3.2-3.4].

- **kriging** je složenija metoda interpolacije koja se koristi u geostatistici i temelji se na teoriji vjerojatnosti i slučajnih procesa za opisivanje varijabilnosti u vremenu i prostoru. Pretpostavlja se da se podaci mogu modelirati tzv. Gaussovskim procesom[18] sa zadanom kovarijancom. Za dobivanje kriging prediktora potrebni su podaci i model funkcije kovarijanci tih podataka.

Osim deskriptivnog pristupa, imamo i Bayesov pristup ([11]), dinamički pristup ([4, poglavlje 5]), analogan je analizi multivarijatnih vremenskih nizova i metode strojnog učenja (engl. *„Machine Learning"*, [9]), ali ih nećemo obraditi u ovom radu. Kako bi odabrali metodu modeliranja, potrebno je postaviti si cilj, što nas zanima u podacima, na koje pitanje želimo dobiti odgovor, a to su:

- procjena vrijednosti ovisne varijable na nekoj lokaciji u danom prostoru unutar vremenskog perioda (regresija)

- procjena vrijednosti u nekoj točki u prostoru i vremenu uz pomoć prostorno-vremenskih ovisnosti među danim podacima (interpolacija)

- predviđanje buduće vrijednosti ovisne varijable na nekoj lokaciji (dinamički modeli)

Razina složenosti modela ovisi o pitanju koje se postavlja. Također, na odabir metode će utjecati i specifične karakteristike prostorno-vremenskih podataka te dostupni računalni resursi. Fokus u ovom radu će biti na tehnike kriginga s primjenom na dnevne temperature u Europi.

[18] **Gaussovski proces** je slučajni proces $X = \{X(\boldsymbol{r}) : \boldsymbol{r} \in \mathbb{D}\}$, gdje je $\boldsymbol{r}$ prostorna, vremenska ili prostorno-vremenska oznaka iz skupa $\mathbb{D} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, kod kojeg su sve konačnodimenzionalne distribucije normalne s funkcijom očekivanja $\mu(\boldsymbol{r})$ i funkcijom kovarijanci $C_X(\boldsymbol{r}, \tilde{\boldsymbol{r}}) = Cov(X(\boldsymbol{r}), X(\tilde{\boldsymbol{r}}))$, $\forall \boldsymbol{r}, \tilde{\boldsymbol{r}} \in \mathbb{D}$.

4.1 Kriging

Ovu metodu interpolacije razvio je Danie Gerhardus Krige, južnoafrički statističar i rudarski inženjer. Razvio ju je za prostorne podatke pa je naknadno proširena na prostorno-vremenske podatke uključivanjem prostorne i vremenske zavisnosti. Potpuno je analogna prostornom krigingu, samo što sada imamo tri komponente umjesto dvije. Osim dvije prostorne, u kriging je uključena i vremenska komponenta. Koristi se za predviđanje željenih varijabli u prostorno-vremenskim točkama gdje nije zabilježeno mjerenje pa se na taj način popunjavaju nedostaci u prostorno-vremenskim podacima. Osim na nepoznatim podacima, kriging se može primijeniti i na već postojeće podatke pa ih na taj način izglađujemo.

Ovom metodom razvijamo tzv. hijerarhijski model jer se sastoji od modela koji u sebi sadrži „skriveni" proces dobiven nekim drugim modelom. Hijerarhijsko statističko modeliranje predstavlja način izgradnje modela unutar nekog drugog modela kroz dobro definirane uvjetne vjerojatnosti[19]. Sastoji se od dva koraka:

1. izglađivanja postojećih podataka jer sumnjamo u njihovu pogrešku pri mjerenju (niti jedan senzor ne može bilježiti točnu vrijednost) - Wikle ga zove „data model" [4, poglavlje 4]. Sastoji se od nekog latentnog („skrivenog") slučajnog procesa[20].

2. modeliranje „pravog" slučajnog procesa, kojemu je uklonjena pogreška pri mjerenju. Gradi se određivanjem težina statističkim modeliranjem kovarijanci (prostorno-vremenska zavisnost među podacima) - tzv. „process model" [4, poglavlje 4].

Moguće je da je i sam model skrivenog procesa sastavljen od podmodela također izražen uvjetnom distribucijom. Model dobiven na temelju podataka se zapravo smatra aditivnom greškom pri mjerenju, a skriveni proces se smatra dekompozicijom na fiksni (deterministički) i slučajni (stohastički) dio.

Neka je $\{z(\boldsymbol{s};t)\}$ $(m\cdot T)$-dimenzionalni niz podataka slučajnog procesa $Z=\{Z(\boldsymbol{s};t),\ \boldsymbol{s}\in\mathbb{D},\ t\in\mathbb{N}\}$, $\mathbb{D}\subseteq\mathbb{R}^2$, diskretnog u vremenu i geostatističkog u prostoru, na lokacijama $\{\boldsymbol{s}_i : i=1,\ldots,m\}\subseteq\mathbb{R}^2$ i u vremenima $\{t_j : j=1,\ldots,T\}\subseteq\mathbb{N}$, $t_1<t_2<\cdots<t_T$. Skup $\{t_j : j=1,\ldots,T\}$ može označavati uzastopne sekunde, minute, sate, dane, mjesece, godine i slično, a skup $\{\boldsymbol{s}_i : i=1,\ldots,m\}$ označava geografsku širina i dužinu. Kako znamo da neke vrijednosti, niti ljudska ruka, niti senzor ili radar ne mogu točno izmjeriti, već u samom početku modeliranja moramo zabilježenim podacima dati aditivnu pogrešku pri mjerenju. Neka je tada proces $\{Z(\boldsymbol{s};t)\}$ zadan s

$$Z(\boldsymbol{s};t)=Y(\boldsymbol{s};t)+\mathcal{E}(\boldsymbol{s};t), \tag{4.1}$$

gdje je $\{Y(\boldsymbol{s};t),\ \boldsymbol{s}\in\mathbb{D},\ t\in\mathbb{N}\}$ slučajni prostorno-vremenski proces stvarnih vrijednosti bez šuma[21] (bez pogreške pri mjerenju) i $\{\mathcal{E}(\boldsymbol{s};t),\ \boldsymbol{s}\in\mathbb{D},\ t\in\mathbb{N}\}\sim IID(0,\sigma^2_{\mathcal{E}})$ je aditivna greška pri mjerenju koja je nezavisna od $\{Y(\boldsymbol{s};t)\}$, a predstavlja nezavisni i jednako distribuirani

[19]Uvjetna vjerojatnost na vjerojatnosnom prostoru $(\Omega,\mathcal{F},\mathrm{P})$ uz uvjet da se dogodio događaj $A\in\mathcal{F}$ je funkcija $\mathrm{P}(\,\cdot\,|\,A)$ definirana na $\mathcal{F}$ izrazom

$$\mathrm{P}(B\,|\,A)=\frac{\mathrm{P}(A\cap B)}{\mathrm{P}(A)},\quad B\in\mathcal{F},$$

pri čemu je $\mathrm{P}(A)>0$.

[20]Latentni proces je proces slučajnih varijabli koje su „prikrivene", tj. procjenjuju se indirektno izgradnjom modela uvjetne vjerojatnosti za dane podatke

[21]Nezavisni i jednako distribuirani šum (n.j.d. šum) je niz $\{X_t\}$ nezavisnih i jednako distribuiranih slučajnih varijabli takvih da je $\mathbf{E}X_t^2\leq\infty$, $\mathbf{E}X_t=0$, $\mathrm{Var}X_t=\sigma^2$. Označavamo ga s $\{X_t\}\sim IID(0,\sigma^2)$.

šum s očekivanjem 0 i varijancom $\sigma^2_\mathcal{E}$. $\{Y(\boldsymbol{s};t)\}$ je zapravo latentni („skriveni") proces dan modelskom jednadžbom

$$Y(\boldsymbol{s};t) = \mu(\boldsymbol{s};t) + \eta(\boldsymbol{s};t), \quad \forall(\boldsymbol{s};t) \in \mathbb{D} \times \mathbb{N}, \tag{4.2}$$

pri čemu $\mu(\boldsymbol{s};t)$ predstavlja očekivanje procesa $\{Y(\boldsymbol{s};t)\}$, i nije slučajan proces, a $\eta(\boldsymbol{s};t)$ predstavlja slučajni proces s očekivanjem 0 i prostorno-vremenskom zavisnošću. Ovisno o tome kakav je $\mu(\boldsymbol{s};t)$, kriging može biti:

(i) jednostavni (engl. *„simple"*) kriging - $\mu(\boldsymbol{s};t)$ je poznata konstanta $\forall\,(\boldsymbol{s};t) \in \mathbb{D} \times \mathbb{N}$,

(ii) obični (engl. *„ordinary"*) kriging - $\mu(\boldsymbol{s};t)$ je nepoznata konstanta $\forall\,(\boldsymbol{s};t) \in \mathbb{D} \times \mathbb{N}$ i

(iii) univerzalni (engl. *„universal"*) kriging - $\mu(\boldsymbol{s};t)$ nije konstanta, nego linearna kombinacija fiksnih, ali nepoznatih parametara.

Glavni cilj kriginga je odrediti statistički optimalnog prediktora slučajne varijable $Y(\boldsymbol{s}_0;t_0)$ na prostoru $\boldsymbol{s}_0 \in \mathbb{R}^2$ i u vremenu $t_0 \in \mathbb{N}$ na temelju slučajnog procesa $\{Y(\boldsymbol{s};t)\}$, a označit ćemo ga s $\widehat{Y}(\boldsymbol{s}_0;t_0)$. Pod optimalnost se misli naći najboljeg linearnog prediktora $\widehat{Y}(\boldsymbol{s}_0;t_0)$, tj. linearnu funkciju koja će minimizirati srednje kvadratnu grešku predikcije, $\mathbf{E}[Y(\boldsymbol{s}_0;t_0) - \widehat{Y}(\boldsymbol{s}_0;t_0)]^2$, između stvarne vrijednosti $Y(\boldsymbol{s}_0;t_0)$ i prediktirane vrijednosti $\widehat{Y}(\boldsymbol{s}_0;t_0)$. Tako dobiveni prediktor će se zvati *„kriging prediktor"*. Zapravo, želimo procijeniti latentnu vrijednost $Y\,(\boldsymbol{s}_0;t_0)$ kao funkciju danih podataka u obliku vektora

$$\begin{aligned}\boldsymbol{z} = \big(&z(\boldsymbol{s}_1;t_1),\, z(\boldsymbol{s}_2;t_1),\, \ldots,\, z(\boldsymbol{s}_m;t_1),\\ &z(\boldsymbol{s}_1;t_2),\, z(\boldsymbol{s}_2;t_2),\, \ldots,\, z(\boldsymbol{s}_m;t_2),\\ &\ldots,\\ &z(\boldsymbol{s}_1;t_T),\, z(\boldsymbol{s}_2;t_T),\, \ldots,\, z(\boldsymbol{s}_m;t_T)\big)^\mathsf{T}\end{aligned}$$

Modelske jednadžbe (4.1) i (4.2) se ekvivalentno mogu prikazati modelskim jednadžbama na temelju podataka u vektorskom obliku:

$$\begin{aligned}\boldsymbol{z} &= \boldsymbol{y} + \boldsymbol{\varepsilon} \text{ - Izglađivanje podataka,}\\ \boldsymbol{y} &= \boldsymbol{\mu} + \boldsymbol{\eta} \text{ - Određivanje težina,}\end{aligned}$$

gdje su

$$\begin{aligned}\boldsymbol{y} = \big(&y(\boldsymbol{s}_1;t_1),\, y(\boldsymbol{s}_2;t_1),\, \ldots,\, y(\boldsymbol{s}_m;t_1),\\ &y(\boldsymbol{s}_1;t_2),\, y(\boldsymbol{s}_2;t_2),\, \ldots,\, y(\boldsymbol{s}_m;t_2),\\ &\ldots,\\ &y(\boldsymbol{s}_1;t_T),\, y(\boldsymbol{s}_2;t_T),\, \ldots,\, y(\boldsymbol{s}_m;t_T)\big)^\mathsf{T},\end{aligned} \qquad \begin{aligned}\boldsymbol{\varepsilon} = \big(&\varepsilon(\boldsymbol{s}_1;t_1),\, \varepsilon(\boldsymbol{s}_2;t_1),\, \ldots,\, \varepsilon(\boldsymbol{s}_m;t_1),\\ &\varepsilon(\boldsymbol{s}_1;t_2),\, \varepsilon(\boldsymbol{s}_2;t_2),\, \ldots,\, \varepsilon(\boldsymbol{s}_m;t_2),\\ &\ldots,\\ &\varepsilon(\boldsymbol{s}_1;t_T),\, \varepsilon(\boldsymbol{s}_2;t_T),\, \ldots,\, \varepsilon(\boldsymbol{s}_m;t_T)\big)^\mathsf{T},\end{aligned}$$

$$\begin{aligned}\boldsymbol{\mu} = \big(&\mu(\boldsymbol{s}_1;t_1),\, \mu(\boldsymbol{s}_2;t_1),\, \ldots,\, \mu(\boldsymbol{s}_m;t_1),\\ &\mu(\boldsymbol{s}_1;t_2),\, \mu(\boldsymbol{s}_2;t_2),\, \ldots,\, \mu(\boldsymbol{s}_m;t_2),\\ &\ldots,\\ &\mu(\boldsymbol{s}_1;t_T),\, \mu(\boldsymbol{s}_2;t_T),\, \ldots,\, \mu(\boldsymbol{s}_m;t_T)\big)^\mathsf{T},\end{aligned} \qquad \begin{aligned}\boldsymbol{\eta} = \big(&\eta(\boldsymbol{s}_1;t_1),\, \eta(\boldsymbol{s}_2;t_1),\, \ldots,\, \eta(\boldsymbol{s}_m;t_1),\\ &\eta(\boldsymbol{s}_1;t_2),\, \eta(\boldsymbol{s}_2;t_2),\, \ldots,\, \eta(\boldsymbol{s}_m;t_2),\\ &\ldots,\\ &\eta(\boldsymbol{s}_1;t_T),\, \eta(\boldsymbol{s}_2;t_T),\, \ldots,\, \eta(\boldsymbol{s}_m;t_T)\big)^\mathsf{T}.\end{aligned}$$

