Univerzitet Crne Gore
Elektrotehnički fakultet

Edin Salković

# DIGITALIZACIJA PEDOLOŠKIH PODATAKA CRNE GORE I GENERISANJE PEDOLOŠKIH KARATA

MAGISTARSKI RAD

Podgorica, 2016.

**Informacije o magistrandu**

| | |
|---|---|
| Ime i prezime: | Edin Salković |
| Datum i mjesto rođenja: | 03.03.1983. godina, Skoplje, Makedonija |
| Završeni osnovni studijski program: | Elektrotehnički fakultet (dipl. ing., 4 godine)<br>Odsjek: Elektronika, telekomunikacije i računari<br>Smjer: Računari<br>Godina završetka studija: 2005. |

**Informacije o magistarskom radu**

| | |
|---|---|
| Naziv postdiplomskog studija: | Akademske magistarske studije, studijski program: Elektronika, telekomunikacije i računari |
| Naslov rada: | Digitalizacija pedoloških podataka Crne Gore i generisanje pedoloških karata |
| Fakultet: | Elektrotehnički fakultet, Podgorica |

**UDK, ocjena i odbrana magistarskog rada**

| | |
|---|---|
| Datum prijave rada: | 8. 12. 2015. |
| Datum prihvatanja teme: | 16. 2. 2016. |
| Komisija za ocjenu teme: | Dr Mirko Knežević<br>Prof. dr Igor Đurović<br>Prof. dr Vesna Popović-Bugarin |
| Mentor: | Prof. dr Igor Đurović |
| Komisija za ocjenu rada: | Dr Mirko Knežević<br>Prof. dr Igor Đurović<br>Prof. dr Vesna Popović-Bugarin |
| Komisija za odbranu rada: | Dr Mirko Knežević<br>Prof. dr Igor Đurović<br>Prof. dr Vesna Popović-Bugarin |
| Lektor: | Elsan Salković, profesor njemačkog jezika i književnosti |
| Datum odbrane: | 19. 7. 2016. |
| Datum promocije: | |

# Predgovor

Rad obrađuje dvije tijesno povezane cjeline: digitalizaciju pedoloških podataka Crne Gore i generisanje pedoloških karata na osnovu tih podataka. Iako je u radu glavni akcenat stavljen na generisanje karata, značajan dio rada obrađuje problem digitalizacije pedoloških podataka Crne Gore, jer su u pitanju podaci od državnog značaja: sakupljani su duži niz godina na teritoriji cijele Crne Gore. Takođe, kvalitetna priprema podataka za izradu karata često oduzme i preko 80% vremena ukupnog procesa generisanja karata.

Rad je organizovan na sljedeći način:

- U uvodnom poglavlju su dati osnovni koncepti pedologije (nauke o zemljištu) kako bi se ona približila inžinjeru elektrotehnike/računarstva.
- U drugom poglavlju je dat pregled tehnika, algoritama i metoda koji se koriste za obradu pedoloških podataka i dobijanje digitalnih pedoloških karata.
- Sljedeće poglavlje je posvećeno opisu prikupljanja digitalnih pedoloških podataka na teritoriji Crne Gore i izazovima digitalizacije ovih podataka.
- Poglavlje posvećeno generisanju tematske karte lako pristupačnog fosfora (hemijska karakteristika) za dio teritorije Crne Gore je najznačajnije poglavlje rada. Premda postoje dvije vrste pedoloških karata, i to a) tematske, koje se bave prikazom pojedinačnih karakteristika zemljišta i b) tipske, koje se bave prikazom klasa zemljišta, odabrana je izrada jedne tematske karte, jer su one jednostavnije za izradu i predstavljaju osnovu za izradu modernih tipskih karata zemljišta, čija izrada prevazilazi okvire ovog rada. Odabrano je nekoliko algoritama, pri čemu je najkompleksniji univerzalni kriging. Za svaki od algoritama je generisana karta, a data su i kvantitativna poređenja uspješnosti algoritama postupkom, tzv. validacije.
- Posljednje poglavlje je poglavlje sa zaključnim razmatranjima gdje je dat rezime čitavog rada, kao i moguće smjernice za dalja istraživanja na ovom polju.

# Naslov

Digitalizacija pedoloških podataka Crne Gore i generisanje pedoloških karata

# Izvod rada

U radu je dat pregled postupka digitalizacije pedoloških podataka Crne Gore i pregled nekoliko geostatističkih interpolacionih algoritama, kao i primjer njihove konkretne primjene u cilju generisanja tematske pedološke karte lako pristupačnog fosfora za jedan dio Crne Gore, uz ocjenu pouzdanosti i kvaliteta tako dobijene karte. Iako je u radu glavni akcenat stavljen na generisanje karata, značajan dio rada obrađuje problem digitalizacije pedoloških podataka Crne Gore, u što spada ne samo unos podataka preko računara, već i njihova obrada, filtriranje i uopšte priprema za korišćenje u algoritmima za izradu mapa. Opis procesa digitalizacije je takođe od posebnog značaja, jer na neki način predstavlja sintezu višedecenijskog obimnog rada na ovom polju.

# Title

Digitalization of Montenegro's pedological data and pedological map generation

# Abstract

This work gives an overview of the process of digitalizing the pedological data of Montenegro and gives an overview of several geostatistical interpolation algorithms, as well as an example of their concrete application in generating a thematic pedological map of phosphorus for a part of Montenegro, together with an assessment of the reliability and quality of that map. Although the main topic is map synthesis, a significant part of the work deals with the problem of digitalizing of pedological data of Montenegro, including not only data entry, but also its processing, filtering and, in general, its preparing for usage in map creating algorithms. The description of the process of digitalization has also a special significance, as it, in a way, represents the synthesis of a decades long work in this field.

# Sadržaj

# Poglavlje 1

# Osnovni koncepti pedologije

U ovom poglavlju će biti dati osnovni koncepti pedologije, kako bi se ona približila inžinjeru elektrotehnike.

Pedologija (πέδον, pedon: zemljište) je nauka koja se bavi proučavanjem zemljišta u njegovom prirodnom ambijentu. Sa aspekta primjene, može se podijeliti na agropedologiju (poljoprivreda), silvopedologiju (šumarstvo), meliorativnu pedologiju (melioracije: unapređenja zemljišta), hidropedologiju, itd. (Resulović i Čustović 2002, str. 11-14).

Tlo, koje je predmet izučavanja pedologije, se može definisati kao površinski sloj Zemljine kore, tj. rastresiti materijal koji naliježe na vršne stijene čvrstog omotača Zemlje — litosfere (λίθος, lithos: kamen). Često je granicu između tla i stijene teško definisati, pa se radi toga, ali i usljed potrebe izučavanja nekog specifičnog tla, granica tla može postaviti pliće ili pak dublje. Na primjer, u šumarstvu, za razliku od poljoprivrede, ova granica mora biti postavljena dublje, usljed većeg uticaja stjenovitog sloja na razvoj drveća (Resulović i Čustović 2002, str. 15).

## 1.1 Ispitivanje zemljišta

Kada se neko zemljište želi istražiti (npr. za potrebe poljoprivrede, i sl.) sprovodi se prvo terensko, a zatim i laboratorijsko ispitivanje (Belić, Nešić i Ćirić 2014, str. 22). Kod terenskog ispitivanja se, za geografsku oblast od interesa, prvo proučavaju činioci vezani za nastanak i razvoj zemljišta: klima, živi organizmi, reljef, matični supstrat i vrijeme (Jenny 1994, str. 10), a potom se planski uzimaju uzorci zemljišta samo na lokacijama koje se proučavanjem procijene kao najznačajnije i

najmjerodavnije za dato istraživanje, jer je naravno praktično neizvodljivo uzimanje uzoraka za svaki djelić zemlje.

Uzorci se uzimaju na dva načina (Belić, Nešić i Ćirić 2014, str. 24):

- tzv. „sondiranjem“ (slika 1.1) koje je jednostavnije, ali manje precizno, gdje se uzima uzorak zemljišta za prvih 60 cm dubine;
- „otvaranjem“, tzv. „pedoloških profila“ (slika 1.2), koji predstavljaju rupe koje se kopaju do dubine dejstva pedogenetskih činilaca (najčešće do dubine 2 m), odnosno do pojave podzemne vode ili čvrste matične stijene, i to na reprezentativnom mestu, tj. što je moguće dalje od mjesta gdje postoji ljudsko djelovanje na zemljište.

Slika 1.1: Sondiranje

Nakon otvaranja profila i jednostavne vizuelne inspekcije, pedolozi bilježe niz podataka o zemljištu, između ostalog i opis spoljašnje (reljef, biljni pokrivač) i unutrašnje morfologije. Kod opisivanja unutrašnje morfologije, posebno je značajno odrediti koliko horizonata ima zemljište, kao i njihove donje i gornje gra-

Slika 1.2: Primjer profila

nice (najčešće u cm). *Horizonti* su horizontalni slojevi[1] zemljišta koji su nastali razvojem pedogenetske podloge, i koji se, osim vizuelno, međusobno razlikuju po morfološkim, fizičkim, hemijskim i biološkim svojstvima (Belić, Nešić i Ćirić 2014, str. 24-26).

Nakon utvrđivanja granica horizonata, uzorci zemljišta se uzimaju iz pojedinih horizonata, a potom se nad uzorcima vrše laboratorijska ispitivanja kojima se utvrđuju fizičke, vodno-fizičke, mehaničko-fizičke i hemijske osobine zemljišta (Belić, Nešić i Ćirić 2014, str. 33-89). Radi jednostavnosti, u radu ćemo fizičke, vodno-fizičke i mehaničko-fizičke osobine nazivati „mehaničko-fizičke osobine“ ili „MP osobine“ (od *Mechanical & Physical*), a hemijske „C osobine“ (od *Chemical*).

[1]Termin „sloj“ se često u pedologiji koristi u drugom, specifičnom značenju: za naslagu određene supstance koja nije nastala razvojem pedogenetske podloge. Ovdje je, u definiciji horizonta, koristimo u opštem značenju.

## 1.2 Mehaničko-fizičke osobine zemljišta

Neke od MP osobina su:

**Mehanički sastav**

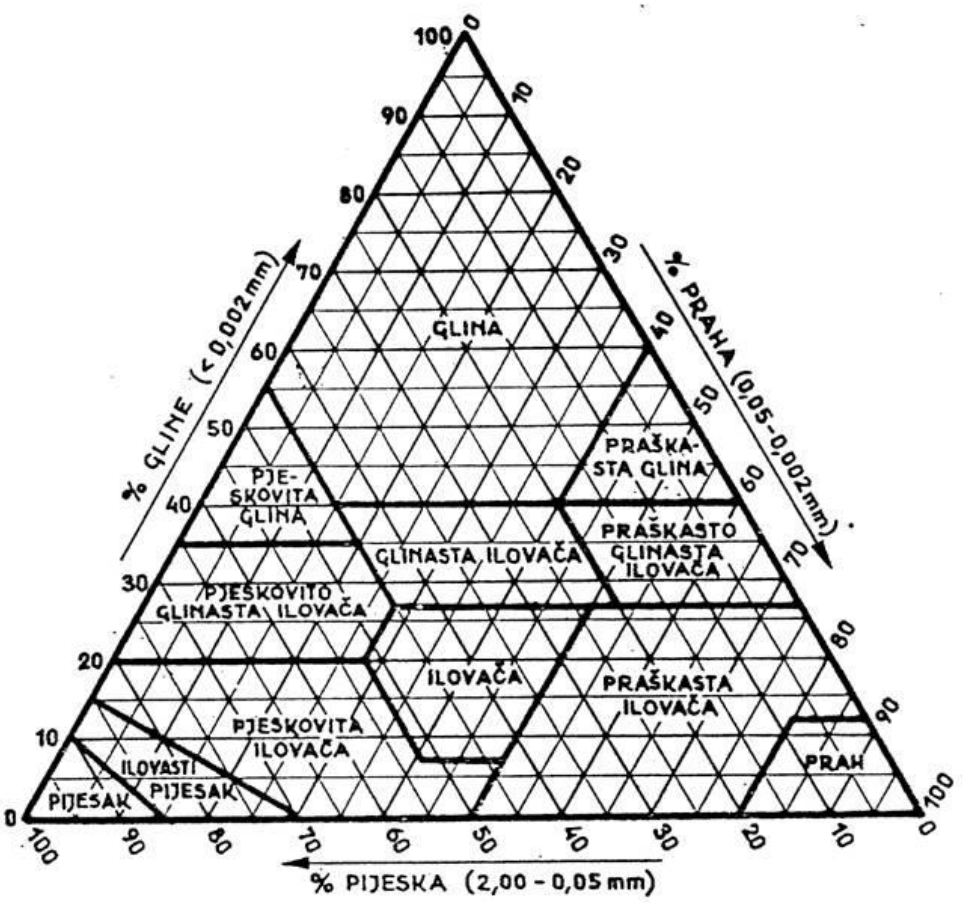
Mehanički sastav zemljišta (Belić, Nešić i Ćirić 2014, str. 33-39) je jedna od njegovih ključnih osobina, a odnosi se na procentualnu zastupljenost čestica različitih prečnika u zemljištu. Same čestice se svrstavaju u frakcije, koje su definisane maksimalnim i minimalnim čestičnim prečnikom. Postoji veliki broj klasifikacija čestica, a na našim prostorima se najčešće koristi klasifikacija po Atterbergu (tabela 1.1). Samo zemljište klasifikujemo na osnovu te-

| **Naziv frakcije** | **Veličina čestice** |
|---|---|
| Skelet | |
| Kamen | >20 mm |
| Šljunak | 20–2 mm |
| Sitna zemlja | |
| Krupan pijesak | 2–0.2 mm |
| Sitan pijesak | 0.2–0.02 mm |
| Prah | 0.02–0.002 mm |
| Glina | <0.002 mm |
| Koloidna glina | <0.0002 mm |
| | ili <0.0005 mm |

Tabela 1.1: Klasifikacija čestica po Atterbergu

ksture, tj. procentualne zastupljenosti različitih frakcija čestica u njemu. I tu postoji više klasifikacija, a na slici 1.3 je data klasifikacija koju koristi *United States Department of Agriculture (USDA)*.