Primijetimo da su to sve vektori $n = m \cdot T$ duljine. Vektor očekivanja je linearna kombinacija $(p+1)$-dimenzionalnog vektora parametara $\boldsymbol{\beta}$ pa pišemo $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. $\boldsymbol{X}$ je $n \times (p+1)$

matrica vrijednosti dobivenih vektorskom funkcijom s tri varijable (prostorno-vremenske komponente), čija je kodomena $\mathbb{R}^{p+1}$, a definirana je na cijeloj domeni prostorno-vremenskih komponenata $\mathbb{D} \times \mathbb{N}$, tj. od svake uređene dvojke pripadnih prostornih lokacija i vremenskih oznaka $(\boldsymbol{s}_i; t_j)$, $\forall i = 1, \ldots, m$ i $j = 1, \ldots, T$, među danim opservacijama $\{z(\boldsymbol{s}; t)\}$. $\boldsymbol{X}$ u matričnom zapisu je dana s

$$\boldsymbol{X} = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1p} \\ x_{20} & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \ldots & x_{np} \end{bmatrix} = [x_{kl}]_{k=1,\ldots,n,\, l=0,\ldots,p}, \quad x_{kl} = x_l(a_k), \quad a_k = (\boldsymbol{s}_i; t_j),$$

pri čemu je

$$a_1 = (\boldsymbol{s}_1; t_1),\ a_2 = (\boldsymbol{s}_2; t_1),\ \ldots,\ a_m = (\boldsymbol{s}_m; t_1),\ a_{m+1} = (\boldsymbol{s}_1; t_2),\ a_{m+2} = (\boldsymbol{s}_2; t_2),\ \ldots,$$
$$a_{m+m} = (\boldsymbol{s}_m; t_2),\ \ldots,\ a_{m(T-1)+1} = (\boldsymbol{s}_1; t_T),\ a_{m(T-1)+2} = (\boldsymbol{s}_2; t_T),\ \ldots,\ a_{mT} = (\boldsymbol{s}_m; t_T)$$

niz oznaka od $m \cdot T = n$ članova, a $x_l(\boldsymbol{s}; t)$, $\forall l = 0, \ldots, p$, je linearna funkcija prostorno-vremenskih komponenti. Na primjer, $x_1(\boldsymbol{s}; t) = s_1 + s_2$, $x_2(\boldsymbol{s}; t) = s_1 + t$ ili $x_2(\boldsymbol{s}; t) = t$, pri čemu je $\boldsymbol{s} = (s_1, s_2)$, $\forall (\boldsymbol{s}; t) \in \mathbb{D} \times \mathbb{N}$. U većini slučajeva je $x_0(\boldsymbol{s}; t) = 1$, tj. slobodni član. Neka je zatim,

- $\boldsymbol{x}(\boldsymbol{s}_0; t_0)$ $(p+1)$-dimenzionalni vektor,
- $\boldsymbol{c}_0$ vektor kovarijanci ($1 \times n$ matrica kovarijanci) između slučajne varijable $Y(\boldsymbol{s}_0; t_0)$ i slučajnog procesa $\{Z(\boldsymbol{s}; t)\}$:

$$\boldsymbol{c}_0 = [Cov(Y(\boldsymbol{s}_0; t_0), Z(\boldsymbol{s}_i; t_j))]^{\intercal}_{i=1,\ldots,m,\, j=1,\ldots,T},$$

- $c_{0,0}$ varijanca slučajne varijable $Y(\boldsymbol{s}_0; t_0)$:

$$c_{0,0} = Var(Y(\boldsymbol{s}_0; t_0)),$$

- $\boldsymbol{C}_Z$, $\boldsymbol{C}_Y$, $\boldsymbol{C}_\varepsilon$, $\boldsymbol{C}_\eta$ funkcije kovarijanci procesa $\{Z(\boldsymbol{s}; t)\}$, $\{Y(\boldsymbol{s}; t)\}$, $\{\varepsilon(\boldsymbol{s}; t)\}$, $\{\eta(\boldsymbol{s}; t)\}$ tim redom, pri čemu je $\boldsymbol{C}_Z = \boldsymbol{C}_Y + \boldsymbol{C}_\varepsilon$ i $\boldsymbol{C}_Y = \boldsymbol{C}_\eta$,
- procesi $\{Z(\boldsymbol{s}; t)\}$ i $\{Y(\boldsymbol{s}; t)\}$ su Gaussovski procesi pa slijedi da slučajne varijable procesa $\{Y(\boldsymbol{s}; t)\}$ i $\{\varepsilon(\boldsymbol{s}; t)\}$ imaju normalnu distribuciju
- slučajni vektor $\left[\begin{smallmatrix} Y(\boldsymbol{s}_0; t_0) \\ Z \end{smallmatrix}\right]$ ima normalnu distribuciju[22], tj.

$$\begin{bmatrix} Y(\boldsymbol{s}_0; t_0) \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}(\boldsymbol{s}_0; t_0)^{\intercal} \\ \boldsymbol{X} \end{bmatrix} \boldsymbol{\beta}, \begin{bmatrix} c_{0,0} & \boldsymbol{c}_0^{\intercal} \\ \boldsymbol{c}_0 & \boldsymbol{C}_z \end{bmatrix} \right). \tag{4.3}$$

Zadatak nam je pronaći optimalnog prediktora za $Y(\boldsymbol{s}_0; t_0)$. Pokazat ćemo da je uvjetno očekivanje $\mathbf{E}[Y(\boldsymbol{s}_0; t_0) \,|Z]$ najbolji prediktor u srednje kvadratnom smislu, a onda ćemo pokazati da je i najbolji linearni prediktor.

[22] Slučajni vektor $\boldsymbol{X} = [X_1, \ldots, X_n]^{\intercal}$ ima **normalnu** ili **Gaussovu** distribuciju ako ima funkciju gustoće

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\intercal} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})},$$

pri čemu je $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^{\intercal} \in \mathbb{R}^n$ vektor očekivanja, a $\boldsymbol{\Sigma}$ $n \times n$ matrica kovarijanci koja je pozitivno definitna i simetrična. Takvu distribuciju označavamo s $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ili $\boldsymbol{X} \sim Gau(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Teorem 4.1. *Neka su Y slučajna varijabla, t.d. $\mathbf{E}[Y]^2 < \infty$ i $\boldsymbol{Z} = [Z_1, Z_2, \ldots, Z_n]^\intercal$ slučajni vektor. Tada je uvjetno očekivanje $m(\boldsymbol{Z}) = \mathbf{E}[Y|\boldsymbol{Z}]$ najbolji prediktor za Y na temelju $\boldsymbol{Z}$ u srednje kvadratnom smislu, odnosno za svaki prediktor $g(\boldsymbol{Z})$ za Y vrijedi*

$$\mathbf{E}[Y - g(\boldsymbol{Z})]^2 \geq \mathbf{E}[Y - m(\boldsymbol{Z})]^2$$

Dokaz.

$$\begin{aligned}
\mathbf{E}[Y - g(\boldsymbol{Z})]^2 =& \mathbf{E}[Y - m(\boldsymbol{Z}) + m(\boldsymbol{Z}) - g(\boldsymbol{Z})]^2 = \\
&= \mathbf{E}[Y - m(\boldsymbol{Z})]^2 + \mathbf{E}[m(\boldsymbol{Z}) - g(\boldsymbol{Z})]^2 + 2\mathbf{E}[(Y - m(\boldsymbol{Z}))(m(\boldsymbol{Z}) - g(\boldsymbol{Z}))] = \\
&= \mathbf{E}[Y - m(\boldsymbol{Z})]^2 + \underbrace{\mathbf{E}[m(\boldsymbol{Z}) - g(\boldsymbol{Z})]^2}_{\geq 0} + 2(m(\boldsymbol{Z}) - g(\boldsymbol{Z}))\underbrace{\mathbf{E}[Y - m(\boldsymbol{Z})]}_{(\star)} = \\
&\geq \mathbf{E}[Y - m(\boldsymbol{Z})]^2
\end{aligned}$$

U $(\star)$ smo iskoristili svojstvo uvjetnog očekivanja da slučajna varijabla $Y - m(\boldsymbol{Z})$ i uvjetno očekivanje $\mathbf{E}[(Y - m(\boldsymbol{Z}))|\boldsymbol{Z}]$ imaju ista očekivanja pa zatim svojstvo da je uvjetno očekivanje $\mathbf{E}[(Y - m(\boldsymbol{Z}))|\boldsymbol{Z}]$ linearno:

$$(\star) = \mathbf{E}[Y - m(\boldsymbol{Z})] = \mathbf{E}\big[\mathbf{E}[(Y - m(\boldsymbol{Z}))|\boldsymbol{Z}]\big] = \mathbf{E}\big[\mathbf{E}[Y|\boldsymbol{Z}] - \mathbf{E}[Y|\boldsymbol{Z}]\big] = 0$$

□

Kako bi dalje odredili uvjetnu distribuciju slučajne varijable $Y(\boldsymbol{s}_0; t_0)$ uz uvjet slučajnog procesa Z, potrebno je iskazati neka svojstva multivarijatne normalne distribucije.

Teorem 4.2. *Neka je $\boldsymbol{X}$ slučajni vektor koji ima multivarijatnu normalnu distribuciju, tj.*

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Za $k < n$ promotrimo particije od $\boldsymbol{X}$, $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1,1)} & \boldsymbol{\Sigma}^{(1,2)} \\ \boldsymbol{\Sigma}^{(2,1)} & \boldsymbol{\Sigma}^{(2,2)} \end{bmatrix},$$

gdje su

$$\begin{aligned}
\boldsymbol{X}^{(1)} &= (X_1, \ldots, X_k)^\intercal & \boldsymbol{X}^{(2)} &= (X_{k+1}, \ldots, X_n)^\intercal \\
\boldsymbol{\mu}^{(1)} &= (\mu_1, \ldots, \mu_k)^\intercal & \boldsymbol{\mu}^{(2)} &= (\mu_{k+1}, \ldots, \mu_n)^\intercal,
\end{aligned}$$

$\boldsymbol{\Sigma}^{(i,i)}$ je matrica kovarijanci vektora $\boldsymbol{X}^{(i)}$, $i = 1, 2$, $\boldsymbol{\Sigma}^{(1,2)} = [Cov(X_i, X_j)]$, $\forall i = 1, \ldots, k$, $j = k+1, \ldots, n$, $i < j$ i $\boldsymbol{\Sigma}^{(2,1)} = (\boldsymbol{\Sigma}^{(1,2)})^\intercal$. Tada uvjetna distribucija slučajnog vektora $\boldsymbol{X}^{(1)}$ uz uvjet $\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}$ ima multivarijatnu normalnu distribuciju

$$\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}}, \boldsymbol{\Sigma}_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}})$$

s vektorom uvjetnog očekivanja

$$\boldsymbol{\mu}_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}} = \mathbf{E}[\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}] = \boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}^{(1,2)}(\boldsymbol{\Sigma}^{(2,2)})^{-1}(\boldsymbol{x}^{(2)} - \boldsymbol{\mu}^{(2)})$$

i matricom uvjetnih kovarijanci

$$\boldsymbol{\Sigma}_{\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}} = Cov(\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)} = \boldsymbol{x}^{(2)}) = \boldsymbol{\Sigma}^{(1,1)} - \boldsymbol{\Sigma}^{(1,2)}(\boldsymbol{\Sigma}^{(2,2)})^{-1}\boldsymbol{\Sigma}^{(2,1)}.$$

Dokaz. Dokaz se može naći na [14]. □

Iz ovog teorema slijedi da je vektor uvjetnog očekivanja linearna funkcija od uvjeta, što nam omogućuje određivanje optimalnog prediktora.

Iz teorema 4.2 i (4.3) nam slijedi da je uvjetna distribucija od $Y(\boldsymbol{s}_0; t_0)$ uz uvjet $Z = \boldsymbol{z}$, $\boldsymbol{z}$ je realizacija procesa Z, također normalna, tj.

$$Y(\boldsymbol{s}_0; t_0) \mid Z = \boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{x}(\boldsymbol{s}_0; t_0)^\intercal \boldsymbol{\beta} + \boldsymbol{c}_0^\intercal \boldsymbol{C}_Z^{-1}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}), c_{0,0} - \boldsymbol{c}_0^\intercal \boldsymbol{C}_Z^{-1}\boldsymbol{c}_0\right), \tag{4.4}$$

s vektorom uvjetnog očekivanja

$$\mathbf{E}[Y(\boldsymbol{s}_0; t_0) | Z = \boldsymbol{z}] = \boldsymbol{x}(\boldsymbol{s}_0; t_0)^\intercal \boldsymbol{\beta} + \boldsymbol{c}_0^\intercal \boldsymbol{C}_Z^{-1}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) \tag{4.5}$$

i matricom uvjetnih kovarijanci

$$Cov(Y(\boldsymbol{s}_0; t_0) | Z = \boldsymbol{z}) = c_{0,0} - \boldsymbol{c}_0^\intercal \boldsymbol{C}_Z^{-1}\boldsymbol{c}_0. \tag{4.6}$$

Prema tome, vidimo da je vektor uvjetnog očekivanja (4.5) linearna funkcija od $\boldsymbol{z}$ pa smo pronašli optimalnog prediktora za $Y(\boldsymbol{s}_0; t_0)$ i označimo ga s $\widehat{Y}(\boldsymbol{s}_0; t_0)$. Varijanca kriging prediktora je minimizirana srednje kvadratna greška predikcije, ali uz pretpostavku normalne distribucije (4.3), to je zapravo matrica uvjetnih kovarijanci 4.6 i označit ćemo ju sa $\sigma^2(\boldsymbol{s}_0; t_0)$.

Primjedba 4.1. Uvjetno očekivanje 4.5 se sastoji od:

1. marginalnog očekivanja slučajne varijable $Y(\boldsymbol{s}_0; t_0)$, $\mathbf{E}[Y(\boldsymbol{s}_0; t_0)] = \boldsymbol{x}(\boldsymbol{s}_0; t_0)^\intercal \boldsymbol{\beta}$ i
2. reziduala, $\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}$, između vektora podataka $\boldsymbol{z}$ i marginalnog očekivanja procesa Z, $\mathbf{E}[Z] = \boldsymbol{X}\boldsymbol{\beta}$, pri čemu su tim rezidualima dodane težine, $\boldsymbol{w}^\intercal = \boldsymbol{c}_0^\intercal \boldsymbol{C}_Z^{-1}$

Primjedba 4.2. Stavimo li matricu $\boldsymbol{X}$ u kontekst regresije, $\boldsymbol{X}$ predstavlja matricu dizajna, a funkcije $x_l(\boldsymbol{s}; t)$, $\forall l = 0, \ldots, p$ prostorno-vremenske prediktore u svim prostorno-vremenskim točkama $(\boldsymbol{s}; t) \in \mathbb{D} \times \mathbb{N}$.

U stvarnosti gotovo nikad nećemo znati $\boldsymbol{C}_Z$, $\boldsymbol{c}_0$ i $c_{0,0}$. Stoga nam je potrebno modelirati funkciju kovarijanci $\boldsymbol{C}_Z$ slučajnog procesa $Z = \{Z(\boldsymbol{s}; t)\}$ kako bismo mogli odrediti kriging prediktora što dovodi do modeliranja pripadnog variograma.

4.2 Modeliranje variograma

U ovom poglavlju pokušavamo modelirati prostorno-vremensku zavisnost otkrivenu u variogramu ili funkciji kovarijanci. Tradicionalni način pronalaženja prikladnog modela variograma je prilagođavanje parametarskog modela empirijskom variogramu (3.10). Kod implementacije klasičnog kriginga se inače pretpostavlja stacionarnost drugog reda (definicija 3.2), tj. da slučajni proces ima konstantno očekivanje i funkciju kovarijance koja se može izraziti u smislu pomaka u prostoru i razlika u vremenu.

Postoji nekoliko mogućih modela koji se razlikuju po načinu na koji kombiniraju vremensku i prostornu zavisnost. U R paketu ***gstat*** imamo pet modela funkcija kovarijanci, a to su separabilni model (engl. *„separable"*), model produkta i sume (engl. *„product-sum"*), metrički model (engl. *„metric"*), model metričke sume (engl. *„sum-metric"*) i model jednostavne metričke sume (engl. *„simple sum-metric"*). U ovom radu ćemo razmotriti tri modela od ovih nabrojanih:

- separabilni model - sastoji se od umnoška prostorne i vremenske funkcije kovarijanci, što daje strukturu koja nema prostorno-vremensku interakciju jer se može rastaviti na faktore pa ima samo multiplikativnu interakciju,
- metrički model - model zajedničke prostorno-vremenske funkcije kovarijanci, pri čemu se vrijeme smatra još jednom komponentom pa ju moramo skalirati kako bi odgovarala prostornoj komponenti,
- i model metričke sume - model sume prostorne, vremenske i zajedničke prostorno-vremenske funkcije kovarijanci uključujući skaliranje vremenske komponente.

U literaturi postoji velik broj klasa prostorne i vremenske funkcije kovarijanci, na primjer eksponencijalna klasa, klasa Matérn, Gaussova klasa, i klasa opća potencija. Klasa prostorne funkcije kovarijanci ne mora biti ista kao i vremenska funkcija kovarijanci. Mi možemo uzeti neke osnovne za primjer, najvjerojatnije nećemo dobiti najbolji model. No ovdje nam je trenutno cilj pokazati postupak modeliranja funkcije kovarijanci, odnosno variograma. Klase modela i klase prostorno-vremenskih funkcija kovarijanci koje ne obradimo, a i više detalja o onima koje jesmo obradili u radu mogu se pogledati u [5].

4.2.1 Separabilni model

Prvi model variograma koji ovdje razmatramo odgovara prostorno-vremenskoj separabilnoj funkciji kovarijanci. Separabilni model pretpostavlja da se prostorno-vremenska funkcija kovarijanci može prikazati kao produkt prostorne i vremenske funkcije kovarijanci

$$C^{(\text{sep})}(\|\boldsymbol{h}\|;|\tau|) = C^{(\boldsymbol{s})}(\|\boldsymbol{h}\|) \cdot C^{(t)}(|\tau|) \tag{4.7}$$

odnosno, pripradni prostorno-vremenski separabilni variogram je dan s

$$\gamma^{(\text{sep})}(\boldsymbol{h};\tau) = \text{sill} \cdot \left(\bar{\gamma}^{(\boldsymbol{s})}(\|\boldsymbol{h}\|) + \bar{\gamma}^{(t)}(|\tau|) - \bar{\gamma}^{(\boldsymbol{s})}(\|\boldsymbol{h}\|)\bar{\gamma}^{(t)}(|\tau|)\right), \tag{4.8}$$

gdje su $\bar{\gamma}^{(\boldsymbol{s})}$ i $\bar{\gamma}^{(t)}$ standardizirani prostorni i vremenski variogrami, a *„sill"* označava prag prostorno-vremenskog variograma.

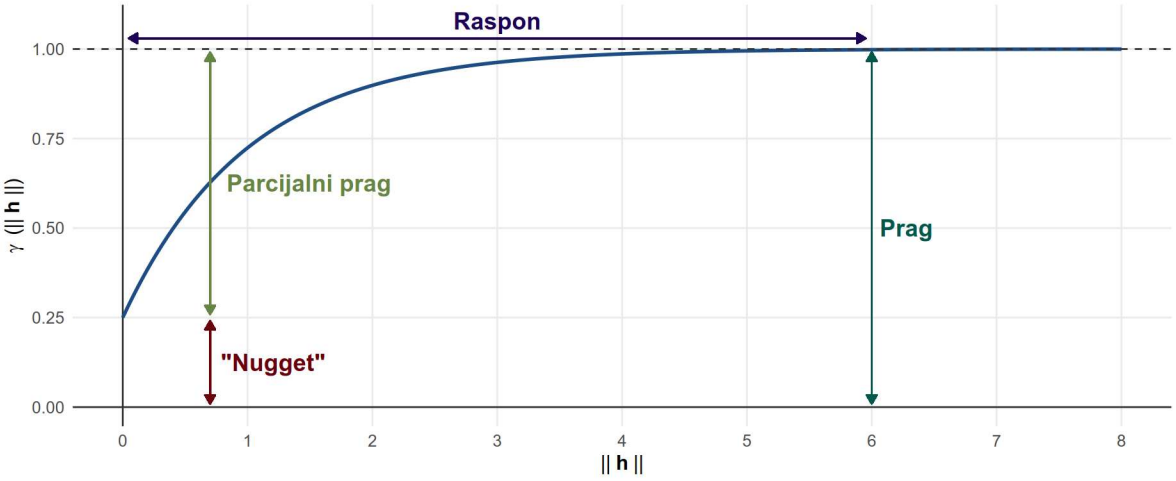
Inače se teorijski (prostorni) variogram sastoji od sljedećih parametara:

- već spomenutog **praga** - horizontalna asimptota eksponencijalne funkcije (teorijskog variograma),
- **„nugget"-a** - iznos $\bar{\gamma}^{(\boldsymbol{s})}(\|\boldsymbol{h}\|)$ kada $\boldsymbol{h} \to \boldsymbol{0}$, što znači iznos variograma za jako male pomake u prostoru. Analogno, „nugget" u modelu vremenskog variograma je iznos $\bar{\gamma}^{(t)}(|\tau|)$ kada $\tau \to 0$, a prostorno-vremenski „nugget" je iznos $\gamma^{(\text{sep})}(\boldsymbol{h};\tau)$ kada $\boldsymbol{h} \to \boldsymbol{0}$ i $\tau \to 0$,
- **raspona** - udaljenost gdje se postiže prag, tj. gdje horizontalna asimptota siječe ordinatu i
- **parcijalnog praga** - razlika između praga i „nugget"-a.

Analogno se pojmovi ovih parametara definiraju i za teorijski prostorno-vremenski variogram. Najčešće klase funkcija[23] koje se koriste kod teorijskih variograma su eksponencijalna, sferna i Gaussova funkcija.

[23]Detalji se mogu pogledati na `https://www.supergeotek.com/Spatial_Statistical_ENG_HTML/index.html?spherical_mode.htm`

Slika 4.1 prikazuje ove parametre na teorijskom prostornom variogramu koji pripada eksponencijalnoj klasi kao funkcija od prostornih pomaka $\|\boldsymbol{h}\|$.



Slika 4.1: Teorijski prostorni variogram eksponencijalne klase.