Agregati su krupne sekundarne čestice koje predstavljaju kompaktne, me-




Slika 1.3: *USDA* klasifikacija zemljišta po teksturi

đusobno povezane skupine primarnih čestica zemljišta. Nazivaju se i strukturni agregati.

**Vlažnost**

**Stabilnost makroagregata**

Makroagregati se definišu kao agregati primarnih čestica većih od 0.25 mm. Stabilnost makroagregata je ocjena otpornosti makroagregata na rasplinjavanje u vodi, i dobar je pokazatelj otpornosti zemljišta na eroziju vodom i vjetrom, a samim tim i održivosti biljne proizvodnje (Belić, Nešić i Ćirić 2014, str. 46-47).

**Zapreminska specifična masa**

„Zapreminska specifična masa predstavlja masu apsolutno suvog zemljišta

u prirodnom, nenarušenom stanju (sa porama) u jedinici zapremine. Praktični značaj zapreminske specifične mase je velik, jer se njene vrijednosti koriste za procjenu stepena sabijenosti zemljišta i za izračunavanje ukupne poroznosti zemljišta, sadržaja vode u zemljištu i zalivnih normi.“ (Belić, Nešić i Ćirić 2014, str. 40-41).

Postoji još značajnih MP osobina kao što su poroznost, vodo-retencioni kapacitet, vodopropustljivost, vazdušni kapacitet itd., ali su ovdje navedene samo neke, radi ilustracije tipa osobina koje spadaju u ovu kategoriju.

## 1.3 Hemijske osobine zemljišta

Neke od hemijskih osobina su:

**Aktivna kisjelost, ili „pH u $H_2O$“**

Određuje se u supspenziji zemljišta sa destilovanom vodom (Belić, Nešić i Ćirić 2014, str. 76).

**Supstituciona kisjelost, ili „pH u KCl“**

Fiziološki aktivna kisjelost, određuje se u supspenziji zemljišta sa 1 M rastvorom KCl (Belić, Nešić i Ćirić 2014, str. 76).

**Sadržaj kalcijum karbonata ($CaCO_3$)**

**Sadržaj humusa**

Humus nastaje od organske materije i u njegov sastav ulaze svi organogeni elementi: kiseonik, vodonik, ugljenik i azot, kao i elementi pepela. Najčešće se količina humusa određuje indirektno preko količine ugljenika ili ugljendioksida (Belić, Nešić i Ćirić 2014, str. 69).

**Sadržaj lako pristupačnog fosfora ($P_2O_5$)**

**Sadržaj lako pristupačnog kalijuma ($K_2O$)**

## 1.4 Zavisnost parametara zemljišta od prirodne sredine

Parametri zemljišta zavise od velikog broja činilaca, tako da je jako teško precizno formalno definisati tu zavisnost.

Prije razvoja savremenih računara, jedna od najpoznatijih formulacija te zavisnosti je bila, tzv. *clorpt* jednačina (Jenny 1994):

$$S = f(cl, o, r, p, t, ...)$$

gdje je:

- $S$: zemljište,
- $cl$: klima,
- $o$: organska aktivnost,
- $r$: reljef,
- $p$: matični supstrat,
- $t$: vrijeme,
- ...: ostali činioci.

Ova jednačina prije svega objašnjava način nastajanja i razvoja zemljišta, pa makar (pedogenetski) činioci bili teško mjerljivi, dok je danas, iz praktičnih razloga, pažnja usmjerena na iskazivanje funkcionalne zavisnosti karakteristika zemljišta od empirijski mjerljivih činilaca. Jedna od formulacija takve zavisnosti je *scorpan* model (A.B. McBratney, Mendonça Santos i B. Minasny 2003):

$$S = f(s; c; o; r; p; a; n)$$

gdje je:

- $S$: klasa zemljišta ili neka osobina zemljišta;
- $s$: slično kao $S$, ali sada u ulozi činioca, npr. ostale osobine zemljišta na dotičnoj lokaciji;

- $c$: klima, tj. klimatski parametri na lokaciji;
- $o$: organizmi — flora, fauna i čovjek — tj. njihov uticaj;
- $r$: reljef;
- $p$: matični substrat, litologija;
- $a$: starost;
- $n$: prostor, lokacija.

Ako eksplicitno prikažemo koordinate i vrijeme, jednačina dobija oblik (A.B. McBratney, Mendonça Santos i B. Minasny 2003):

$$S[x, y, \tilde{t}] = f(s[x, y, \tilde{t}]; c[x, y, \tilde{t}]; o[x, y, \tilde{t}]; r[x, y, \tilde{t}]; p[x, y, \tilde{t}]; a[x, y]; [x, y])$$

gdje je $\tilde{t}$ približno vrijeme. Ako postoji zavisnost od vertikalne koordinate (nadmorska visina ili dubina zemljišta) onda se u navedenu jednačinu može dodati još i $z$.

Gore navedenu formulu je McBratney izveo nakon opširne analize literature na temu funkcionalne zavisnosti parametara zemljišta od činilaca prirodne sredine (A.B. McBratney, Mendonça Santos i B. Minasny 2003).

Ovdje je važno napomenuti da se za matematičku, tj. računarsku, predstavu reljefa koristi *DEM* (*Digital Elevation Model — Digitalni Model Elevacije*). Koriste se još i izrazi *DTM* (*Digital Terrain Model — Digitalni Model Terena*) i *DSM* (*Digital Surface Model — Digitalni Model Površine*), dok neki autori prave razliku između tih pojmova.

U svom osnovnom značenju DEM se vezuje samo za nadmorsku visinu površine Zemlje u nekoj tački. Međutim, u ovom radu pojam DEM obuhvata sve različite aspekte reljefa koji se mogu predstaviti (i generisati) računarski, a koji se mogu koristiti u analizi i „sintezi“ (tj. prognozi, tamo gdje nisu poznati) parametara zemljišta.

DEM modeli su značajni za prognozu parametara zemljišta, jer su lako dostupni: postoje besplatne baze na Internetu sa relativno preciznim DEM modelima (<100m), koji se uglavnom generišu iz satelitskih snimaka, npr. Landsat misije i dr.

Od satelitskih snimaka interesantni su i, npr. snimci iz opsega zelene boje, jer nam pokazuju količinu biljaka u nekoj tački.

# Poglavlje 2

# Pregled algoritama prostorne prognoze

U ovom poglavlju će biti dat pregled najvažnijih algoritama prostorne prognoze, sa posebnim osvrtom na geostatističke (tj. kriging) algoritme i njihovu primjenu u pedologiji.

## 2.1 Uvod

Savremeno terensko ispitivanje zemljišta je nezamislivo bez GPS uređaja koji omogućavaju precizno bilježenje koordinata profila zemljišta. Nakon izvršenih laboratorijskih ispitivanja, dobijeni podaci o MP i C osobinama nekog profila se mogu unijeti u bazu podataka zajedno sa njegovim geografskim koordinatama. Naravno, plansko ispitivanje zemljišta podrazumijeva otvaranje i ispitivanje velikog broja profila, pa unos MP i C podataka i koordinata za svaki profil ponaosob u jedinstvenu bazu omogućava dublju skupnu (tj. statističku) analizu svih profila i dobijanje novih korisnih informacija koje je teško dobiti iz pojedinačne analize ma kojeg profila.

Nauka koja se bavi statističkom obradom, analizom i interpretacijom prostorno (i vremenski) referenciranih podataka uopšte se naziva geostatistika (Hengl 2009, str. 2), i ona je relativno mlada oblast statistike, nastala u okrilju rudarstva (Krige 1951), a danas se koristi u geo-naukama (uključujući pedologiju), hidrologiji, ekologiji, poljoprivredi itd. (Zhou i dr. 2007). Neke od poznatijih istraživačkih grupa na ovom polju su (Hengl 2009): *International Association of Mathematical Geosciences*

(IAMG)[1], geoENVia[2], *pedometrics*[3] itd.

Naravno, podaci koje proučava geostatistika ne moraju biti vezani za karakteristike zemljišta, već mogu biti ma koji podaci koji su na neki način korelisani sa geografskim koordinatama, kao što su širenje neke epidemije, koncentracija jedinki neke životinjske vrste u nekoj oblasti i sl.

Može se reći da geostatistika ima 3 zadatka (Hengl 2009):

- **Proračun (procjena, estimacija) modela koji objašnjava funkcionalnu zavisnost izmjerenog prostornog parametra od ostalih parametara**.
- **Prognoza (interpolacija ili ekstrapolacija)[4] prostornog parametra za mjesta gdje parametar nije mjeren:** u zavisnosti od toga šta je cilj prognoze, možemo prognozirati parametar za pojedine tačke ili pak čitave površine. Prognoziranje parametra za neku površinu je u suštini generisanje rasterske karte na osnovu proračunatog modela.
- **Provjera hipoteze:** odnosi se na određivanje stepena tačnosti prognoziranog parametra za željene lokacije. Time, npr. utvrđujemo koliko je interpolisana mapa nekog parametra tačna.

Ovdje treba napomenuti da kod tradicionalne izrade karata vezanih za parametre prirodne sredine (uključujući i zemljište) nisu korišteni kompleksni proračuni kao što su oni koje zahtijeva geostatistika, jer jednostavno nije bilo računarskih resursa za to. Izrada karte je bila isključivno zasnovana na mentalnom modelu eksperta-pedologa, koji je u mnogome zavisio od iskustva samog pedologa. Kod geostatističke izrade karata se u značajnoj mjeri koriste precizno definisani algoritmi, ali je i dalje nophodno aktivno učešće čovjeka-eksperta, posebno u pripremi, obradi i kvalitativnoj evaluaciji polaznih podataka, kao i kod izbora konkretnih algoritama koji će biti korišteni za izradu karte.

## 2.2 Koraci pri izradi karte

Moguće je uočiti sljedeće korake prilikom izrade karte (Hengl 2009, str. 3):

[1] `http://iamg.org`

[2] `http://geoenvia.org`

[3] `http://pedometrics.org`

[4] U nastavku će se često pod interpolacijom podrazumijevati i ekstrapolacija, radi jednostavnosti.

1. planiranje uzorkovanja;
2. terensko prikupljanje podataka i laboratorijska analiza;
3. estimacija modela na osnovu prikupljenih podataka (kalibracija);
4. interpolacija na osnovu modela;
5. validacija interpoliranih podataka korištenjem validacionih podataka;
6. generisanje konačne verzije karte.

Ako se podaci kontinualno prikupljaju i unose u trajnu bazu podataka, npr. preko laboratorija za analizu zemljišta u poljoprivredne i druge svrhe, onda je potrebno, nakon dobijanja novih podataka, ponoviti neke od navedenih koraka da bi se generisala „svježa“ karta.

## 2.3 Geostatističko modeliranje varijabilnosti

U geostatistici se parametar prirodne sredine posmatra kao *signalni proces*[5] $Z$, koji se sastoji od tri komponente (Hengl 2009):

$$Z(\vec{s}) = m(\vec{s}) + \varepsilon'(\vec{s}) + \varepsilon''$$

gdje je $m(\vec{s})$ deterministička, $\varepsilon'(\vec{s})$ stohastička komponenta, dok je $\varepsilon''$ inherentni šum, izazvan uglavnom mjerenjem. $\vec{s}$ predstavlja lokaciju (najčešće dvodimenzionalnu). Kaže se da su konkretni uzorci $z(\vec{s}_1), z(\vec{s}_2), ..., z(\vec{s}_n)$ *realizacije* procesa $Z$.

Deterministička komponenta je poznata i kao *trend-površ* (*trend surface*), i ona daje informaciju kako se globalno, tj. na čitavoj posmatranoj površini ponaša parametar, tj. kakav je makro „trend“ parametra. Stohastička komponenta govori kako se parametar ponaša na lokalizovanom, tj. mikro nivou.