Na primjeru GSOD podataka smo napravili empirijski (engl. *„sample"*) variogram (slika 3.8). Želimo naći model koji će najbolje opisati taj variogram. Za teorijski variogram će nam biti potrebno sedam parametara, za variogram svake od komponenata trebat će po dva parcijalna praga, po dva raspona i po dva „nugget"-a te prostorno-vremenski prag. Prostorno-vremenski prag je zapravo zamijenio zasebni prostorni i vremenski prag te se može procijeniti iz empirijskog prostorno-vremenskog variograma kao 80%-tni kvantil dobivenih vrijednosti, a u našim podacima iznosi 26.61. Dobijemo ga u R na sljedeći način:

```
(estimated.sill <- quantile(na.omit(var$gamma), .8))
```

Zatim odredimo teorijski separabilni variogram funkcijom `vgmST()` kojoj zadamo željeni model (separabilni u ovom slučaju), kako da izgledaju teorijski prostorni i vremenski variogrami te parametar praga kojeg smo procijenili. Teorijski prostorni, odnosno, vremenski variogram određujemo funkcijom `vgm()` i dajemo joj parametre parcijalnog praga ($1-nugget$), klasu funkcije kovarijanci (eksponencijalna u ovom slučaju), raspon i „nugget" (većinom ne bude veći od 0.1). Za odabir raspona je bitno samo da je pozitivan. Mi ćemo postaviti neke inicijalne vrijednosti, a onda ćemo kasnije vidjeti koje ćemo optimalne vrijednosti dobiti algoritmom u R.

```
sepVar <- vgmST(stModel = "separable",
                space = vgm(0.9, "Exp", 500, 0.1),
                time = vgm(0.9, "Exp", 1, 0.1),
                sill = estimated.sill)
```

Takav teorijski separabilni variogram ćemo iskoristiti za prilagodbu našem empirijskom variogramu R funkcijom `fit.StVariogram()`. Kako bi bili sigurni da optimizacijski algoritam ne bi vratio negativne parametre, funkciji `fit.StVariogram()` ćemo dati dodatne parametre `lower` i `upper` kao donju i gornju granicu parametara teorijskog variograma. R kod nam daje sljedeći *„output"*:

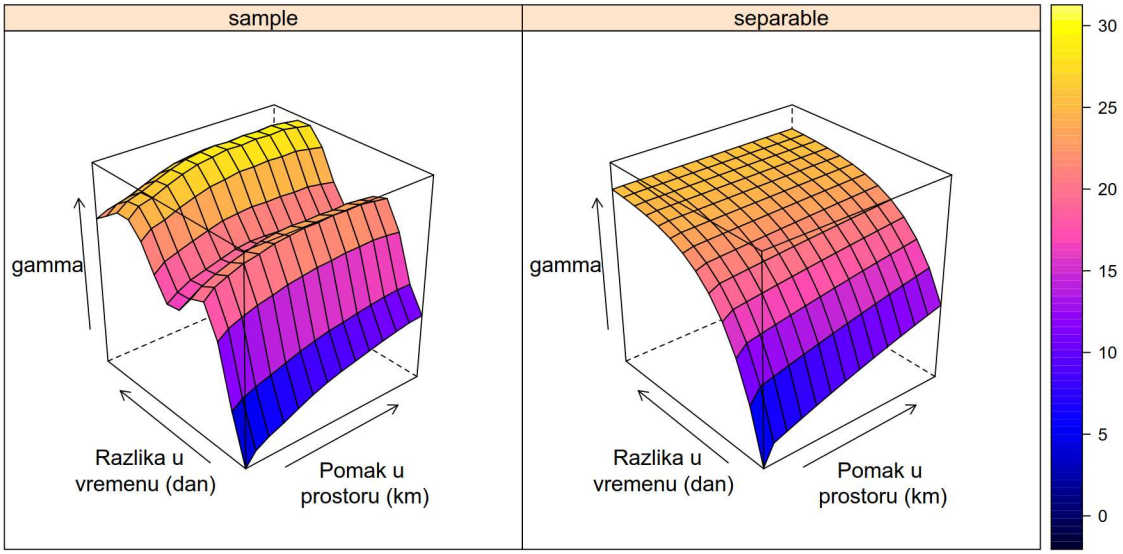
```
(sepVarFit <-
    fit.StVariogram(var, sepVar,
                    #       c(space range, space nugget, time range, time nugget, sill)
                    lower = c(     10,           0,          0.1,          0,        0.1),
                    upper = c(    2000,          1,           12,          1,        200)))

# output:
space component:
  model      psill range
1   Nug 0.09041858     0
2   Exp 0.90958142  1100
time component:
  model psill    range
1   Nug     0 0.000000
2   Exp     1 2.863111
sill: 25.9664784394401
```

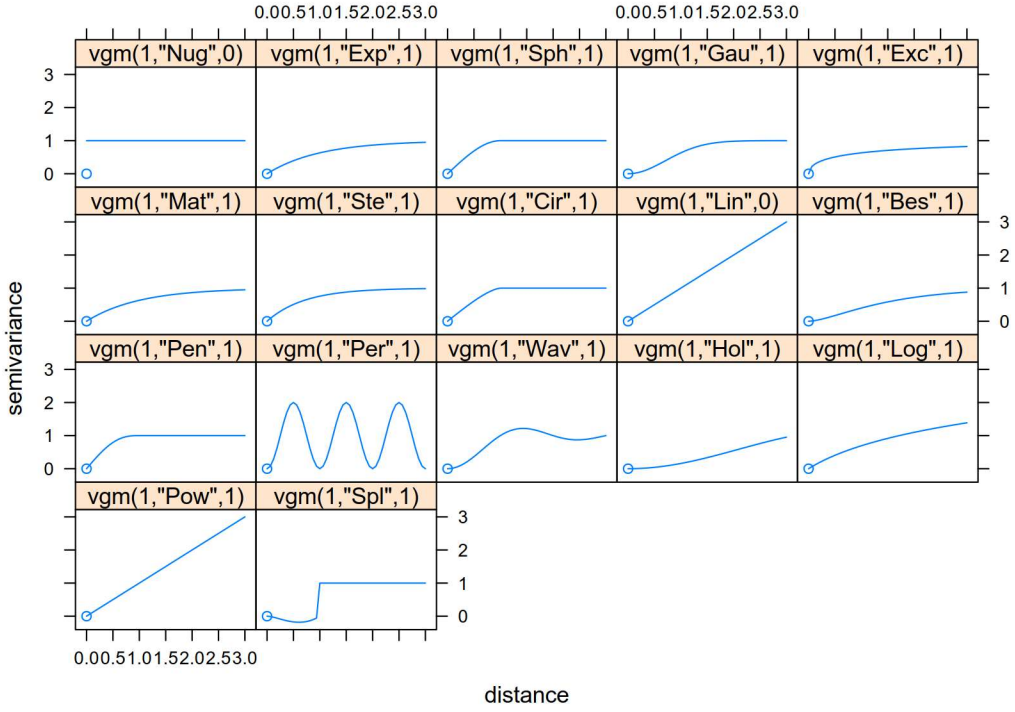
Vidimo da prostorno-vremenski prag ipak nije 26.61 nego 25.97. Da smo postavili bilo koji drugi pozitivan broj umjesto kvantila, opet bi nam algoritam vratio isti *„output"*. I raspone i „nugget"-e smo dobili drugačije od inicijalnih.

Primjedba 4.3. Nakon ručne provjere srednje kvadratne greške modela (MSE) pri izmjeni parametara, može se reći da odabir parametara u `vgm()` neće utjecati na MSE nakon prilagodbe modela empirijskom variogramu pa nema onda ni smisla zamarati se dodatno time nego procjenu parametara treba prepustiti funkciji `fit.StVariogram()` koja u sebi poziva funkciju `optim()` za numeričku optimizaciju parametara. Potrebno je samo postaviti pozitivne raspone i pragove kako R ne bi vraćao grešku.

Na sljedećoj slici 4.2 vidimo empirijski variogram i njemu najbolje prilagođeni separabilni model variograma u smislu optimiziranih parametara i najmanje srednje kvadratne greške modela nakon unosa svih klasa variograma definiranih u R paketu ***gstat*** (slika 4.3).



Slika 4.2: Lijevo je empirijski variogram dnevnih maksimalnih temperatura u srpnju 2022., a desno je njemu prilagođen teorijski variogram dobiven separabilnim modelom.



Slika 4.3: Zadani prikaz klasa variograma definiranih u R paketu ***gstat***.

4.2.2 Metrički model

Drugi model koji ćemo prilagoditi (engl. *„fit"*) našim podacima ima zajedničku prostorno-vremensku funkciju kovarijanci, pri čemu se vremenska komponenta treba skalirati kako bi se vremenski pomak mogao koristi kao varijabla u funkciji kovarijanci, a da pri tome ne bi previše ili premalo utjecao na cjelokupnu funkciju. Vremenska komponenta je anizotropna, što znači da je zavisnost kroz vrijeme na drugačijoj skali od one kroz prostor (u ovom slučaju dvodimenzionalni). Pretpostavljena struktura funkcije kovarijanci u ovom modelu je:

$$C^{(\mathrm{m})}(\boldsymbol{h};\tau) = C^{(\mathrm{zaj})}\left(\sqrt{\boldsymbol{h}^{\intercal}\boldsymbol{h} + (\kappa\tau)^2}\right),$$

i pripadni variogram je dan s

$$\gamma^{(\mathrm{m})}(\mathbf{h};\tau) = \gamma^{(\mathrm{zaj})}\left(\sqrt{\boldsymbol{h}^{\intercal}\boldsymbol{h} + (\kappa\tau)^2}\right),$$

pri čemu je κ faktor za skaliranje vremenske komponente, dan kao prostorni jedinični pomak po vremenskoj jediničnoj razlici. Ako variogram kroz prostor promatramo po pomacima od 100 kilometara, a kroz vrijeme po razmacima od jedan dan, tada za parametar κ možemo postaviti vrijednost 100, u R se κ označava sa `stAni`. Osim parametra κ modelu dajemo i vrijednosti parametara praga, raspona i „nugget"-a zajedničkog teorijskog prostorno-vremenskog variograma. Znači da bi se metrički model prilagodio empirijskom variogramu, za procjenu parametara ima tri parametra manje od separabilnog modela.

R kod je analogan separabilnom modelu, samo što sada imamo samo jedan variogram za definirati, zajednički. Kod separabilnog smo imali zasebni prostorni i vremenski variogram za definirati u funkciji `vgmST()`. Parametre teorijskog variograma zapravo pogađamo dok ne dobijemo graf sličan grafu empirijskog variograma. Svakako će algoritam dati optimalne

vrijednosti. Prije nagađanja vrijednosti raspona i anizotropije, možemo uzeti da je parcijalni prag maksimum od variograma u njegovoj najvećoj vremenskoj razlici, a „nugget" vidimo iz 3.8 da je 0.

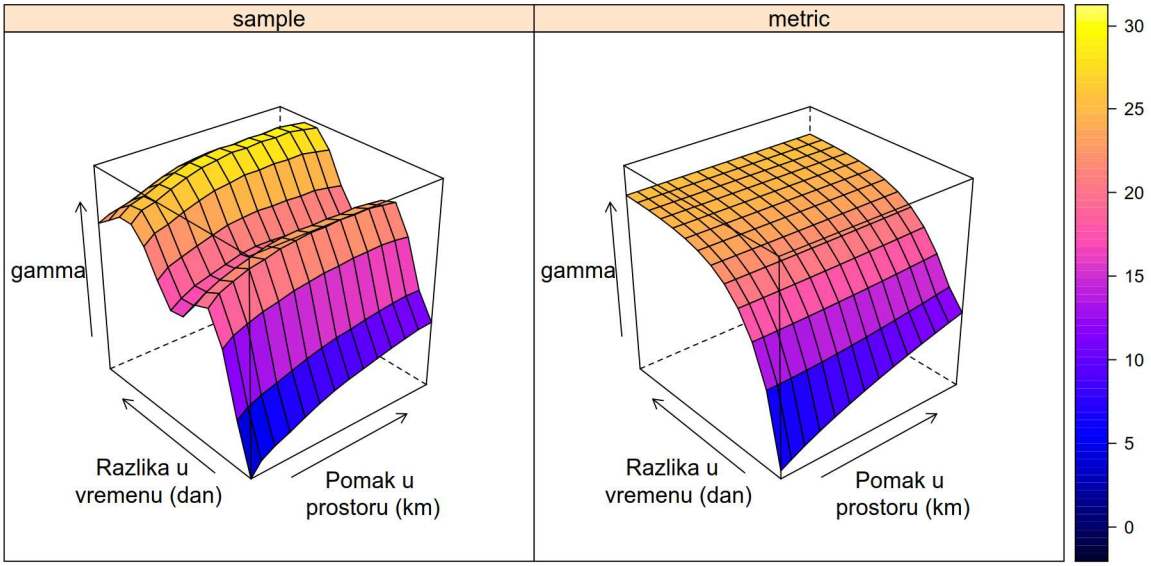
```
(estimated.psill <- max(var[var$id=="lag12",]$gamma))
metricVar <- vgmST(stModel = "metric",
                   joint = vgm(estimated.psill, "Exp", 1000, nugget = 0),
                   stAni = 500)
```

Pozovemo funkciju `fit.StVariogram()` pa dobijemo optimalne vrijednosti parametara.

```
(metricVarFit <- fit.StVariogram(var, metricVar))

# output:
joint component:
  model     psill    range
1   Nug  1.688473    0.000
2   Exp 23.515366 1005.125
stAni: 487.741071117858
```

Na slici 4.4 vidimo usporedbu empirijskog variograma dnevnih maksimalnih temperatura u Europi u 2022. godini i metričkog modela prilagođenog tom empirijskom variogramu.



Slika 4.4: Lijevo je empirijski variogram dnevnih maksimalnih temperatura u srpnju 2022., a desno je njemu prilagođen teorijski variogram dobiven metričkim modelom.