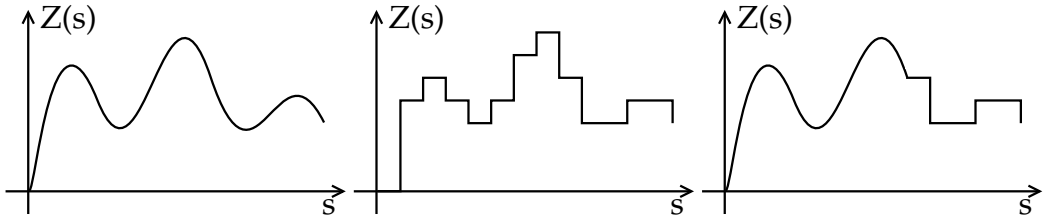
## 2.4 Aspekti varijabilnosti koji se često previđaju

U poglavlju 1 je već pomenuto da je matematičko modeliranje parametara zemljišta izuzetno kompleksno i nabrojani su glavni činioci koji se uzimaju u obzir

[5] U nastavku se koriste još i izrazi „signal“ i „proces“.

prilikom modeliranja. Međutim, postoje još neki aspekti (ne mogu se zvati činiocima) varijabilnosti koji značajno utiču na kvalitet modela, a često se, u većoj ili manjoj mjeri, namjerno (iz praktičnih razloga) ili nenamjerno previđaju. Neki od tih aspekata su (Hengl 2009):

- **Geografska 2D varijacija** Neki parametri sredine — kao što je dubina nekog horizonta zemljišta, tip zemljišta itd. — ne zavise od treće koordinate, dok drugi parametri zavise, kao što je temperatura koja se može mjeriti na različitim visinama u vazduhu, pa čak i na različitim dubinama u zemljištu. 2D varijacija se može modelovati kontinualnom i diskretnom funkcijom, ili kombinacijom, u zavisnosti od same prirode parametra (slika 2.1).




(a) Kontinualni signal (b) Diskretni signal (c) Kombinovani signal

Slika 2.1: Tipovi signala koji se koriste za modelovanje parametara koji imaju samo 2D varijabilnost.

- **Vertikalna (3D) varijacija** Ovaj aspekt je značajan kod parametara koji zavise od dubine i/ili visine. Već smo pomenuli da ovdje spada temperatura, a jedan od parametara koji zavisi od dubine je i prelaz između horizonata zemljišta koji može biti nagao ili postepen. Još jedan primjer je količina ugljen-monoksida (CO) u zemljištu (zavisnost od dubine) ili vazduhu (zavisnost od visine).

- **Vremenska varijacija** Ovaj vid varijacije je jako značajan jer je vrlo čest, doduše u različitom stepenu kod različitih parametara. Parametri zemljišta mogu varirati iz godine u godinu, od godišnjeg doba do godišnjeg doba, pa čak i od jednog doba dana do drugog. Primjer toga bi bila recimo temperatura zemljišta, ili količina vode u zemljištu. Sa druge strane, prilikom modelovanja nekog parametra radi njegove prognoze, često kao ulazne parametre koristimo parametre reljefa (DEM), koji se uglavnom sporo mijenjaju

sa vremenom. Još jedan primjer parametra sa specifičnom vremenskom varijacijom je biljni pokrivač. On se na neki način mijenja tokom godišnjih doba, a sa druge strane, imamo i sporu, ali značajnu promjenu tokom dužeg niza godina. Nasuprot tome, broj životinja u nekoj „tački", pa čak i oblasti, je toliko podložan vremenskoj varijaciji, da se često promijeni i za nekoliko minuta, usljed pokretljivosti životinja.

- **Veličina potpore (*Support size*)** Veličina potpore je površina ili zapremina za koju je jedno mjerenje nekog parametra vezano, kao i smjer, orijentacija i sl. prilikom mjerenja (kada to utiče na mjerenje).

Za dobre rezultate prognoze je neophodno što veće podudaranje svih mjerenja po navedenim aspektima (tamo gdje to ima smisla). Sa druge strane, ako su razlike među mjerenjima velike (npr. satelitski snimci biljnog pokrivača iz raznih godišnjih doba), onda prognoza pomoću takvih podataka može biti malo korisna, ili čak potpuno beskorisna.

## 2.5 Modeli prostorne prognoze

Kao što se iz dosadašnjeg izlaganja može zaključiti, nemoguće je da napravimo savršeni model nekog prostornog parametra, jer to podrazumijeva tačno poznavanje apsolutno svih činilaca koji utiču na njega, kao i pronalaženje precizne funkcionalne zavisnosti između parametra i činilaca. Ono što možemo uraditi je napraviti približan model (na osnovu izmjerenih vrijednosti), validirati ga i po potrebi usavršavati.

Sada ćemo dati preciznu definiciju modela prostorne prognoze (Hengl 2009, str. 9). Neka je:

$$z(\vec{s}_1), z(\vec{s}_2), ..., z(\vec{s}_n)$$

skup izmjerenih vrijednosti prostornog parametra $Z$, pri čemu su:

$$\vec{s}_i = (x_i, y_i), i \in \{1, ..., n\}, \vec{s} \in \mathbb{A}$$

lokacije uzoraka[6], a $n$ je broj mjerenja. Neka je $\gamma(\vec{h})$ ($\vec{h} \in \mathbb{A}$ je vektor[7] daljine između dvije tačke) neki model kovarijansi koje opisuju prostornu autokorelaciju

[6] $(x_i, y_i)$ su koordinate u nekom geografskom prostoru $\mathbb{A}$.

[7] U nastavku ćemo vektor često koristiti i kao matricu dimenzija $n \times 1$.

između izmjerenih vrijednosti. Pod pretpostavkom da su mjereni uzorci reprezentativni, ne-preferencijalni i konzistentni, model prostorne prognoze je onaj model koji predviđa vrijednost parametra u nekoj tački $\vec{s}_0$ na osnovu:

- uzoraka $z(\vec{s}_i), i \in \{1, ..., n\}$;
- ulaznih parametara (*deterministic predictors*) $q_k(\vec{s}_0)$ u tački $\vec{s}_0$, koji se nazivaju još i *kovarijatima* (*covariates*), a potrebno je da budu definisani $\forall \vec{s} \in \mathbb{A}$;
- modela kovarijansi $\gamma(\vec{h})$.

U praktičnom, tj. GIS[8] smislu, domen $\mathbb{A}$ se svodi na pravilnu pravougaonu mrežu tačaka (*grid*, raster), koja pokriva oblast gdje su vršena mjerenja.

Postoji veliki broj konkretnih modela prostorne prognoze koji potpadaju pod navedenu apstraktnu definiciju. U opsežnom pregledu modela (Li i Heap 2008), Li i Heap navode da su opisali preko 40. Međutim, većina tih modela je na neki način povezana, a mnogi linearni modeli su samo specijalan slučaj nekog opštijeg modela, pa poznavanje hijerarhije modela umnogome pomaže u efikasnom odabiru najboljeg modela za neki konkretan problem.

### 2.5.1 Klasifikacija modela prostorne prognoze

Jedan od načina da se modeli prostorne prognoze klasifikuju je prema količini statističke analize koju zahtijevaju (Hengl 2009, str. 11):

- **Prosti, ne-geostatistički, tj. „mehanički" (deterministički) modeli**: to su modeli gdje nema statističke analize, kao ni procjene greške i gdje najčešće nema strogog uslovljavanja varijabilnosti parametra. Neki od ovih modela su:
  - **Thiessen-ovi poligoni**,
  - **model inverzne udaljenosti (*Inverse Distance Weighting — IDW*)**,
  - **regresija nad koordinatama**,
  - **prirodni susjedi**,
  - **splajnovi (*splines*)**.

[8] Geografski Informacioni Sistem

- **Linearni statistički modeli**: kod ovih modela je prisutna statistička analiza parametara modela, kao i procjena greške prognoze. Međutim, to uslovljava i strožije uslove koje ulazni parametri moraju da zadovolje.

  Ovi modeli se dalje mogu podijeliti na sljedeće grupe:

  - **kriging**;
  - **modeli proste korelacije parametara** (npr. linearnom regresijom);
  - **modeli bazirani na Bayes-ovoj statistici**;
  - **hibridni modeli** (npr. regresija-kriging).

- **Ekspertski modeli**: tu spadaju modeli koji ne spadaju u ostale grupe i koji podrazumijevaju kompleksnu, ali nestatističku, analizu (bilo ljudsku, bilo računarsku) međuzavisnosti parametara:

  - **modeli pretežno zasnovani na znanju čovjeka** (npr. ručno pravljene karte);
  - **modeli pretežno zasnovani na podacima** (npr. bazirani na *neuronskim mrežama*);
  - **modeli zasnovani na mašinskom učenju** (tj. u potpunosti zasnovani na podacima).

## 2.6 Prosti modeli

### 2.6.1 Model inverzne udaljenosti

Ovo je jedan od najstarijih i najjednostavnijih modela interpolacije prostornih podataka (Shepard 1968), (Hengl 2009). Kod ovog modela procijenjena vrijednost parametra na nekoj lokaciji $\vec{s}_0$ je:

$$\hat{z}(\vec{s}_0) = \sum_{i=1}^{n} \lambda_i(\vec{s}_0) \cdot z(\vec{s}_i), \qquad \sum_{i=1}^{n} \lambda_i(\vec{s}_0) = 1$$

gdje je $\lambda_i$ težinski koeficijent za susjednu tačku rednog broja $i$.

Model je dobio naziv po načinu određivanja težinskih koeficijenata:

$$\lambda_i(\vec{s}_0) = \frac{\frac{1}{d^{\beta}(\vec{s}_0,\vec{s}_i)}}{\sum_{i=0}^{n} \frac{1}{d^{\beta}(\vec{s}_0,\vec{s}_i)}}, \qquad \beta > 1$$

tj. što je daljina $d(\vec{s}_0, \vec{s}_i)$ susjedne tačke veća, njen uticaj na prognozu je manji. Koeficijent $\beta$ omogućava podešavanje stepena uticaja srazmjerno daljini. Takođe, često se uzimaju u obzir samo tačke u okviru nekog radijusa $D$.

### 2.6.2 Regresija nad koordinatama

Ovaj model pretpostavlja da su parametri zavisni od lokacije tačke, i pokušava da pronađe neku funkciju $f(x, y)$, uz naravno $\vec{s} = (x, y)^\top$, koja će što bolje prekriti mjerene (tj. diskretne) vrijednosti parametra na datim lokacijama. Za ovaj model se koriste još i nazivi trend površi (*trend surfaces*), kada je regresija globalna, i interpolacija pomoću pokretne površi (*moving surface interpolation*), kada se iterativno vrši regresija nad samo dijelom cijele površine. Model je sljedeći (R. Webster i M. A. Oliver 2001, str. 40–41), (Hengl 2009):

$$Z(\vec{s}) = f(x, y) + \varepsilon$$

pri čemu se prognoza vrši pomoću polinoma:

$$\hat{z}(\vec{s}_0) = \sum_{r,v} a_{rv} \cdot x^r y^v$$

uz $r + v \leq p$, gdje je $p$ red polinoma. Koeficijenti $a_{rv}$ se pronalaze metodom najmanjih kvadrata (*Ordinary Least Squares*), tj. minimizacijom ukupne greške procjene:

$$\sum_{i=1}^{n} \left(\hat{z}(\vec{s}_i) - z(\vec{s}_i)\right)^2$$

tako da dobijamo:

$$\vec{a} = (\mathbf{S}^\top \cdot \mathbf{S})^{-1} \cdot (\mathbf{S}^\top \cdot \vec{z})$$

gdje je $\vec{z}$ vektor svih uzoraka $z_i, i \in \{1, ..., n\}$, a $\mathbf{S}$ matrica sa $n$ redova, čiji su redovi vektori sastavljeni iz svih članova $x_i^r y_i^v$, pri čemu je $r \in \{0, 1, ...\}$, $v \in \{0, 1, ...\}$, uz uslov $r + v \leq p$.

### 2.6.3 Splajnovi (*splines*)

Splajnovi (Hengl 2009, str. 14), (Unser 1999) su jedna vrsta parcijalnih, tj. dio-po-dio (*piecewise*) polinoma, koji omogućavaju bolje lokalizovano poklapanje sa polaznim mjerenim podacima, kao i preciznije ravnanje (*smoothing*). Neke od splajn tehnika koje su popularne u geo-naukama su *thin-plate splines* (Hutchinson

1995) i *regularized spline with tension and smoothing* (Mitasova i Mitas 1993), (Mitasova, Mitas i Harmon 2005), kod koje prognozirani parametar dobijamo kao:

$$\hat{z}(\vec{s}_0) = a_1 + \sum_{i=1}^{n} w_i \cdot R(v_i)$$

gdje je $R(v_i)$ (radijalna bazna funkcija):

$$R(v_i) = -\left[E_1(v_i) + \ln(v_i) + C_E\right]$$

, a $v_i$:

$$v_i = \left[\varphi \cdot \frac{\vec{h}_i}{2}\right]^2$$

pri čemu je $E_1(v_i)$ eksponencijalna integralna funkcija, $C_E = 0.577215$ Euler-ova konstanta, $\varphi$ generalizovani tenzioni parametar i $\vec{h}_i$ daljina od tačke 0 do tačke $i$.

Koeficijenti $w_i$ i parametar $a_1$ se mogu dobiti rješavanjem sljedećih jednačina:

$$\sum_{i=1}^{n} w_i = 0$$

$$a_1 + \sum_{i=1}^{n} w_i \cdot \left[R(v_i) + \delta_{ij} \cdot \frac{\omega_0}{\omega_i}\right] = z(\vec{s}_i)$$

gdje su $\omega_0/\omega_i$ pozitivni težinski faktori koji služe za ravnanje, tj. kontrolisanje vertikalne devijacije tačke 0 od tačke $i$. Tenzioni parametar $\varphi$ kontroliše daljinu do koje date tačke utiču na rezultujuću površ (Hengl 2009).

Ova vrsta splajnova daje rezultate slične univerzalnom krigingu (Hengl 2009).

## 2.7 Geostatistički modeli — kriging

Kao što je već rečeno, karakteristika geostatističkih modela je određeni stepen statističke analize koji usložnjava cjelokupni proračun, ali zato omogućava bolje rezultate, preciznije određivanje prirode greški u polaznim podacima i sl., (Hengl 2009).

Sinonim za geostatističke modele je kriging (ili krigovanje, od fr. *krigeage*), i to je pojam koji obuhvata čitavi niz međusobno srodnih modela. Nastao je iz potrebe sa jedne strane meteorologa da interpoliraju klimatske podatke iz rijetkih (*sparse*) podataka (npr. onih koje se bilježe u meteo stanicama), a sa druge strane inžinjera

rudarstva da procijene količinu korisnog materijala u određenom dijelu zemljine kore na osnovu rijetkih uzoraka rude (M. Oliver, Richard Webster i Gerrard 1989).

Ipak, iz razloga prije svega komercijalne prirode, kriging je najbrže razvijan u rudarstvu. Otkrio ga je Danie Gerhardus Krige (kasnije je i imenovan po njemu) prilikom istraživačkog rada za potrebe rudarske industrije u Južnoj Africi (Krige 1951). Međutim, on je objavio samo empirijske rezultate svog istraživanja, bez dublje teorijske analize, i tek je Matheron tokom 60-ih godina XX vijeka teorijski, pod imenom *teorija regionalizovanih varijabli* (Matheron 1971), uobličio dotadašnja istraživanja u Južnoj Africi i šire, i time postao utemeljivač moderne geostatistike (Matheron i Serra 2002). Otprilike u isto vrijeme (1963) sličnu tehniku je otkrio Lev Semenovich Gandin u Sovjetskom Savezu za potrebe meteorologije (Cressie 2015, str. 106). U (M. Oliver, Richard Webster i Gerrard 1989) se navodi da je prva formulacija nekog vida kriginga *metod optimalne interpolacije* Andreja Nikolaeviča Kolmogorova, 1941. godine (Shiryayev 1992), a u (Lichtenstern 2013) se tvrdi da je termin *optimalna linearna predikcija* koristio i Herman Ole Andreas Wold još 1938. godine.

Čak i kada su u pitanju kompletni, tj. „puni“ (*non-sparse*) podaci (npr. signal sa satelita), kriging može pomoći kod, recimo, skladištenja ovakvih podataka, jer skladištenje gotovo uvijek zahtijeva određeni stepen uzorkovanja originalnog signala, pa kriging može pomoći u procjeni optimalne učestalosti odabiranja, tj. one učestalosti koja će omogućiti kvalitetnu reprodukciju signala za ne-skladištene tačke (M. Oliver, Richard Webster i Gerrard 1989).

Matheron je svojom teorijom obuhvatno i matematički precizno opisao prostorno zavisnu promjenu ne samo procenta određenog materijala u zemljinoj kori, nego praktično ma kojeg prostorno zavisnog prirodnog parametra i time omogućio njenu primjenu u širokom spektru prirodnih nauka. Njeno detaljno matematičko izvođenje se može naći u, npr. (Lichtenstern 2013), a ovdje su dati samo izvodi iz teorije. Za kriging se koristi još i naziv *optimalni nepristrasni linearni prognostički model* za prostorne podatke (*Best Linear Unbiased Predictor — BLUP*).

I dok i kao ranije navedeni prostiji linearni modeli kriging koristi linearnu kombinaciju izmjerenih uzoraka parametra za prognostički model, ono što odlikuje kriging je da, prilikom računanja težinskih koeficijenata, koristi informaciju o prostornoj korelacionoj strukturi[9], koju modeluje preko statističkih parametara dru-

[9]Pod pojmom korelacione strukture podrazumijevamo kvanitativno izražen „uticaj“ okolnih

gog reda: tzv. variograma ili kovarijanse nepoznatog prostornog signala $Z(\vec{s})$. Pri tome, kriging opravdava epitet „nepristrasni“ tako što je očekivana vrijednost razlike između vrijednosti prognozirane modelom i stvarne vrijednosti 0:

$$\mathrm{E}\left[z(\vec{s}_0) - \hat{z}(\vec{s}_0)\right] = 0$$

a epitet „optimalni“ tako što je varijansa $\mathrm{Var}\left(z(\vec{s}_0) - \hat{z}(\vec{s}_0)\right)$ minimalna.

Ono što je zajedničko za sve tipove kriginga je da je model prostornog parametra kod svih njih sljedeći:

$$Z(\vec{s}) = m(\vec{s}) + \varepsilon'(\vec{s})$$

gdje je $m(\vec{s})$ deterministička, a $\varepsilon'(\vec{s})$ stohastička komponenta. Slijedeći podjelu kriging metoda datu u (Hengl 2009), možemo izdvojiti, tzv. „čiste“ kriging metode, tj. one kod kojih je deterministička komponenta konstantna, odnosno koji modeliraju praktično čisto stohastičke prostorne parametre, i hibridne metode, tj. one koji modeluju i determinističku komponentu.

### 2.7.1 Intrinzična stacionarnost — uslov validnosti kriginga

Iz razloga matematičke opravdanosti[10] prognostičkog modela, kriging algoritmi zahtijevaju od prostornog parametra koji se prognozira da zadovolji dva uslova[11], a to su:

1. da je očekivana vrijednost razlike $Z(\vec{s} + \vec{h}) - Z(\vec{s})$ jednaka nuli, $\forall \vec{s}$:

   $$\mathrm{E}\left[Z(\vec{s} + \vec{h}) - Z(\vec{s})\right] = 0$$

   gdje je $\vec{h}$ daljina, koja se naziva i zaostajanjem (*lag*).

2. da varijansa razlike $Z(\vec{s} + \vec{h}) - Z(\vec{s})$ bude konačna, tj. ograničena, i da zavisi isključivo od daljine $\vec{h}$, a ne od $\vec{s}$:

   $$\mathrm{Var}\left(Z(\vec{s} + \vec{h}) - Z(\vec{s})\right) = \mathrm{E}\left[\left(Z(\vec{s} + \vec{h}) - Z(\vec{s})\right)^2\right] = 2\gamma(\vec{h})$$

tačaka na vrijednost prostornog parametra u nekoj datoj tački.

[10] Tj. da bi prognostički model bio nepristrasan i optimalan.

[11] Navedeni uslovi se mogu neznatno razlikovati od jednog tipa kriginga do drugog.

Navedene dvije pretpostavke se pravdaju empirijskim iskustvom iz prirodnih nauka, tj. time da ih većina prostornih parametara prirodne sredine zadovoljava barem približno, ali se u (Hengl 2009) navodi da se vrlo rijetko provjeravaju u praksi, već ih geostatističari najčešće jednostavno uzimaju kao date. Za prostorni parametar koji zadovoljava ove uslove se kaže da je *intrinzično stacionaran*, stacionarnošću drugog reda.

### 2.7.2 Teorijski variogram

Ako su navedeni uslovi ispunjeni, onda funkciju $\gamma(\vec{h})$ nazivamo *teorijski variogram*, i on tada predstavlja alat kojim se precizno opisuje prostorna korelaciona struktura parametra $Z$, ili, što se prognoze tiče, on je precizna mjera kako susjedne tačke (definisane raznim vrijednostima daljine $\vec{h}$) prognostički „utiču“ na neku datu tačku $\vec{x}_0$. Kasnije će biti navedeno kako se variogram konkretno koristi u prognostičkim kriging modelima.

Međutim, u praksi praktično nikada nemamo precizno dat signal prostornog parametra, pa samim tim ni variogram, već imamo samo uzorke parametra za određeni broj lokacija. Stoga je potrebno iz tih uzoraka dobiti barem približnu verziju teorijskog variograma, koja se zatim može koristiti u prognostičkim modelima. To pronalaženje približnog variograma je jedan od najznačajnijih koraka prilikom kriginga i biće objašnjeno u nastavku, kroz primjer metoda običnog kriginga. U stvari, upravo relativno jednostavan postupak pronalaženja približne verzije teorijskog variograma i čini variogram najpogodnijim alatom za opisivanje prostorne korelacione strukture.

### 2.7.3 Obični kriging

Neki od čistih kriging metoda su, npr. *kriging srednje vrijednosti* (*kriging the mean*), *jednostavni kriging* (*simple kriging*), a ovdje će biti prikazan *obični kriging* (*ordinary kriging*), jer je najznačajniji i najopštiji „čisti“ kriging metod, a, može se koristiti i umjesto dva prvonavedena metoda.

Uslovi koje mora da zadovolji prostorni parametar su:

1. globalna srednja vrijednost parametra $Z(\vec{s})$ je nepoznata, ali konstantna i iznosi $\mu$, tj. model parametra je:

$$Z(\vec{s}) = \mu + \varepsilon'(\vec{s})$$

2. $Z(\vec{s})$ je intrinzično stacionaran, stacionarnošću drugog reda, pri čemu je variogram:

$$\gamma(\vec{h}) = \frac{1}{2}\,\mathrm{Var}\left(Z(\vec{s}+\vec{h}) - Z(\vec{s})\right) = \frac{1}{2}\,\mathrm{E}\left[\left(Z(\vec{s}+\vec{h}) - Z(\vec{s})\right)^2\right]$$

Ovdje je važno napomenuti da je konstantnost srednje vrijednosti parametra nešto što se rijetko srijeće u prirodi, ali je ona vrlo često približno ostvariva za neke manje oblasti. Slično je i sa intrinzičnom stacionarnošću. Hibridni kriging metodi pokušavaju da prevaziđu ova ograničenja običnog kriginga.

Takođe, pošto je pretpostavljeno da je $\mu$ konstantno, važe sljedeće jednakosti:

$$Z(\vec{s}_i) - Z(\vec{s}_j) = \varepsilon'(\vec{s}_i) + \mu - \varepsilon'(\vec{s}_j) - \mu = \varepsilon'(\vec{s}_i) - \varepsilon'(\vec{s}_j)$$

tj. kada god u formuli imamo razliku vrijednosti parametara za ma koje dvije lokacije to, u suštini, označava razliku vrijednosti stohastičke komponente za te lokacije, i obratno.

Prognoza se vrši slično kao kod modela inverzne daljine:

$$\hat{z}(\vec{s}_0) = \sum_{i=1}^{n} w_i(\vec{s}_0) \cdot z(\vec{s}_i) = \vec{\lambda}_0^{\top} \cdot \vec{z}$$

gdje je $\vec{\lambda}_0$ vektor (za sada nepoznatih) težinskih (ili kriging) koeficijenata ($w_i, i \in \{1, \ldots, n\}$) susjednih tačaka.

Da bi se ispunio uslov nepristrasnosti dovoljno je da je zbir težinskih koeficijenata 1:

$$\sum_{i=1}^{n} w_i(\vec{s}_0) = 1$$

jer se time dobija:

$$\begin{aligned} \mathrm{E}[\hat{z}(\vec{s}_0) - z(\vec{s}_0)] &= \mathrm{E}\left[\textstyle\sum_{i=1}^{n} w_i(\vec{s}_0) z(\vec{s}_i) - z(\vec{s}_0) \underbrace{\sum_{i=1}^{n} w_i(\vec{s}_0)}_{=1}\right] \\ &= \textstyle\sum_{i=1}^{n} w_i(\vec{s}_0) \underbrace{\mathrm{E}[z(\vec{s}_i) - z(\vec{s}_0)]}_{=0} = 0 \end{aligned}$$

gdje je $\mathrm{E}[z(\vec{s}_i) - z(\vec{s}_0)] = 0$, pošto je pretpostavljeno da je parametar $Z$ intrinzično stacionaran, a to, po ranije navedenoj definiciji, podrazumijeva da je očekivana vrijednost razlike $Z(\vec{s}+\vec{h}) - Z(\vec{s})$ jednaka nuli, $\forall \vec{s}$.