4.2.3 Model metričke sume

Kombinacija prostornog, vremenskog i prostorno-vremenskog metričkog modela, uključujući parametar anizotropije κ, čini model funkcije kovarijanci dan s

$$C^{(\mathrm{ms})}(\boldsymbol{h};\tau) = C^{(\boldsymbol{s})}(\|\mathbf{h}\|) + C^{(t)}(|\tau|) + C^{(\mathrm{zaj})}\left(\sqrt{\boldsymbol{h}^{\intercal}\boldsymbol{h} + (\kappa\tau)^2}\right),$$

Pripadni variogram je

$$\gamma^{(\mathrm{ms})}(\mathbf{h};\tau) = \gamma^{(\boldsymbol{s})}(\|\mathbf{h}\|) + \gamma^{(t)}(|\tau|) + \gamma^{(\mathrm{zaj})}\left(\sqrt{\boldsymbol{h}^{\intercal}\boldsymbol{h} + (\kappa\tau)^2}\right),$$

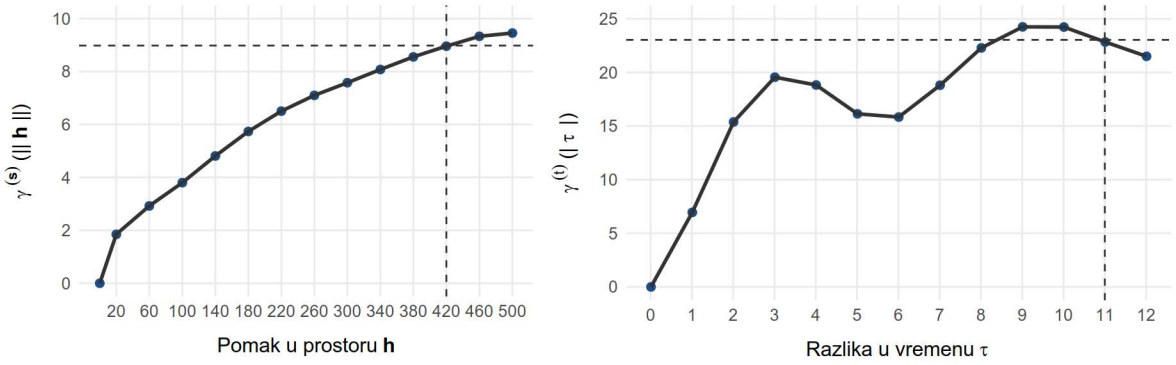
gdje su $\gamma^{(\boldsymbol{s})}$, $\gamma^{(t)}$ i $\gamma^{(\text{zaj})}$ prostorni, vremenski i zajednički prostorno-vremenski variogram tim redom. Svaki od ovih variograma ima svoj raspon, parcijalni prag i „nugget". U ovom modelu još procjenjujemo i zajednički parametar anizotropije κ. Kao i dosadašnja dva modela, pogledajmo kako ga dobiti uz pomoć R koda. Kako bi odredili vrijednosti parametara u ovom modelu, odredimo prvo prostorni variogram s vremenskom razlikom 0 i vremenski variogram s prostornim pomakom 0 iz prostorno-vremenskog variograma.

```
(v.sp <- var[var$id == "lag0", c("spacelag", "gamma")])
(v.t <- var[var$spacelag == 0, c("timelag", "gamma")])
```

Budući da koristimo eksponencijalnu klasu za izradu modela variograma, dat ćemo parametru raspona vrijedost od $\frac{1}{3}$ efektivnog raspona kod sva tri variograma, $\gamma^{(\boldsymbol{s})}$, $\gamma^{(t)}$ i $\gamma^{(\text{zaj})}$. Efektivni raspon je udaljenost na kojoj variogram doseže 95% svog maksimuma, [7, poglavlje 4.3] pa ih izračunajmo uz pomoć R.

```
(max(v.sp$gamma)*0.95) #8.982043
(max(v.t$gamma)*0.95) #23.03572
```

Promatrajući lijevi graf 4.5 prostornog variograma s vremenskom razlikom 0, vidimo da se empirijska varijanca od 8.98, što je prostorni parcijalni prag, postiže u prostornom pomaku od 420 km, a promatrajući desni graf na istoj slici vremenskog variograma s prostornim pomakom 0, vidimo da se empirijska varijanca od 23.04, što je vremenski parcijalni prag, postiže u razlici u vremenu od 11 dana.



Slika 4.5: Lijevo je prostorni variogram s vremenskom razlikom 0, a desno vremenski variogram s prostornim pomakom 0 iz prostorno-vremenskog empirijskog variograma.

Što znači da ćemo za parametar raspona u teorijskom prostornom variogramu postaviti na vrijednost $\frac{420}{3} = 140$, a u vremenskom $\frac{11}{3} = 3.67$. Parametar anizotropije je u ovom slučaju omjer prostornog i vremenskog raspona, što znači $\frac{140}{3.67} = 38.18$.

Promatrajući sliku 3.8b paralelnih prostornih variograma, jedan po razlici u vremenu τ, vidimo da svi variogrami imaju konstantnu varijancu nakon pomaka $\boldsymbol{h} = 200$ pa će nam teorijski zajednički variogram poprimiti vrijednost $\frac{200}{3} = 66.67$ u parametru raspona. Parcijalni prag u tom slučaju iznosi 6 jer toliko otprilike iznosi prostorni variogram s vremenskom razlikom 0 u pomaku 200. „Nugget"-e ćemo postaviti na nulu jer toliko iznose u empirijskim variogramima.

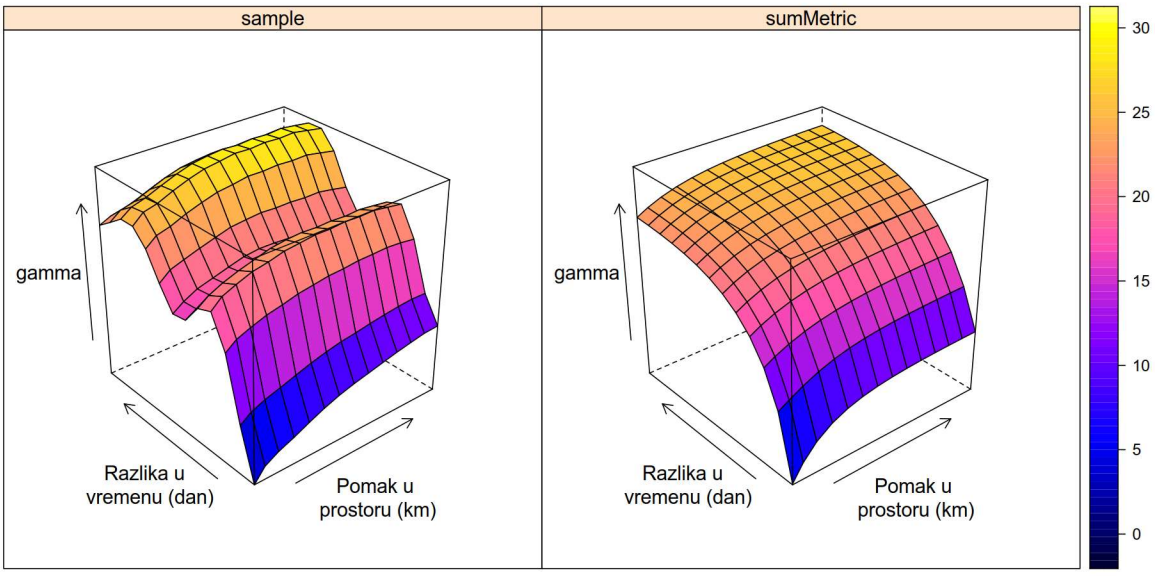
```
sumMetricVar <- vgmST(stModel = "sumMetric",
                      space = vgm(0.95*max(v.sp$gamma), "Exp", 420/3, 0),
                      time = vgm(0.95*max(v.t$gamma), "Exp", 11/3, 0),
                      joint = vgm(6, "Exp", 200/3, 0),
                      stAni = 420/11)
```

Opet pozivamo funkciju `fit.StVariogram()` kako bi model prilagodili empirijskom variogramu pa dobijemo optimalne vrijednosti parametara.

```
(sumMetricVarFit <- fit.StVariogram(var, sumMetricVar))

# output:
space component:
  model    psill   range
1   Nug 0.000000   0.000
2   Exp 4.142056 126.464
time component:
  model    psill    range
1   Nug  0.00000 0.000000
2   Exp 18.35086 2.881774
joint component:
  model    psill    range
1   Nug 0.000000  0.00000
2   Exp 4.496638 67.42674
stAni: 94.2895597628096
```

Na slici 4.6 vidimo usporedbu empirijskog variograma dnevnih maksimalnih temperatura u Europi u 2022. godini i modela metričke sume prilagođenog tom empirijskom variogramu.



Slika 4.6: Lijevo je empirijski variogram dnevnih maksimalnih temperatura u srpnju 2022., a desno je njemu prilagođen teorijski variogram dobiven modelom metričke sume.

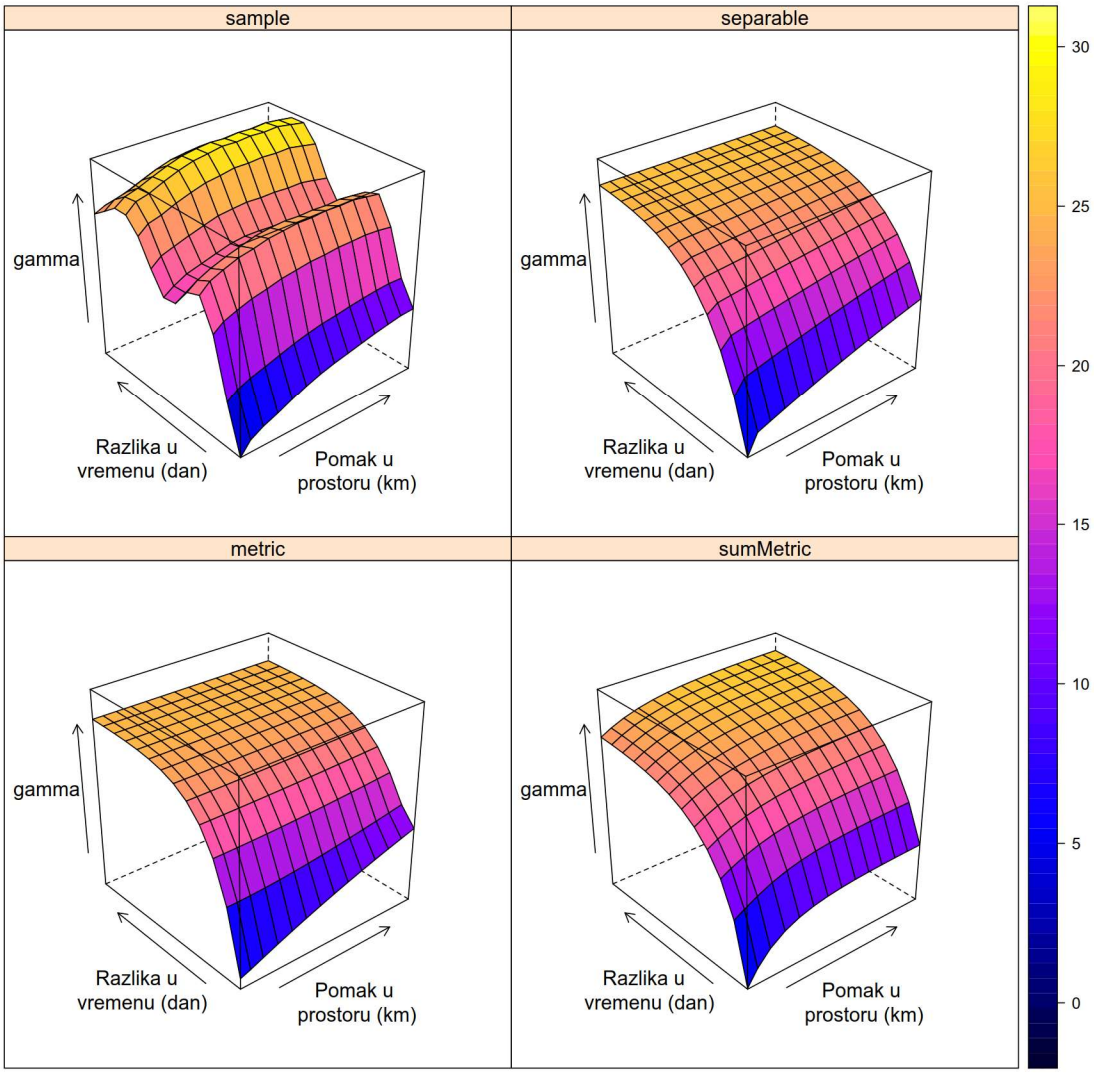
4.2.4 Odabir modela

Kako bi odredili najbolji model u smislu koji bi bolje opisao empirijski variogram, postoji metrika koja nam može pomoći u odluci. Možemo usporediti srednje kvadratne greške (MSE)

od svakog modela nakon što model prilagodimo empirijskom variogramu i vidjeti koji će nam dati najmanju vrijednost. MSE lako izračunamo u R pa pogledajmo.

```
(sepMSE <- attr(sepVarFit, "optim")$value)
(metricMSE <- attr(metricVarFit, "optim")$value)
(sumMetricMSE <- attr(sumMetricVarFit, "optim")$value)
```

Separabilni model u našem primjeru nam je vratio 7.03, metrički model 7.70, a model metričke sume 6.09, što nam ukazuje na to da je model metričke sume bolji od ostala dva. Pogledajmo grafičku usporedbu na slici 4.7.

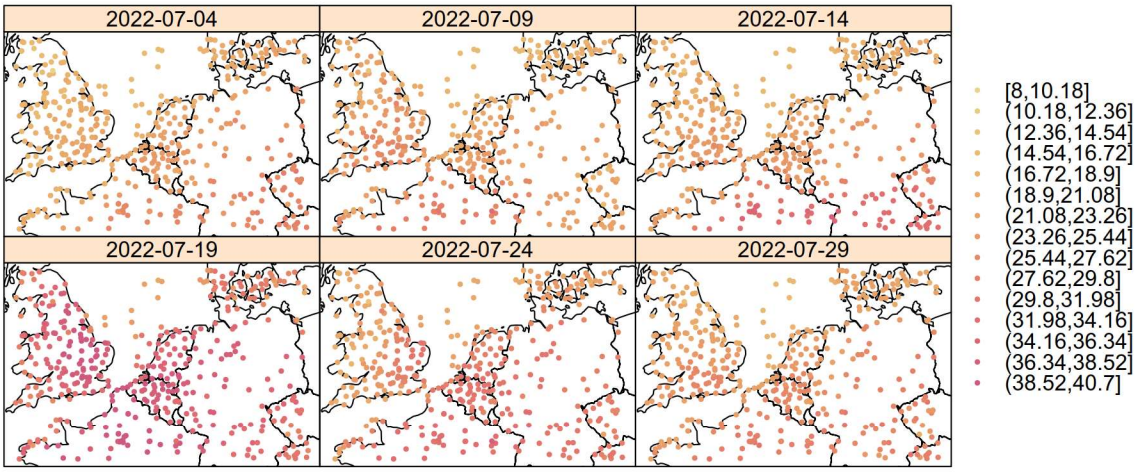


Slika 4.7: Lijevo gore je empirijski variogram dnevnih maksimalnih temperatura u srpnju 2022., desno gore je njemu prilagođen teorijski variogram dobiven separabilnim modelom, lijevo dolje metričkim modelom, a desno dolje modelom metričke sume.

Sada kada imamo model variograma, potrebno je još napraviti rešetku na kojoj ćemo primijeniti interpolaciju pa smo spremni za kriging.

4.2.5 Rešetka za interpolaciju

Za potrebe ovog rada, ranije smo se ograničili na „centar" Europe (od 5° zapadne zemljopisne dužine do 15° istočne zemljopisne dužine i od 48° sjeverne zemljopisne širine do 56° sjeverne zemljopisne širine). Na ovom području imamo 433 lokacije na kojima su postavljene meteorološke stanice, kao što vidimo i na slici 4.8, tj. nemamo stanice baš na svakom djeliću promatranog područja, a nisu niti podjednako udaljene jedna od druge tako da nemamo regularnu rešetku.



Slika 4.8: Stvarni GSOD podaci dnevnih maksimalnih temperatura na području „centra" Europe u šest jednako razmaknutih dana u srpnju na 433 lokacije postavljenih meteoroloških stanica.

Stoga nam je sada zadatak napraviti prostornu regularnu rešetku kojoj ćemo kasnije dodati vrijednosti interpolacije.

```
spat_grid <- expand.grid(
  LONGITUDE = seq(-5, 15, length = 20),
  LATITUDE = seq(48, 56, length = 20)) %>%
  SpatialPoints(proj4string = CRS(proj4string(STFDF_Eu)))
gridded(spat_grid) <- TRUE
```

Zbog zahtjevnosti izvođenja kod-a u R-u, interpolaciju, odnosno, kriging ćemo izvesti za šest vremenskih točaka.

```
temp_grid <- as.Date("2022-07-01") + seq(3, 28, length = 6)
```

Ova dobivena klasa je vektor od šest elemenata. Zatim prostornu rešetku i vremenski vektor povežemo u prostorno-vremensku rešetku ***spacetime*** funkcijom u R `STF()` koja nam daje raspored prostorno-vremenskih podataka u punu mrežu (rešetku) i ovu klasu zovemo STF.

```
STF_EU <- STF(sp = spat_grid,     # prostorna rešetka
              time = temp_grid)  # vremenski vektor
```

Dobili smo regularnu prostornu rešetku koju ćemo kasnije popuniti vrijednostima interpolacije. Za kriging će nam biti još potrebna iregularna prostorno-vremenska struktura vrijednosti temperatura na točnim lokacijama i vremenima u kojima su zabilježeni. Ta struktura podataka neće imati „**NA**" vrijednosti (vrijednosti koje nisu dostupne, odnosno "prazne vrijednosti") pa je zato iregularna. Ovakva struktura nema zabilježene vrijednosti na jednako udaljenim lokacijama po geografskoj širini i dužini. Što znači da prelazimo iz STFDF klase u STIDF klasu pri čemu ćemo jednu vremensku točku (14. srpnja 2022.) izbaciti tako da vidimo točnost kriginga u tom danu.

```
# Uzimamo podskup srpanj 2022.
STFDF_EuJuly <- STFDF_Eu[, "2022-07-01::2022-07-31"]
str(STFDF_EuJuly)

STIDF_EuJuly <- as(STFDF_EuJuly[, -14], "STIDF") # konverzija STFDF klase u STIDF klasu
STIDF_EuJuly <- subset(STIDF_EuJuly, !is.na(STIDF_EuJuly$MAX)) # izbacujemo NA vrijednosti
```

Sada imamo sve spremno za prostorno-vremenski kriging.

4.3 Jednostavni kriging

Vrsta kriginga kod kojeg je $(m \cdot T)$-dimenzionalni vektor $\boldsymbol{\mu}$ poznat na svakoj lokaciji u prostoru $\boldsymbol{s} \in \mathbb{D}$ i u svakom vremenskom trenutku $t \in \mathbb{N}$. Ovo je dobra vrsta kriginga za prostorno-vremensku interpolaciju i predviđanje podataka ako su zadovoljene pretpostavke stacionarnosti drugog reda (definicija 3.2). Većinom u stvarnosti ove pretpostavke neće vrijediti pa napredniji oblici kriginga, poput običnog ili univerzalnog kriginga, mogu biti prikladniji.

Prediktor prostorno-vremenskog jednostavnog kriginga je uvjetno očekivanje (4.5), tj.

$$\widehat{Y}_{SK}\left(\boldsymbol{s}_0; t_0\right) = \boldsymbol{x}\left(\boldsymbol{s}_0; t_0\right)^{\intercal} \boldsymbol{\beta} + \boldsymbol{c}_0^{\intercal} \boldsymbol{C}_Z^{-1} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}), \tag{4.9}$$

a uvjetna varijanca (4.6) je varijanca prostorno-vremenskog jednostavnog kriginga

$$\sigma^2_{SK}\left(\boldsymbol{s}_0; t_0\right) = c_{0,0} - \boldsymbol{c}_0^{\intercal} \boldsymbol{C}_Z^{-1} \boldsymbol{c}_0. \tag{4.10}$$

Sjetimo se da je $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ i da se uvjetno očekivanje sastoji od reziduala kojima su dodane težine $\boldsymbol{w}^{\intercal} = \boldsymbol{c}_0^{\intercal} \boldsymbol{C}_Z^{-1}$. U vidu toga i pretpostavke da je očekivanje konstanta, npr. $\mu = \mathbf{E}Z(\boldsymbol{s}; t)$, $\forall\, (\boldsymbol{s}; t) \in \mathbb{D} \times \mathbb{N}$, imamo alternativni zapis prediktora jednostavnog kriginga

$$\widehat{Y}_{SK}\left(\boldsymbol{s}_0; t_0\right) = \mu + \boldsymbol{w}^{T} (\boldsymbol{z} - \mu \mathbf{1}). \tag{4.11}$$

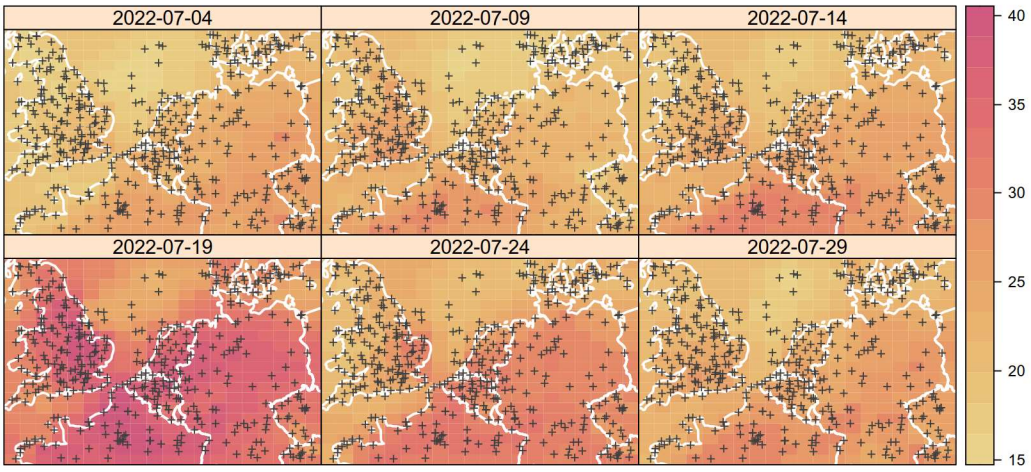
Interpolacija se lako može izvršiti uz pomoć funkcije `krigeST()`. Budući da ne znamo stvarno očekivanje, a potrebno nam je znati ga kako bi mogli provesti jednostavni kriging, možemo pretpostaviti, tj. procijeniti ga aritmetičkom sredinom opservacija na danom području „centra" Europe u srpnju.

```
mu <- mean(STIDF_EuJuly$MAX) #23.82076

simple.kriging <- krigeST(MAX ~ 1,
                          data = STIDF_EuJuly,
                          newdata = STF_EU,
                          modelList = sumMetricVarFit,
                          beta = mu,
                          computeVar = TRUE) # želimo da nam izračuna varijance
```

Pozivanjem ***spacetime*** funkcije `stplot()` kreiramo grafove rezultata kriginga u obliku rešetkastog prikaza prostorno-vremenskih klasa.

```
stplot(simple.kriging,
       main = NULL,
       layout = c(3, 2), # raspored dnevnih grafova
       sp.layout = layout, # granice država
       col.regions = color_pal) # paleta boja
```

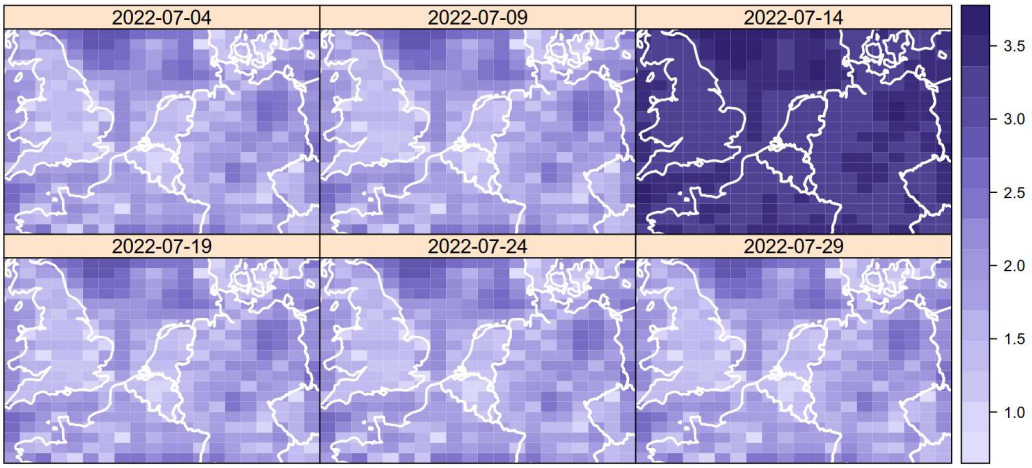


Slika 4.9: Rešetkasti prikaz predikcija dobivenih prostorno-vremenskim jednostavnim kriginom za šest dana u srpnju 2022. na području „centra" Europe. Bijele linije označuju granice država, a križići meteorološke stanice.