Da bi se ispunio uslov optimalnosti, potrebno je prije svega imati neku mjeru kvaliteta prognoze, a to je varijansa greške prognoze:

$$\begin{aligned}\sigma_e^2 &= \mathrm{Var}\left(\hat{z}(\vec{s}_0) - z(\vec{s}_0)\right) = \mathrm{E}\left[\left(\hat{z}(\vec{s}_0) - z(\vec{s}_0) - \mathrm{E}\left[\hat{z}(\vec{s}_0) - z(\vec{s}_0)\right]\right)^2\right] \\ &= \mathrm{E}\left[\left(\hat{z}(\vec{s}_0) - z(\vec{s}_0)\right)^2\right]\end{aligned}$$

gdje je $\mathrm{E}\left[\hat{z}(\vec{s}_0) - z(\vec{s}_0)\right] = 0$ usljed nepristrasnosti. Da bi prognoza bila optimalna, treba postići minimalnu varijansu greške prognoze.

Sljedeći cilj je izražavanje $\sigma_e^2$ kao funkcije kriging koeficijenata, a potom minimizacija tako dobijene funkcije, tj. pronalazak konkretnih vrijednosti koeficijenata za koje će funkcija imati minimalnu vrijednost. Daljim izvođenjem[12] se dobija:

$$\begin{aligned}\sigma_e^2 &= -\sum_{i=1}^{n}\sum_{j=1}^{n} w_i(\vec{s}_0) w_j(\vec{s}_0)\gamma(\vec{s}_i - \vec{s}_j) + 2\sum_{i=1}^{n} w_i(\vec{s}_0)\gamma(\vec{s}_i - \vec{s}_0) \\ &= -\vec{\lambda}_0^\top \mathbf{G}\vec{\lambda}_0 + 2\vec{\lambda}_0^\top\vec{\gamma}_0 = \vec{\lambda}_0^\top\left(2\vec{\gamma}_0 - \mathbf{G}\vec{\lambda}_0\right)\end{aligned}$$

gdje su $\gamma(\vec{s}_i - \vec{s}_j)$ semivarijanse[13] između pojedinačnih uzoraka međusobno, $\gamma(\vec{s}_i - \vec{s}_0)$ semivarijanse između pojedinačnih uzoraka i tačke za koju prognoziramo parametar, $\mathbf{G}$ simetrična matrica svih semivarijansi $\gamma(\vec{s}_i - \vec{s}_j)$, a $\vec{\gamma}_0$ vektor svih semivarijansi $\gamma(\vec{s}_i - \vec{s}_0)$. Ovdje, naravno, pretpostavljamo da je funkcija teorijskog variograma $\gamma(\vec{s})$ poznata.

Nakon minimizacije[14] $\sigma_e^2$ po $\vec{\lambda}_0$, dobija se sljedeća formula za vektor optimalnih koeficijenata:

$$\vec{\lambda}_0 = \mathbf{G}^{-1}\left[\vec{\gamma}_0 - \vec{1}\left(\frac{\vec{1}^\top\mathbf{G}^{-1}\vec{\gamma}_0 - 1}{\vec{1}^\top\mathbf{G}^{-1}\vec{1}^\top}\right)\right]$$

a odatle je optimalna varijansa:

$$\sigma_e^2 = \vec{\gamma}_0^\top\mathbf{G}^{-1}\vec{\gamma}_0 - \frac{\left(\vec{1}^\top\mathbf{G}^{-1}\vec{\gamma}_0\right)^2}{\vec{1}^\top\mathbf{G}^{-1}\vec{1}}$$

dok je prognozirana vrijednost parametra za lokaciju $\vec{s}_0$, po početnoj definiciji:

$$\hat{z}(\vec{s}_0) = \vec{\lambda}_0^\top \cdot \vec{z} = \left[\vec{\gamma}_0 - \vec{1}\left(\frac{\vec{1}^\top\mathbf{G}^{-1}\vec{\gamma}_0 - 1}{\vec{1}^\top\mathbf{G}^{-1}\vec{1}^\top}\right)\right]^\top \mathbf{G}^{-1}\vec{z}$$

[12] Detaljno izvođenje se može naći u (Lichtenstern 2013, str. 52).

[13] *Semivarijansa* je naziv za numeričku vrijednost teorijskog variograma za neko konkretno $\vec{h}$, u ovom slučaju $\vec{h} = \vec{s}_i - \vec{s}_j$. U (Bachmaier i Backes 2008) se navodi da prefiks „semi“ označava da je u pitanju iznos varijanse jedne tačke u paru, tj. pola varianse para.

[14] Detalji procesa minimizacije su dati u, recimo, (Lichtenstern 2013, str. 53-56).

### 2.7.4 Osobine teorijskog variograma

Kao što se iz prethodnih formula vidi, moguće je, poznavajući samo teorijski variogram nekog intrinzično stacionarnog parametra, dobiti optimalni linearni nepristrasni prognostički model na osnovu određenog broja uzoraka parametra.

Međutim, kako je već pomenuto, u praksi nikad nemamo zadat tačan teorijski variogram parametra, vać uvijek koristimo neku približnu, tj. estimiranu verziju[15]. Taj približni variogram će morati da ima neke osobine koje mora da ima i teorijski variogram ma kojeg intrinzično stacionarnog parametra, a to su:

1. $\gamma(\vec{0}) = 0$
2. $\gamma(\vec{h}) \geq 0$
3. $\gamma(\vec{h}) = \gamma(-\vec{h})$
4. $\gamma(\vec{h})$ raste sporije nego $|\vec{h}|^2$, tj.:

$$\lim_{|\vec{h}|\to\infty} \frac{\gamma(\vec{h})}{|\vec{h}|^2} = 0$$

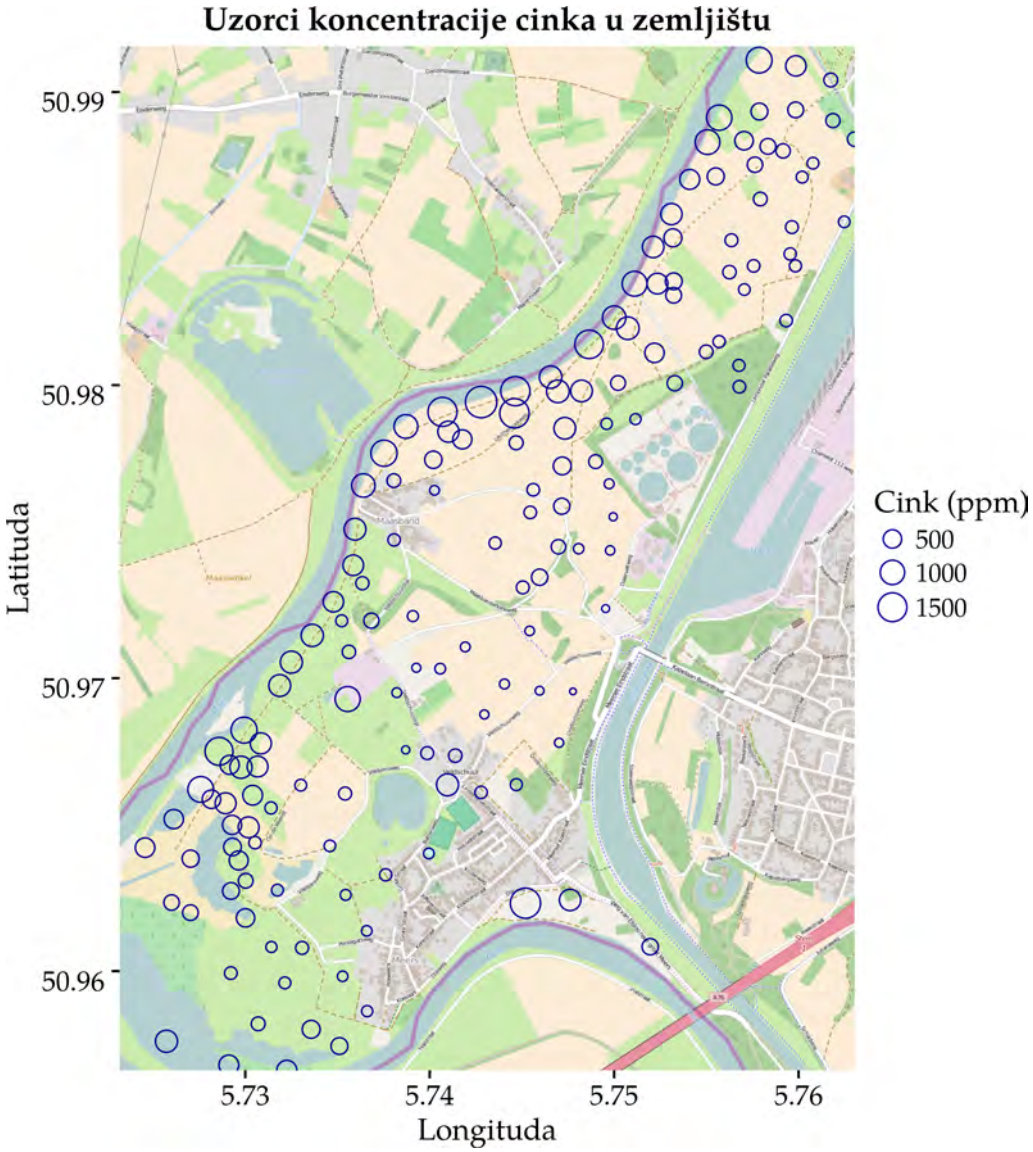
U narednom dijelu će postupno biti prikazan proces estimacije variograma. Neki koncepti će biti dati i vizuelno, radi boljeg razumijevanja, pri čemu će na prikazanim graficima biti korišten *Meuse* (francuski naziv za rijeku Meza) skup podataka (Pebesma 2004), koji se često koristi u primjerima u geostatističkim paketima programskog jezika R (R Core Team 2015). Kao prostorni parametar ćemo koristiti koncentraciju cinka u ppm (*parts-per-million*, koristićemo i logaritam ove vrijednosti) u zemljištu oko rijeke Meze, čija je karta uzoraka prikazana na slici 2.2.

### 2.7.5 Variogramski oblak

Do sada smo govorili o variogramu kao funkciji kontinualne promjenljive $\vec{h}$. Međutim, ograničeni broj uzoraka koje posjedujemo je diskretne prirode, pa, kao prvi korak analize, možemo posmatrati sljedeće diskretne iznose, koje ćemo nazivati semivarijansama[16]:

$$\gamma_{i,j} = \frac{\left(z(\vec{s}_i) - z(\vec{s}_j)\right)^2}{2}$$

[15]To takođe znači da će krajnji prognostički model odstupati u većoj ili manjoj mjeri od BLUP-a.
[16]Naravno, ove semivarijanse nisu identične semivarijansama teorijskog variograma.




Slika 2.2: Koncentracija cinka (u ppm) u zemljištu oko rijeke Meuse

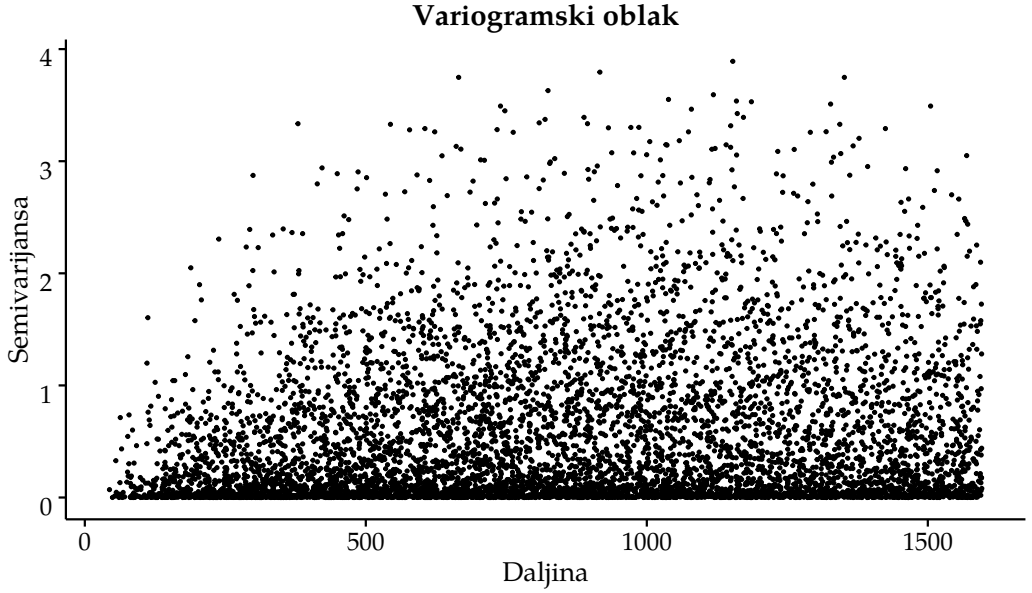
ili, drugačije zapisano:

$$\begin{aligned} \gamma_{i,j} &= \gamma_i(\vec{h}) &&= \frac{\left(z(\vec{s}_i) - z(\vec{s}_i - \vec{h})\right)^2}{2} \\ &= \gamma_j(-\vec{h}) &&= \frac{\left(z(\vec{s}_j) - z(\vec{s}_j + \vec{h})\right)^2}{2} \end{aligned}$$

Navedene vrijednosti, u općem slučaju, zavise i od lokacije i od smjera vektora daljine. Radi jednostavnosti, pretpostavićemo da je inherentni teorijski variogram

izotropan[17], pa tako u narednom koraku možemo posmatrati tačkasti grafik na kome se prikazuju gore navedene vrijednosti u funkciji samo dužine vektora $|\vec{h}|$, tj. ne uzimajući u obzir smjer. Taj grafik se naziva *variogramski oblak* (*variogram cloud*, primjer se može vidjeti na slici 2.3) i daje nam prvu vizuelnu, iako grubu, informaciju o stepenu porasta razlike među uzorcima sa porastom daljine $|\vec{h}|$.