Izračunamo standardne devijacije predikcija kriginga i prikažemo ih grafički.

```
simple.kriging$std <- sqrt(simple.kriging$var1.var)
stplot(simple.kriging[, , "std"],
       main = NULL,
       layout = c(3, 2),
       sp.layout = layout.std,
       col.regions = rev(color_pal.std))
```

Na slici 4.10 vidimo da su najmanje standardne devijacije na mjestima gdje su postavljene meteorološke stanice. Najveće su dana 14. srpnja 2022. jer smo te podatke izbacili iz kriginga pa je napravilo predikciju s podacima koji su mu dani, no problem je što su veliki vremenski razmaci za precizniju predikciju.



Slika 4.10: Rešetkasti prikaz standardnih devijacija dobivenih prostorno-vremenskim jednostavnim krigingom za šest dana u srpnju 2022. na području „centra" Europe. Bijele linije označuju granice država.

4.4 Obični kriging

Poseban slučaj kriginga u kojem za razliku od jednostavnog kriginga, nije pretpostavljeno poznavanje očekivanja i funkcija kovarijanci, ali pretpostavljamo da je očekivanje konstantno, $\mu = \mathbf{E}Z(\boldsymbol{s};t)$, $\forall\,(\boldsymbol{s};t) \in \mathbb{D} \times \mathbb{N}$, $\mu \in \mathbb{R}$, pa imamo $(m \cdot T)$-dimenzionalni vektor očekivanja $\boldsymbol{\mu} = \mu\mathbf{1}$ fiksnih, ali nepoznatih vrijednosti na svakoj lokaciji $\boldsymbol{s} \in \mathbb{D}$ i u svakom vremenskom trenutku $t \in \mathbb{N}$. Pa će nam prediktor običnog kriginga biti isti kao prediktor jednostavnog kriginga $\widehat{Y}_{SK}$ (4.9) u koji uvrštavamo supstituciju očekivanja μ s procjeniteljem $\widehat{\mu}_{GLS}$ za μ dobiven generaliziranom metodom najmanjih kvadrata. Prediktor običnog kriginga je dan s

$$\widehat{Y}_{OK}\,(\boldsymbol{s}_0;t_0) = \widehat{\mu}_{GLS} + \boldsymbol{c}_0^\intercal \boldsymbol{C}_Z^{-1}(\boldsymbol{z} - \widehat{\mu}_{GLS}\mathbf{1}), \tag{4.12}$$

gdje je $\widehat{\mu}_{GLS}$ dan s

$$\widehat{\mu}_{GLS} = (\mathbf{1}^\intercal \boldsymbol{C}_Z^{-1}\mathbf{1})^{-1}\mathbf{1}^\intercal \boldsymbol{C}_Z^{-1}\boldsymbol{z}.$$

Varijanca prostorno-vremenskog običnog kriginga je

$$\sigma^2_{OK}\,(\boldsymbol{s}_0;t_0) = \underbrace{c_{0,0} - \boldsymbol{c}_0^\intercal \boldsymbol{C}_Z^{-1}\boldsymbol{c}_0}_{=\sigma^2_{SK}} + \kappa, \tag{4.13}$$

gdje

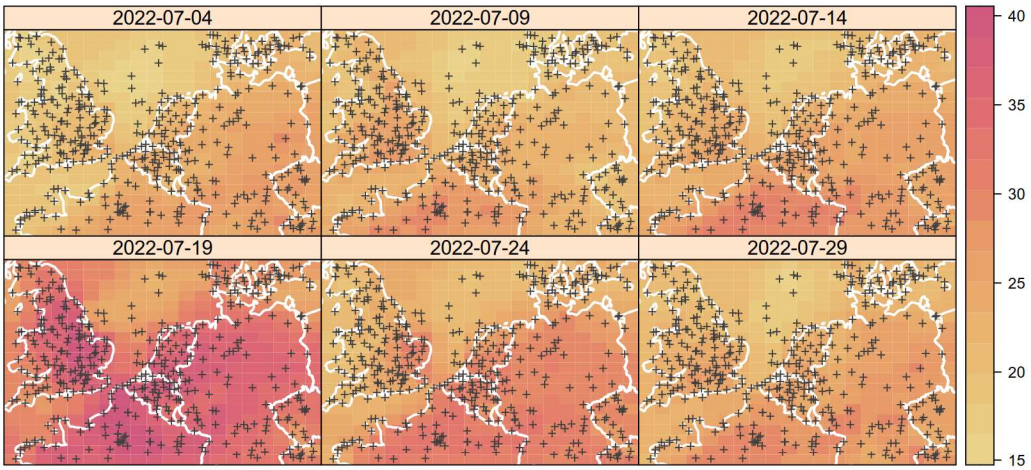
$$\kappa = (1 - \mathbf{1}^\intercal \boldsymbol{C}_Z^{-1}\boldsymbol{c}_0)^\intercal (\mathbf{1}^\intercal \boldsymbol{C}_Z^{-1}\mathbf{1})^{-1}(1 - \mathbf{1}^\intercal \boldsymbol{C}_Z^{-1}\boldsymbol{c}_0) \tag{4.14}$$

predstavlja dodatnu varijancu procjenitelja $\widehat{\mu}_{GLS}$ za parametar μ, a vidimo i da je sigurno $\sigma^2_{OK} \geq \sigma^2_{SK}$, što znači da imamo veću nesigurnost u procjenu vrijednosti varijable koju promatramo.

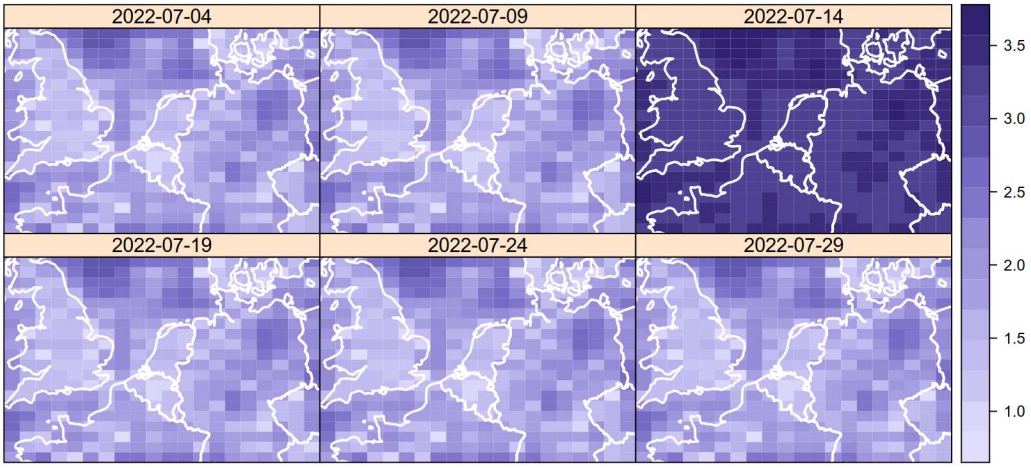
Interpolacija se u R-u analogno izvršava kao jednostavni kriging uz pomoć funkcije `krigeST()` samo što sada ne znamo očekivanje i ne navodimo ga kao dodatan parametar.

```
ordinary.kriging <- krigeST(MAX ~ 1,
                            data = STIDF_EuJuly,
                            newdata = STF_EU,
                            modelList = sumMetricVarFit,
                            computeVar = TRUE)
```

Analogno jednostavnom krigingu kreiramo grafove rezultata kriginga u obliku rešetkastog prikaza prostorno-vremenskih klasa funkcijom `stplot()`.



Slika 4.11: Rešetkasti prikaz predikcija dobivenih prostorno-vremenskim običnim krigingom za šest dana u srpnju 2022. na području „centra" Europe. Bijele linije označuju granice država, a križići meteorološke stanice.



Slika 4.12: Rešetkasti prikaz standardnih devijacija dobivenih prostorno-vremenskim običnim krigingom za šest dana u srpnju 2022. na području „centra" Europe. Bijele linije označuju granice država.

4.5 Univerzalni kriging

Generalizirani model običnog kriginga kod kojeg je pretpostavka konstantnog očekivanja zamijenjena s pretpostavkom linearnog modela. Sada je očekivanje procesa $Y(\boldsymbol{s};t)$, $\mathbf{E}Y(\boldsymbol{s};t) = \boldsymbol{X}\boldsymbol{\beta}$, gdje je $\boldsymbol{\beta}$ $(p+1)$-dimenzionalni vektor parametara nepoznat. Prediktor univerzalnog kriginga od $Y(\boldsymbol{s}_0;t_0)$ je

$$\widehat{Y}_{UK}\left(\boldsymbol{s}_0;t_0\right) = \boldsymbol{x}\left(\boldsymbol{s}_0;t_0\right)^{\intercal}\widehat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{c}_0^{\intercal}\boldsymbol{C}_Z^{-1}(\boldsymbol{z} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{GLS}), \tag{4.15}$$

gdje je $\widehat{\boldsymbol{\beta}}_{GLS}$ procjenitelj za $\boldsymbol{\beta}$ dobiven generaliziranom metodom najmanjih kvadrata dan s

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^{\intercal}\boldsymbol{C}_Z^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{C}_Z^{-1}\boldsymbol{z}.$$

Varijanca prostorno-vremenskog univerzalnog kriginga je

$$\sigma^2_{UK}\left(\boldsymbol{s}_0;t_0\right) = c_{0,0} - \boldsymbol{c}_0^{\intercal}\boldsymbol{C}_Z^{-1}\boldsymbol{c}_0 + \kappa, \tag{4.16}$$

gdje

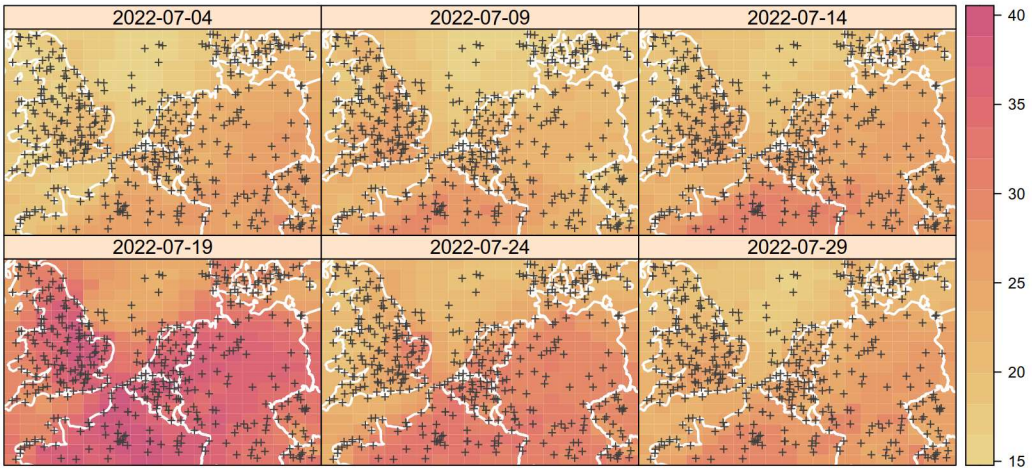
$$\kappa = (\boldsymbol{x}\left(\boldsymbol{s}_0;t_0\right) - \boldsymbol{X}^{\intercal}\boldsymbol{C}_Z^{-1}\boldsymbol{c}_0)^{\intercal}(\boldsymbol{X}^{\intercal}\boldsymbol{C}_Z^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x}\left(\boldsymbol{s}_0;t_0\right) - \boldsymbol{X}^{\intercal}\boldsymbol{C}_Z^{-1}\boldsymbol{c}_0) \tag{4.17}$$

predstavlja dodatnu varijancu kao matricu kovarijanci procjenitelja $\widehat{\boldsymbol{\beta}}_{GLS}$ za parametar $\boldsymbol{\beta}$.