Na grafiku se vidi da za ma koju određenu vrijednost $|\vec{h}|$ imamo veći broj vrlo različitih vrijednosti semivarijansi. To je i očekivano, jer variogramski oblak prikazuje konkretne, tj. realne, semivarijanse svakog para uzoraka, a one, naravno, odstupaju od idealnog teorijskog variograma, po kome bi svaki takav par trebao da ima istu semivarijansu. Ipak, i na ovom grafiku se intuitivno može uočiti da postoji trend blagog rasta semivarijansi i da bi se nekim vidom usrednjavanja mogla dobiti neka kontinualna funkcija koja bi opisivala rast varijanse, odnosno koja bi estimirala teorijski variogram.



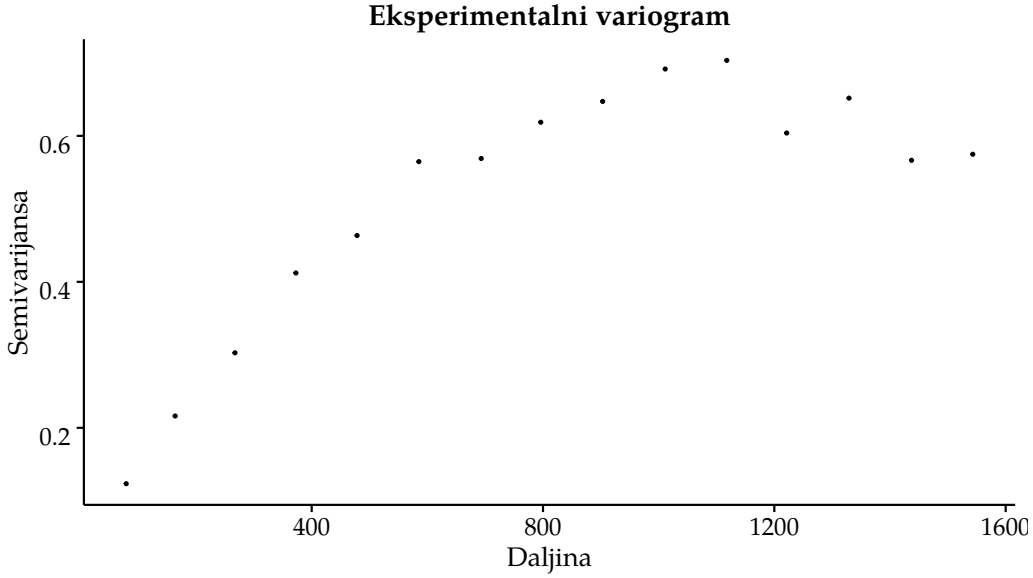

Slika 2.3: Variogramski oblak logaritma koncentracije cinka

Tako je sljedeći korak, tj. korak prije dobijanja konačnog estimiranog modela kontinualnog variograma, usrednjavanje diskretnih vrijednosti semivarijansi variogramskog oblaka. Takav variogram se naziva *eksperimentalni variogram*. Iako bi to usrednjavanje moglo da se vrši tako da za svako diskretno $|\vec{h}|$ dobijeno iz pa-

[17]Tj. ne zavisi od smjera i pravca vektora $\vec{h}$, već samo od inteziteta $|\vec{h}|$. Postoje i anizotropni modeli variograma, ali oni prevazilaze okvire ovog rada.

rova uzoraka usrednjimo semivarijanse, u praksi se to rijetko radi, jer bi srednje vrijednosti daljina sa malim brojem parova bile podložne uticaju *outlier*-a. Ono što se najčešće radi, usljed robusnosti, je usrednjavanje semivarijansi za vrijednosti $|\vec{h}|$ koje su inkrement neke osnovne daljine (npr. 100 m), tako što se usrednjavaju sve varijanse koje su između tih inkremenata. Npr. usrednjena semivarijansa za daljinu 100 m bi obuhvatila semivarijanse svih daljina koje su manje od 100 m, usrednjena semivarijansa za daljinu 200 m bi obuhvatila semivarijanse svih daljina koje su između 100 m i 200 m itd. Ovakvo usrednjavanje na neki način čini eksperimentalni variogram analognim histogramu.

Na slici 2.4 je dat grafik eksperimentalnog variograma logaritma koncentracije cinka oko rijeke Meze. Kao što se i očekivalo, semivarijanse su pretežno manje za male daljine, a veće za velike, s tim što stepen rasta opada, a na kraju čak prelazi u pad[18].



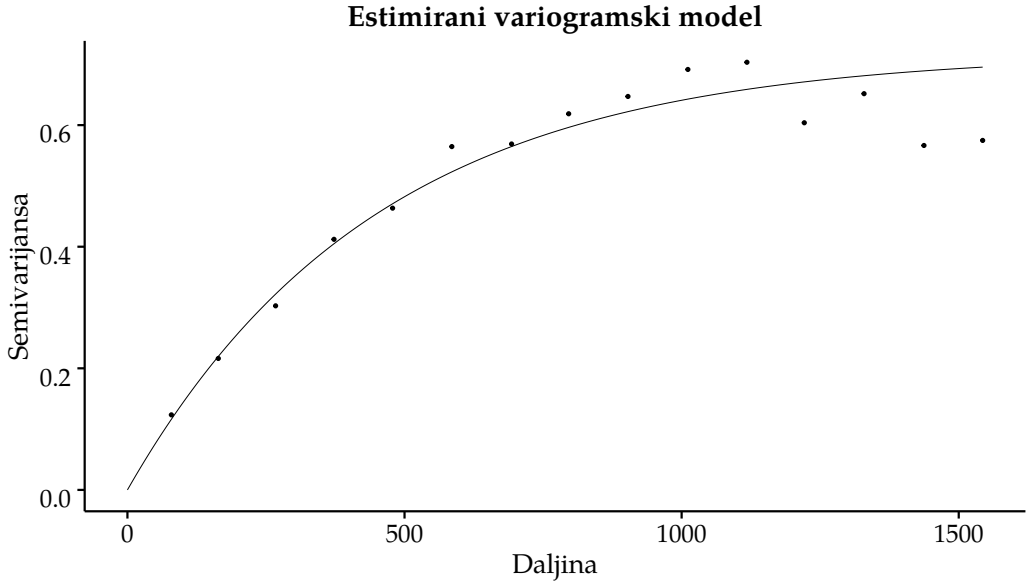

Slika 2.4: Eksperimentalni variogram logaritma koncentracije cinka

Posljednji korak je konačno dobijanje kontinualnog modela variograma, koji treba da se što bolje poklopi sa eksperimentalnim. On će predstavljati estimiranu verziju teorijskog variograma i može se koristiti za prognozu vrijednosti prostornog parametra u ma kojoj tački. Funkcija koja se koristi prilikom poklapanja se

[18]Za očekivati je, nakon određene daljine koja je veća od daljine najudaljenijeg para uzoraka ovog skupa podataka, da se vrijednosti „smire“ oko neke konstante koja predstavlja, tzv. globalnu semivarijansu.

ne bira nasumično, već postoje, tzv. autorizovani modeli variograma koji zadovoljavaju (prethodno pomenute) uslove koje zadovoljava i teorijski variogram. Neki od tih modela su: linearni, sferni, eksponencijalni, cirkularni, Gausov, Beselov, Maternov itd.

Autorizovani modeli variograma se datom eskperimentalnom variogramu prilagođavaju iterativno, a detalji tog postupka prevazilaze okvire ovog rada. Na slici 2.5 se vidi primjer variograma prilagođenog pomoću Maternovog modela za naš eksperimentalni variogram. Važno je napomenuti da, kao što se mogu dobiti različiti eksperimentalni variogrami za istu oblast, u zavisnosti od toga kako se vrši uzorkovanje, tako se mogu dobiti i različiti teorijski variogrami, a čak i za isti eksperimentalni variogram mogu se ponekad napraviti nekoliko različitih prilagođenih, u zavisnosti od toga koji je autorizovani model odabran.

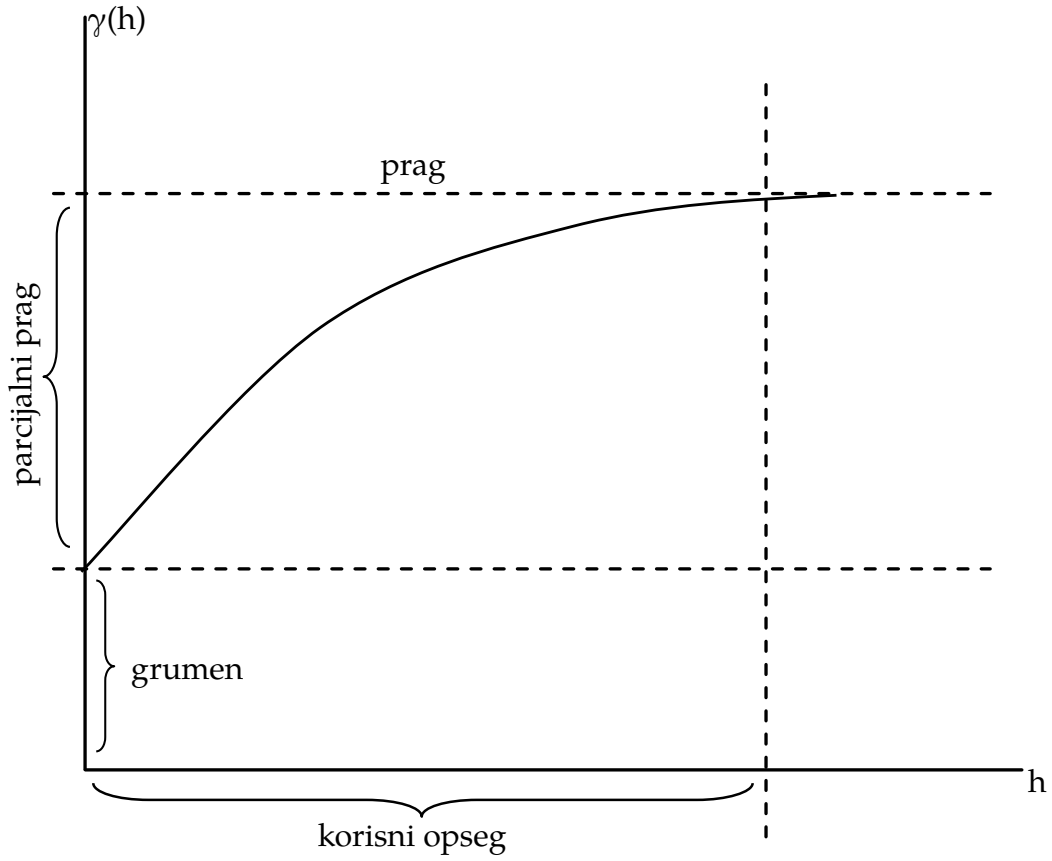



Slika 2.5: Prilagođeni (estimirani) variogram model eksperimentalnog variograma logaritma koncentracije cinka

Radi lakšeg proučavanja variograma, uvedeni su sljedeći parametri (slika 2.6):

- **grumen (*nugget*)**: vrijednost variograma za nultu daljinu;
- **prag (*sill*)**: vrijednost kojoj variogram asimptotski prilazi za $|\vec{h}| \to \infty$;
- **parcijalni prag (*partial sill*)**: razlika između praga i grumena;

- **korisni opseg (*practical range*)**: daljina na kojoj variogram iznosi 95% praga.




Slika 2.6: Parametri variograma

## 2.8 Modeli bazirani na ambijentalnoj korelaciji

U odjeljku 1.4 je već govoreno o korelaciji parametara zemljišta sa parametrima prirodne sredine (ambijent), kao i o nekim primjerima matematičkog formulisanja te korelacije. Ova korelacija je posebno važna kada postoje lako dostupni ambijentalni parametri, jer su parametri zemljišta uglavnom teže dostupni usljed većih troškova njihovog prikupljanja — neki od ambijentalnih parametara su često dostupni potpuno besplatno, sa zadovoljavajućom prostornom rezolucijom. Ambijentalna korelacija, slično krigingu, omogućava prognozu parametara zemljišta za tačke u kojima nemamo parametre zemljišta. Međutim, ona je uglavnom korisna

kod prognoze determinističke komponente signala, nasuprot krigingu, koji je pogodan za prognozu stohastičke komponente. Takođe, kod čistih kriging modela, ulazni parametri su u suštini koordinate (odnosno daljina među tačkama), dok su kod modela čiste ambijentalne korelacije ulazni parametri samo parametri sredine, bez koordinata. Model koji se koristi kod ambijentalne korelacije je (Hengl 2009, str. 20):

$$Z(s) = f\left(q(\vec{s})\right) + \varepsilon$$

gdje su sa $q$ označeni svi ambijentalni parametri pomoću kojih vršimo prognozu, dok $\varepsilon$ je stohastička komponenta, uključujući i šum.

Gore navedeni model je u suštini analogan *clorpt* modelu (1.4), tj. baziran je na pedogenetskim faktorima. U (Budiman Minasny i Alex.B. McBratney 2016) se navodi da se od 40-ih godina 20. vijeka *clorpt* model koristio i na kvantitativan način, tj. istraživanjem inherentne funkcionalne zavisnosti zemljišta od faktora korištenjem konkretnih numeričkih podataka, i da je čak i sam autor modela Jenny 1968. godine objavio značajan, iako začuđujuće slabo citiran, rad u kojem utvrđuje zavisnost nekoliko parametara zemljišta od više ambijentalnih faktora pomoću multivarijatne linearne regresije, a nakon izvršene *analize osnovnih komponenti* (*Principal Component Analysis — PCA*), na sljedeći način:

$$S = a + k1 \cdot MAP + k2 \cdot MAT + k3 \cdot substrat + k4 \cdot nagib + k5 \cdot vegetacija + k6 \cdot latituda$$

gdje $S$ predstavlja više osobina zemljišta, $a, k1, ..., k6$ su konstante koje je autor utvrdio istraživanjem, $MAP$ srednja godišnja količina padavina, a $MAT$ srednja godišnja temperatura.

Međutim, gore navedeni model je u radu korišten samo za objašnjenje ambijentalne zavisnosti, bez ikakvog govora o mogućnosti generisanja mapa. U (Budiman Minasny i Alex.B. McBratney 2016) se navodi da su tek devedesetih godina 20. vijeka Arrouays i Pelissier u Francuskoj koristili regresioni algoritam za prognozu količine ugljenika u zemljištu na osnovu ambijentalnih i drugih podataka. Jedan od razloga tako kasne primjene u generisanju karata je dotadašnja nedovoljna brzina računara[19].

[19](Budiman Minasny i Alex.B. McBratney 2016) navode podatak da se u posljednje vrijeme, usljed razvoja računara, kao i pod uticajem projekta *GlobalSoilMap*, broj obrađenih piksela u projektima zemljišnog mapiranja udvostručava svake dvije godine.

Funkcionalna zavisnost između ambijentalnih parametara i zemljišta je izuzetno kompleksna. Često značajno varira od oblasti do oblasti, od jednog godišnjeg doba do drugog i sl., što uslovljava ponovno računanje prognostičkog modela za svaku situaciju posebno. Naravno, iako se kaže da model varira od oblasti do oblasti, to ne znači da izlazni parametar direktno zavisi od koordinata, već prosto u nekim oblastima postoji specifična združena zavisnost izlaznog parametra od više ulaznih parametara.

Iako kompleksni, ambijentalni modeli su uzročno-posljedični, tj. omogućavaju mnogo bolje suštinsko razumijevanje pojava u prirodi, što nije slučaj sa čistim kriging modelima. Na primjer, bolje razvijena zemljišta će postojati tamo gdje je nagib terena mali i gdje ima dosta akumulacije vode; neki tipovi šuma će biti karakteristični za određene nagibe; hladna i vlažna klima će obično usloviti zemljište sa većim sadržajem organske materije i sl.

Postoji više tipova ambijentalnih prognostičkih modela, a najvažniji su sljedeći (Hengl 2009, str. 21):

- **Klasifikacioni modeli**: koriste se prije svega kod izlaznih parametra diskretnog tipa, kao što je tip zemljišta, tip biljnog pokrivača i slično. Takođe postoji podjela ovih modela na osnovu toga da li se koristi precizna ili *fuzzy* klasifikacija. Klasa izlaznog parametra može biti određena centrom i standardnom devijacijom (od centra) ili pomoću skupa pravila koji jednoznačno određuju svaku klasu.

- **Modeli bazirani na, tzv. drvetu odlučivanja (*decision-tree*)**: kod ovih modela generiše se struktura podataka drvo, koja je u suštini hijerarhijska predstava nekog parametra (npr. klase zemljišta, ali se drvetom može predstaviti i neki kontinualni parametar). Korijen drveta sadrži sve ostale podgrupe (grane), a što se ide bliže listovima, to je izlazni parametar preciznije definisan. Takođe se, na osnovu ulaznih parametara i njihovih opsega, određuju pravila koja neku konkretnu tačku ili oblast na karti smještaju u drvo. Drvo odlučivanja je robustan model, i može se koristiti i u situacijama kada se praktično ništa ne zna o inherentnom regresionom modelu. Neke od mana su to što zahtijeva veliki broj ulaznih tačaka i što, osim ako se ne kombinuje sa nekim drugim modelom, uopšte ne koristi koordinate ulaznih tačaka.

- **Regresioni modeli**: u regresionoj analizi se koristi grupa funkcija koje se na-

zivaju Generalizovanim Linearnim Modelima (GLM-ovi). Korišćenjem ovih modela dobijaju se u konačnici regresioni koeficijenti, koji utvrđuju linearnu zavisnost između izlaznog parametra i ulaznih parametara. Postoje i regresioni modeli koji se koriste za prikazivanje nelinearne zavisnosti, a koji se nazivaju Generalnim Aditivnim Modelima (GAM-ovi). U ovu grupu modela se mogu ubrojiti i neuronske mreže, koje implicitno utvrđuju koeficijente zavisnosti (odnosno čitavu mrežu koeficijenata zavisnosti) između ulaznih i izlaznih parametara.

Jedan od najjednostavnijih i najčešće korišćenih modela regresije je višeparametarska linearna regresija (*Multiple Linear Regression*), gdje se izlazni parametar za lokaciju $\vec{s}_0$ estimira pomoću (Hengl 2009, str. 22):

$$\hat{z}(\vec{s}_0) = \hat{b}_0 + \hat{b}_1 \cdot q_1(\vec{s}_0) + ... + \hat{b}_p \cdot q_p(\vec{s}_0) = \sum_{k=0}^{p} \hat{b}_k \cdot q_k(\vec{s}_0)$$

pri čemu su sa $q_k, k \in \{0, ..., p\}$ označeni ulazni parametri (uz $q_0(\vec{s}_0) = 1$), a $\hat{b}_k$ su regresioni koeficijenti. Matrično:

$$\hat{z}(\vec{s}_0) = \vec{\hat{b}}^{\top} \cdot \vec{q}(s_0)$$

Regresioni koeficijenti[20] se dobijaju pomoću metoda najmanjih kvadrata:

$$\vec{\hat{b}} = \left(\mathbf{q}^{\top} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\top} \cdot \vec{z}$$

gdje je $\mathbf{q}$ matrica dimenzija $n \times p+1$ koja sadrži vrijednosti svih ulaznih parametre za sve ulazne tačke (tj. tačke gdje je izlazni parametar uzorokovan), a $\vec{z}$ vektor svih $n$ uzorkovanih vrijednosti izlaznog parametra.

Greška prognoze je tada:

$$\hat{\sigma}^2(\vec{s}_0) = MSE \cdot \left[1 + \vec{q}^{\top}(\vec{s}_0) \cdot \left(\mathbf{q}^{\top} \cdot \mathbf{q}\right)^{-1} \cdot \vec{q}^{\top}(\vec{s}_0)\right]$$

pri čemu je $MSE$ (*Mean Squared Error*) rezidualna *srednja kvadratna greška* oko regresione linije:

$$MSE = \frac{\sum_{i=1}^{n} \left[z(\vec{s}_i) - \hat{z}(\vec{s}_i)\right]^2}{n-2}$$

Kao što je već rečeno, problem proste linearne regresije prostornih podataka je to što uopšte ne uzima u obzir koordinate uzoraka, a jedan od načina njegovog

[20] Date su samo konačne formule, bez izvođenja.

prevazilaženja jeste korišćenjem regresionih koeficijenata koji su lokalizovatni, tj. na neki način zavise od lokacije uzorka. Takva regresija se naziva geografski prilagođena regresija (*geographically weighted regression*) i kod nje se regresioni koeficijenti dobijaju sa (Hengl 2009):

$$\vec{\hat{b}} = \left(\mathbf{q}^\top \cdot \mathbf{W} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^\top \cdot \mathbf{W} \cdot \vec{z}$$

gdje je $\mathbf{W}$ matrica težinskih koeficijenata, čiji se članovi računaju pomoću neke od funkcija čija vrijednost opada sa udaljenošću, recimo:

$$w_i(\vec{s}_i, \vec{s}_j) = \sigma_E^2 \cdot \exp\left[-\frac{1}{2} \cdot \left(\frac{d(\vec{s}_i, \vec{s}_j)}{D}\right)^2\right]$$

pri čemu je $\sigma_E^2$ nivo varijacije greške, $d(\vec{s}_i, \vec{s}_j)$ Euklidova daljina između tačaka ($\vec{s}_i$ i $\vec{s}_j$), a $D$ je opseg uticaja — konstanta koja odražava uticaj udaljenih tačaka — što veća vrijednost $D$, to manji uticaj udaljenih tačaka na vrijednost koeficijenta. Ono što je glavna mana osnovne verzije geografski prilagođene regresije je to što je odabir vrijednosti $D$ u suštini proizvoljan. Ipak, uz malo podešavanja, recimo, isprobavanjem više vrijednosti opsega uticaja u kombinaciji sa pomjerajućim prozorom, mogu se postići dosta dobri rezultati, s tim što se neki inherentni problemi ovog modela, kao što je pretjerano ravnanje, ne mogu prevazići. (Hengl 2009, str. 23) navodi još neke radove koji se bave problemima (i preporukama za prevazilaženje istih) ovog tipa regresije.

## 2.9 Hibridni modeli

Ovi modeli nastaju kombinacijom različitih osnovnih modela, determinističkih i stohastičkih, u pokušaju da se što preciznije opiše zavisnost parametara zemljišta od pedogenetskih faktora sa jedne, i geostatističke prostorne strukture sa druge strane. Ovi modeli su u trenutku pisanja ovog rada *state of the art*, jer je većina istraživanja u ovoj oblasti danas fokusirana na ove modele. Hibridni modeli se grubo mogu podijeliti na (Budiman Minasny i Alex.B. McBratney 2016):

- univerzalni kriging,
- regresija-kriging (*regression-kriging*),
- linearne kombinovane modele (*linear mixed models*),
- kokriging (*cokriging*).

### 2.9.1 Univerzalni kriging

Ovo je najkompleksniji hibridni kriging metod koji će biti izložen u ovom radu. Odabran je baš ovaj hibridni metod, jer ga je još Matheron otkrio (Matheron 1971), a i usljed toga što je od hibridnih kriging metoda samo on eksplicitno implementiran u R biblioteci gstat koja je korištena u ovom radu za generisanje karata.

Pretpostavljeni model prostornog parametra kod ovog tipa kriginga je malo drugačiji nego kod običnog kriginga:

1. $Z(\vec{s})$ se može razložiti na determinističku i stohastičku komponentu koje zavise od lokacije:
$$Z(\vec{s}) = m(\vec{s}) + \varepsilon'(\vec{s})$$

2. stohastička komponenta je intrinzično stacionarna, sa nultom srednjom vrijednošću i variogramom $\gamma(\vec{h})$, koju možemo zvati i *rezidualnom variogram funkcijom* parametra $Z(\vec{s})$, odnosno, $\forall \vec{s}, \vec{h}$ važe sljedeće jednakosti:
$$\mathrm{E}\left[Z(\vec{s})\right] = \mathrm{E}\left[m(\vec{s})\right] + \underbrace{\mathrm{E}\left[\varepsilon'(\vec{s})\right]}_{=0} = m(\vec{s})$$
$$\gamma(\vec{h}) = \frac{1}{2}\mathrm{Var}\left(\varepsilon'(\vec{s}+\vec{h}) - \varepsilon'(\vec{s})\right) = \frac{1}{2}\mathrm{E}\left[\left(\varepsilon'(\vec{s}+\vec{h}) - \varepsilon'(\vec{s})\right)^2\right]$$

3. diksretna komponenta je linearna kombinacija $p$ različitih, prostorno zavisnih diskretnih komponenti:
$$m(\vec{s}) = \sum_{k=0}^{p} b_k \cdot q_k(\vec{s}) = \vec{q}_{\vec{s}}^{\top}\vec{b}$$
gdje je $\vec{q}_{\vec{s}} = (q_0(\vec{s}), q_1(\vec{s}), ..., q_p(\vec{s}))^{\top}$, uz $q_0(\vec{s}) = 1$.