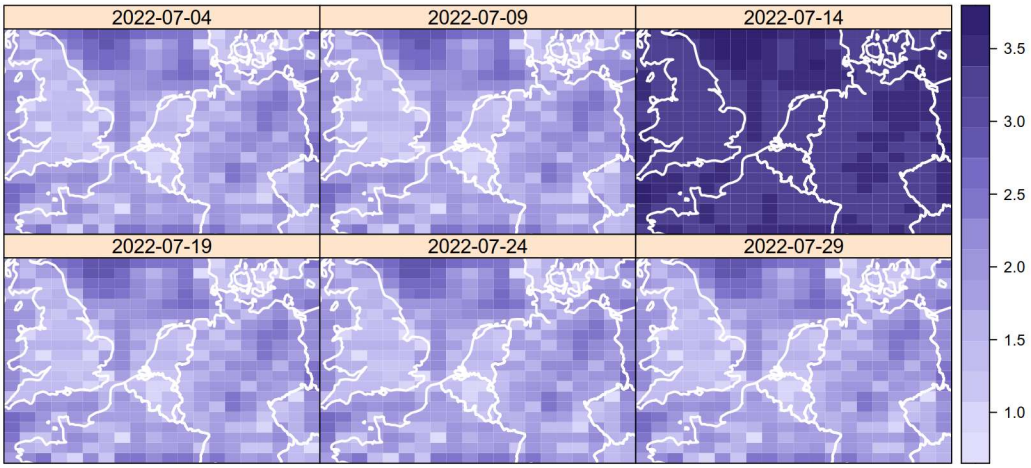
Kao u protekla dva kriginga, i u ovom pozivamo funkciju `krigeST()`, ali sada navodimo dodatnu nezavisnu varijablu *LATITUDE* u GLS formulu jer je pokazala rastući trend na slici 3.3.

```
universal.kriging <- krigeST(MAX ~ 1 + LATITUDE, # GLS formula
                             data = STIDF_EuJuly,
                             newdata = STF_EU,
                             modelList = sumMetricVarFit,
                             computeVar = TRUE)
```

Vizualizaciju dobijemo pozivanjem funkcije `stplot()` kao i kod ostala dva kriginga.



Slika 4.13: Rešetkasti prikaz predikcija dobivenih prostorno-vremenskim univerzalnim krigingom za šest dana u srpnju 2022. na području „centra" Europe. Bijele linije označuju granice država, a križići meteorološke stanice.



Slika 4.14: Rešetkasti prikaz standardnih devijacija dobivenih prostorno-vremenskim univerzalnim krigingom za šest dana u srpnju 2022. na području „centra" Europe. Bijele linije označuju granice država.

4.6 Usporedbe kriginga

Već iz dobivenih grafova kriginga smo uočili dosta sličnosti, ako ne i iste vizualne prikaze među ove tri vrste kriginga pa ćemo ih prikazati jedne pored drugih tako da barem pokušamo lakše uočiti razlike među njima. Prikazat ćemo njihove predikcije i standardne devijacije predikcija na istom grafu (slika 4.15). Uspoređivat ćemo rezultate dana 14. srpnja 2022. jer smo taj dan izbacili iz kriginga pa nam on ima i puno veću varijancu od ostalih dana. Kako bismo dobili ove grafove, pozvali smo ***sp*** funkciju `spplot()`. Sljedeći R kod je primjer kako dobiti ovakav graf usporedbe predikcija. Analogno se dobije graf usporedbe standardnih devijacija predikcija.

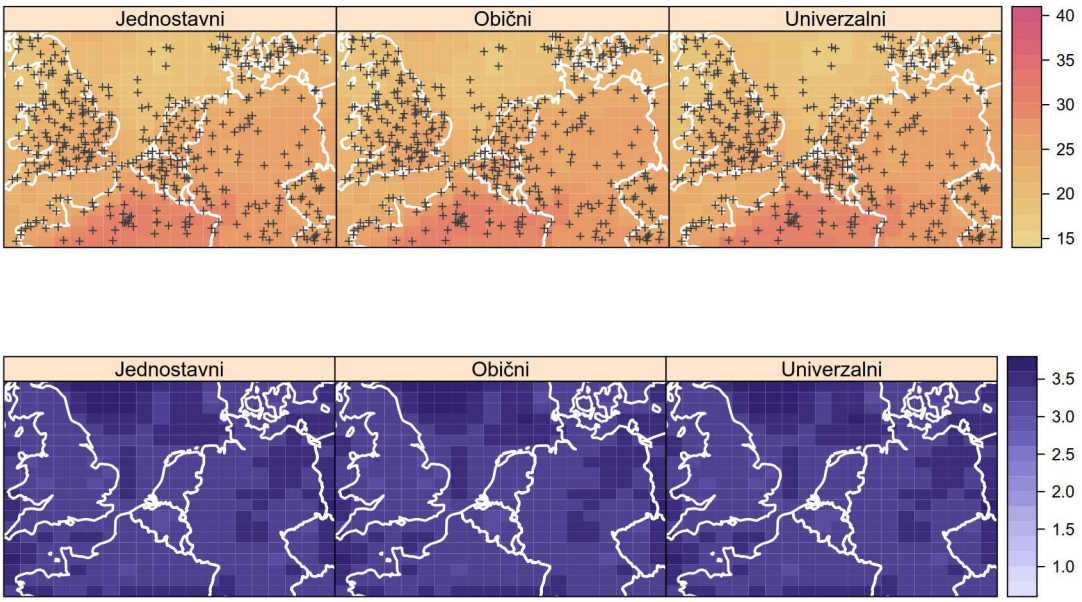
```
# Definiramo podskup dana 14. srpnja
predikcije14 <- universal.kriging[,"2022-07-14"]

# Spremimo predikcije dobivene vrstama kriginga
predikcije14$var1.pred <- simple.kriging[,"2022-07-14"]$var1.pred
predikcije14$var1.var <- ordinary.kriging[,"2022-07-14"]$var1.pred
predikcije14$std <- universal.kriging[,"2022-07-14"]$var1.pred
names(predikcije14@data) <- c("Jednostavni","Obični","Univerzalni")

# Definiramo skalu iznosa temperatura da bude jednaka kao u prethodnim slikama
plot.zlim <- seq(floor(min(universal.kriging$var1.pred,
                           simple.kriging$var1.pred,
                           ordinary.kriging$var1.pred)),
                 ceiling(max(universal.kriging$var1.pred,
                             simple.kriging$var1.pred,
                             ordinary.kriging$var1.pred)),
                 by = 1.8)

# Kreiramo graf
spplot(predikcije14,
       main = NULL,
       layout = c(3, 1),
```

```
        sp.layout = layout,
        col.regions = color_pal,
        at = plot.zlim)
```



Slika 4.15: Usporedba rezultata dobivenih krigingom 14. srpnja 2022. Gore su predikcije, a dolje standardne devijacije.

No ipak se teško uočavaju razlike golim okom pa pogledajmo sada deskriptivnu statistiku predikcija u tablici 4.1 i standardnih devijacija predikcija u tablici 4.2.

Vrsta kriginga	**Minimum**	**Donji kvartil**	**Medijan**	**Prosjek**	**Gornji kvartil**	**Maksimum**
Jednostavni	17.10	20.94	23.85	23.68	25.85	31.98
Obični	17.07	20.91	23.82	23.65	25.82	31.95
Univerzalni	16.95	20.66	23.87	23.60	25.89	31.97

Tablica 4.1: Usporedba predikcija dobivenih krigingom.

Vrsta kriginga	**Minimum**	**Donji kvartil**	**Medijan**	**Prosjek**	**Gornji kvartil**	**Maksimum**
Jednostavni	3.171	3.253	3.323	3.347	3.420	3.773
Obični	3.173	3.254	3.324	3.349	3.422	3.776
Univerzalni	3.173	3.254	3.324	3.350	3.422	3.794

Tablica 4.2: Usporedba standardnih devijacija predikcija dobivenih krigingom.

Razlike su male, ali postoje. Pokazali smo ovim putem da povećanjem nepoznatih parametara u modelu povećavamo standardnu devijaciju predikcije.

Literatura

[1] R. S. Bivand, E. Pebesma, V. Gomez-Rubio, *Applied spatial data analysis with R*, Springer, 2013.

[2] C. P. Chen, *Positive Definite Matrix*, National Sun Yat-sen University (Dostupno na: `http://slpl.cse.nsysu.edu.tw/chiaping/la/pdm.pdf`)

[3] N. Cressie, C. K. Wikle, *Statistics for Spatio-Temporal Data*, John Wiley & Sons, 2011.

[4] N. Cressie, C. K. Wikle, A. Zammit-Mangion, *Spatio-Temporal Statistics with R*, John Wiley & Sons, 2019.

[5] B. Gräler, G. Heuvelink, E. Pebesma, *Spatio-Temporal Interpolation using gstat*, The R Journal, 2016. (Dostupno na: `https://cran.r-project.org/web/packages/gstat/vignettes/spatio-temporal-kriging.pdf`)

[6] D. Jukić, *Realna analiza*, Sveučilište Josipa Jurja Strossmayera u Osijeku - Odjel za matematiku, 2020. (Dostupno na: `https://www.mathos.unios.hr/images/homepages/jukicd/skripta.pdf`)

[7] E. Pebesma, *gstat user's manual*, Dept. of Physical Geography, Utrecht University, 2014. (Dostupno na: `https://www.gstat.org/gstat.pdf`)

[8] E. Pebesma, *Spatio-Temporal Data in R*, Journal of Statistical Software, 2012. (Dostupno na: `https://www.jstatsoft.org/article/view/v051i07`)

[9] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, Massachusetts Institute of Technology, 2006. (Dostupno na: `https://gaussianprocess.org/gpml/chapters/RW.pdf`)

[10] D. G. Rossiter, *Applied geostatistics*, National Cornell University, 2020. (Dostupno na: `https://www.css.cornell.edu/faculty/dgr2/_static/files/R_PDF/exC.pdf`)

[11] S. K. Sahu, *Bayesian Modeling of Spatio-Temporal Data with R*, CHAPMAN & HALL, 2022. (Dostupno na: `https://www.soton.ac.uk/~sks/bmbook/9780429318443_webpdf.pdf`)

[12] M. Sherman, *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*, John Wiley & Sons, 2011.

[13] *Spatio-Temporal Modeling*, 2017., URL: `http://pipiras.sites.oasis.unc.edu/timeseries/Multivariate_7_-_Spatio_Temporal_-_Menu.html#what_is_this_all_about`

[14] *The Book of Statistical Proofs*, URL: `https://statproofbook.github.io/P/mvn-cond`

Prostorno-vremenski modeli u geostatistici

Sažetak

U ovom radu se upoznajemo s prostorno-vremenskim podacima kao proširenje vremenskih nizova, definiramo pripadne slučajne procese i navodimo razlike među tipovima baza podataka. Nakon upoznavanja ove vrste podataka, u radu se zadržavamo na geostatističkom tipu podataka („point referenced"). Krećemo od analize geostatističkih podataka koja se sastoji od vizualizacije i numeričkih karakteristika. Definiramo empirijsko prostorno i vremensko očekivanje, empirijsku prostornu kovarijancu i prostorno-vremenski variogram. Nakon toga podatke opisujemo modelom koji se najčešće koristi u geostatistici, a to je kriging, metoda interpolacije koja se dijeli na tri vrste, jednostavni, obični i univerzalni kriging. Sve prikazujemo na primjeru dnevnih maksimalnih temperatura u Europi u 2022. godini zabilježenih u klimatološkom centru američkog ratnog zrakoplovstva. Dobiveni rezultati i grafovi su dobiveni programskim jezikom R (inačica 4.2.2) te su neki dijelovi koda uključeni u radu.

Ključne riječi: empirijska kovarijanca, empirijsko očekivanje, funkcija kovarijanci, funkcija očekivanja, geostatistika, jednostavni kriging, metrički model variograma, model metričke sume variograma, obični kriging, „point referenced" podaci, separabilni model variograma, univerzalni kriging, variogram

Spatio-temporal models in geostatistics

Abstract

In this thesis, the spatio-temporal data are introduced as an extension of time series models, we define the corresponding random processes and state the differences among the types of databases. After getting acquainted with spatio-temporal data, our focus shifts to geostatistical data ("point referenced"). We begin by analyzing geostatistical data, which involves data visualization and exploratory data analysis. We establish concepts like empirical spatial and temporal mean, empirical spatial covariance and spatio-temporal variogram. Following that, we describe the data using a widely used statistical technique called kriging, an interpolation method that comes in three variations, simple, ordinary and universal kriging. To illustrate these concepts, we use an example of daily maximum temperatures in Europe during the year 2022 obtained from the United States Air Force Climatology Center. We present our findings and graphs generated using the R programming language (version 4.2.2), and we also include some code in the thesis.

Keywords: empirical covariance, empirical expectation, covariance function, expectation function, geostatistics, simple kriging, metric variogram model, ordinary kriging, point referenced data, separable variogram model, sum-metric variogram model, universal kriging, variogram

Životopis

Rođena sam 29. studenog 1996. godine u Đakovu. Pohađala sam osnovnu školu "Ivan Goran Kovačić" te nakon toga jezičnu gimnaziju "Antuna Gustava Matoša" u Đakovu. Preddiplomski studij Matematike na Odjelu za matematiku u Osijeku upisujem 2015. godine. Završavam ga 2019. godine s temom završnog rada "Newton – Cotesove formule" pod mentorstvom izv. prof. dr. sc. Tomislava Maroševića. Diplomski studij Financijska matematika i statistika upisujem 2019. godine. Stručnu praksu sam radila 2020. u Escape d.o.o., a nakon prakse sam ostala raditi kao student. Od 2021. godine sam zaposlena u istoj firmi kao podatkovni analitičar.