Vidimo da je model diskretne komponente u potpunosti ambijentalan, dok je model stohastičke komponente u suštini model običnog kriginga, sa $\mu = 0$, tj. potpuno geostatistički.

Ako prostorni parametar $Z(\vec{s})$ zadovoljava gore navede uslove, i ako imamo uzorkovane vrijednosti parametra sa $n$ lokacija, $z(\vec{s}_i), i \in \{1, 2, ..., n\}$, onda njegovu vrijednost na nekoj lokaciji $\vec{s}_0$ možemo prognozirati pomoću:
$$\hat{z}(\vec{s}_0) = \sum_{i=1}^{n} w_i(\vec{s}_0) \cdot z(\vec{s}_i) = \vec{\lambda}_0^{\top} \cdot \vec{z}$$

gdje su sve oznake iste kao kod običnog kriginga. Slično kao kod običnog kriginga, potrebno je pronaći $\vec{\lambda}_0$ tako da model prognoze bude nepristrasan i optimalan. Detaljno izvođenje se može pronaći u (Lichtenstern 2013, str. 63-74), a prije nego što damo konačni rezultat, uvešćemo nekoliko novih oznaka.

Po definiciji, uz označavanje $\vec{q}_{\vec{s}_i}$ sa $\vec{q}_i$ radi jednostavnosti, vrijednosti parametra za pojedinačne uzorke možemo da izrazimo kao:

$$z(\vec{s}_i) = m(\vec{s}_i) + \varepsilon'(\vec{s}_i) = \vec{q}_i^{\top}\vec{b} + \varepsilon'(\vec{s}_i)$$

odnosno za svih $n$ uzoraka:

$$\vec{z} = \mathbf{Q}\vec{b} + \vec{\varepsilon}'$$

gdje je $\mathbf{Q}$ matrica kojoj su pojedinačni redovi transponovani vektori:

$$\vec{q}_i^{\top} = (q_0(\vec{s}_i), q_1(\vec{s}_i), ..., q_p(\vec{s}_i)), i \in \{1, 2, ..., n\}$$

Tada su za univerzalni kriging konačni izrazi za koeficijente $\vec{\lambda}_0$ i varijansu greške prognoze $\sigma_e^2$:

$$\begin{aligned}
\vec{\lambda}_0 &= \mathbf{G}^{-1}\left[\vec{\gamma}_0 - \mathbf{Q}\left(\mathbf{Q}^{\top}\mathbf{G}^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{Q}^{\top}\mathbf{G}^{-1}\vec{\gamma}_0 - \vec{q}_0\right)\right] \\
\sigma_e^2 &= \vec{\gamma}_0^{\top}\mathbf{G}^{-1}\vec{\gamma}_0 - \left(\mathbf{Q}^{\top}\mathbf{G}^{-1}\vec{\gamma}_0 - \vec{q}_0\right)^{\top}\left(\mathbf{Q}^{\top}\mathbf{G}^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{Q}^{\top}\mathbf{G}^{-1}\vec{\gamma}_0 - \vec{q}_0\right)
\end{aligned}$$

dok je prognozirana vrijednost parametra za lokaciju $\vec{s}_0$:

$$\hat{z}(\vec{s}_0) = \vec{\lambda}_0^{\top}\vec{z} = \left[\vec{\gamma}_0 - \mathbf{Q}\left(\mathbf{Q}^{\top}\mathbf{G}^{-1}\mathbf{Q}\right)^{-1}\left(\mathbf{Q}^{\top}\mathbf{G}^{-1}\vec{\gamma}_0 - \vec{q}_0\right)\right]^{\top}\mathbf{G}^{-1}\vec{z}$$

### 2.9.2 Napomene o univerzalnom krigingu

U literaturi i radovima koji obrađuju kriging postoji blaga konfuzija oko terminologije, pa se često za univerzalni kriging koriste i nazivi *kriging sa driftom* i *kriging sa trendom*.

Kao što se vidi iz krajnjih formula, teorijski prognostički model univerzalnog kriginga zahtijeva unaprijed poznavanje linearnog ambijentalnog modela determinističke komponente, tj. matricu $\mathbf{Q}$ i vektor $\vec{q}_0$. Takođe, potrebno je unaprijed poznavati variogramsku prostornu strukturu stohastičke komponente, tj. matricu $\mathbf{G}$ i vektor $\vec{\gamma}_0$. Međutim, u praksi nam oni praktično nikada nisu unaprijed poznati, već je potrebno da ih na neki način estimiramo. U (Matheron 1971, str. 188) čak postoji poglavlje naslova „*The Indeterminability of the Underlying Variogram*“,

koje govori o nemogućnosti egzaktnog estimiranja inherentnog variograma u prisustvu drifta $m(\vec{s})$.

Mi ćemo u ovom radu, za potrebe generisanja karata, koristiti neke od besplatno dostupnih estimacionih tehnika za linearni model i variogramsku strukturu programskog jezika R, bez ulaženja u detalje načina njihovog rada.

### 2.9.3 Regresija-kriging

Kod nekih geostatističara je popularan regresija-kriging metod, kod koga se deterministička i stohastička komponenta modeluju posebno, a zatim se sabiraju da bi se dobio krajnji prognostički model. To omogućava korištenje ma kog determinističkog modela[21], npr. neuronskih mreža, *slučajnih šuma* (*random forests*) itd. U tom slučaju bi se univerzalni kriging mogao smatrati specijalnim slučajem regresija-kriginga[22], s tim što treba imati na umu da su koraci dolaska do krajnjeg prognostičkog modela različiti kod oba metoda, jer to u praksi može da utiče na krajnji rezultat.

## 2.10 Validacija prognostičkih modela

### 2.10.1 Validacija kontinualnih parametara

Varijansa greške prognoze koju smo dobili za obični i univerzalni kriging nam daje samo procjenu greške prognostičkog modela, tj. ne daju nam stvarnu grešku modela. Stvarna mjera greške se može dobiti jedino mjerenjem stvarnih vrijednosti nekog parametra u tačkama za koje smo vršili prognozu (a koje nismo koristili za dobijanje prognostičkog modela), i upoređivanjem tih vrijednosti sa prognoziranim. Taj proces se zove *validacija*, a tačke nad kojima je vršena validacija se nazivaju *validacione tačke*.

Navešćemo nekoliko veličina koje se koriste za iskazivanje greške prognoze (Hengl 2009, str. 25). Najčešće se koriste srednja greška prognoze $ME$ (*Mean Error*):

$$ME = \frac{1}{l} \cdot \sum_{j=1}^{l} \left[ \hat{z}(\vec{s}_j) - z^*(\vec{s}_j) \right]; \quad E[ME] = 0$$

[21] Univerzalni kriging koristi samo linearni model.

[22] Vidjeti, npr. (Hengl 2009, str. 36) za matematički dokaz ekvivalentnosti.

i korijen srednje kvadratne greške prognoze $RMSE$ (*Root Mean Square Error*):

$$RMSE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^{l} \left[\hat{z}(\vec{s}_j) - z^*(\vec{s}_j)\right]^2}; \quad E[RMSE] = \sigma(\vec{h} = 0)$$

gdje je $l$ ukupan broj validacionih tačaka.

Takođe se može računati i normalizovana $RMSE$:

$$RMNSE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^{l} \left[\frac{\hat{z}(\vec{s}_j) - z^*(\vec{s}_j)}{\hat{\sigma}_j}\right]^2}; \quad E[RMNSE] = 1$$

gdje je $\hat{\sigma}_j$ procijenjena greška prognoze za datu tačku.

Zatim se može koristiti i normalizacija totalnom varijansom parametra za mjerene tačke, i ona se često koristi za poređenje sa drugim modelima (pa čak i za druge parametre), a i kao mjera koliko model objašnjava varijabilnost parametra:

$$RMSE_r = \frac{RMSE}{\sigma_t}$$

Po (Hengl 2009, str. 25), $RMSE_r$ od oko 40% znači dosta pozdanu prognozu, dok $RMSE_r$ već od 71% ukazuje da model objašnjava samo 50% varijabilnosti.

Još jedna mjera je *koeficijent determinacije* $R^2$ (*coefficient of determination*), za koji se kaže da daje informaciju o tome koliko je „varijanse objašnjeno“ (Hengl 2009, str. 23):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

gdje je $SS_{res}$ (*sum of squares of residuals*) zbir kvadrata reziduala (greški prognoze):

$$SS_{res} = \sum_{i=1}^{n} \left(z(\vec{s}_i) - \hat{z}(\vec{s}_i)\right)^2$$

a $SS_{tot}$ ukupni zbir kvadrata (*total sum of squares*)

$$SS_{tot} = \sum_{i=1}^{n} \left(z(\vec{s}_i) - \mu)\right)^2$$

pri čemu je $\mu$ srednja vrijednost parametra $z(\vec{s})$.

Iz definicije $R^2$ slijedi:

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_z^2}$$

gdje je $\sigma_e^2$ varijansa greški prognoze (reziduala), a $\sigma_z^2$ varijansa uzoraka.

Treba napomenuti da, iako se prilikom računanja gore navedenih veličina koriste stvarne mjerene vrijednosti, opet se u tom procesu koristi samo ograničeni skup tačaka, a i samo mjerenje sadrži u sebi određenu grešku.

Iz praktičnih razloga, tačke u kojima se vrše mjerenja za potrebe validacije se najčešće ne uzimaju naknadno, već se uglavnom vrši *unakrsna validacija* (*cross-validation*), tj. za validaciju se koristi podskup originalnog skupa tačaka. Sa obzirom na to kako se bira podskup za validaciju, postoji više tehnika validacije, od kojih su neke (Hengl 2009, str. 25):

- *k-segmentna validacija* (*k-fold*), kod koje se originalni skup tačaka dijeli na $k$ podskupova, i onda se interpolacioni algoritam izvšava $k$ puta, pri čemu se svaki put koristi novi podskup za validaciju, a preostalih $k - 1$ podskupova za proračun modela;

- *izuzev-jednog validacija* (*leave-one-out — LOO*), kod koje se u svakoj iteraciji jedna tačka izostavlja prilikom proračuna modela, a zatim se ta tačka koristi za validaciju modela;

- *britvanje* (*jackknifing*), kod koga se, slično *LOO* validaciji, u svakoj iteraciji koristi jedna tačka, ali ne za validaciju prognoze, već za računanje statističke pristrasnosti (*bias*).

Rezultati validacije se koriste za pronalaženje problematičnih tačaka, npr. onih koje za nekoliko standardnih devijacija odskaču od normalizovane vrijednosti prognostičke greške. Kod jako problematičnih, tj. neprečišćenih, podataka *LOO* validacija zna da često da čudne rezultate. U praksi se preporučuje 10-segmentna validacija (Hengl 2009, str. 26).

Još jedan problem koji se može javiti kod unakrsne validacije je taj da validacioni podskup nije nezavisan od originalnog skupa, tj. da zavisi od načina na koji je originalni plan uzorkovanja dizajniran. Ako uzorkovanje nije bilo reprezentativno, onda validacioni podskup neće biti nezavisan, pa će rezultati unakrsne validacije biti pristrasni. Kod proizvoljnog uzorkovanja ne postoji taj problem, jer je proizvoljan podskup proizvoljnog skupa sam po sebi nepristrasan.

### 2.10.2 Validacija kategoričkih parametara

Za procjenu tačnosti prognoze kategoričkih parametara se koristi *kappa statistika* (Hengl 2009), koja upoređuje procenat utvrđene (*observed*) podudarnosti prognozirane i stvarne vrijednosti $p_o$ sa jedne strane, i očekivani procenat slučajne (*by chance*) poudarnosti $p_c$ na sljedeći način (Foody 2004):

$$\kappa = \frac{p_o - p_c}{1 - p_c}$$

Procenat podudarnosti prognoze se može računati i za svaku klasu ponaosob kao (Hengl 2009):

$$p_{oc} = \frac{N_{oc}}{N_c}$$

gdje je $N_c$ ukupan broj tačaka za koje je prognozirano da pripadaju klasi $c$, a $N_{oc}$ ukupan broj tačaka tačno svrstanih pod klasu $c$.

Ono što je još značajno kod kappa statistike je da se $\kappa$ koeficijenti za različite nezavisne prognostičke modele mogu upoređivati međusobno (Foody 2004).

# Poglavlje 3

# Digitalizacija pedoloških podataka Crne Gore

## 3.1 Uvod

U ovom poglavlju će biti dat kratak pregled procesa digitalizacije pedoloških podataka za teritoriju Crne Gore, a dio tih podaka će biti korišten u poglavlju 4 za izradu tematske karte lako pristupačnog fosfora za definisanu oblast na jugu Crne Gore. Opis procesa digitalizacije je značajan jer je njegov konačni rezultat baza[1] pedoloških podataka sa unosima za više hiljada profila širom Crne Gore, tako da uvid u proces digitalizacije u suštini daje uvid u kvalitet digitalizovanih podataka i ideje za dalje unapređenje istih. Više detalja o ovom procesu se može naći u (Salković 2015).

## 3.2 Digitalizacija ručno rađene ekspertske karte

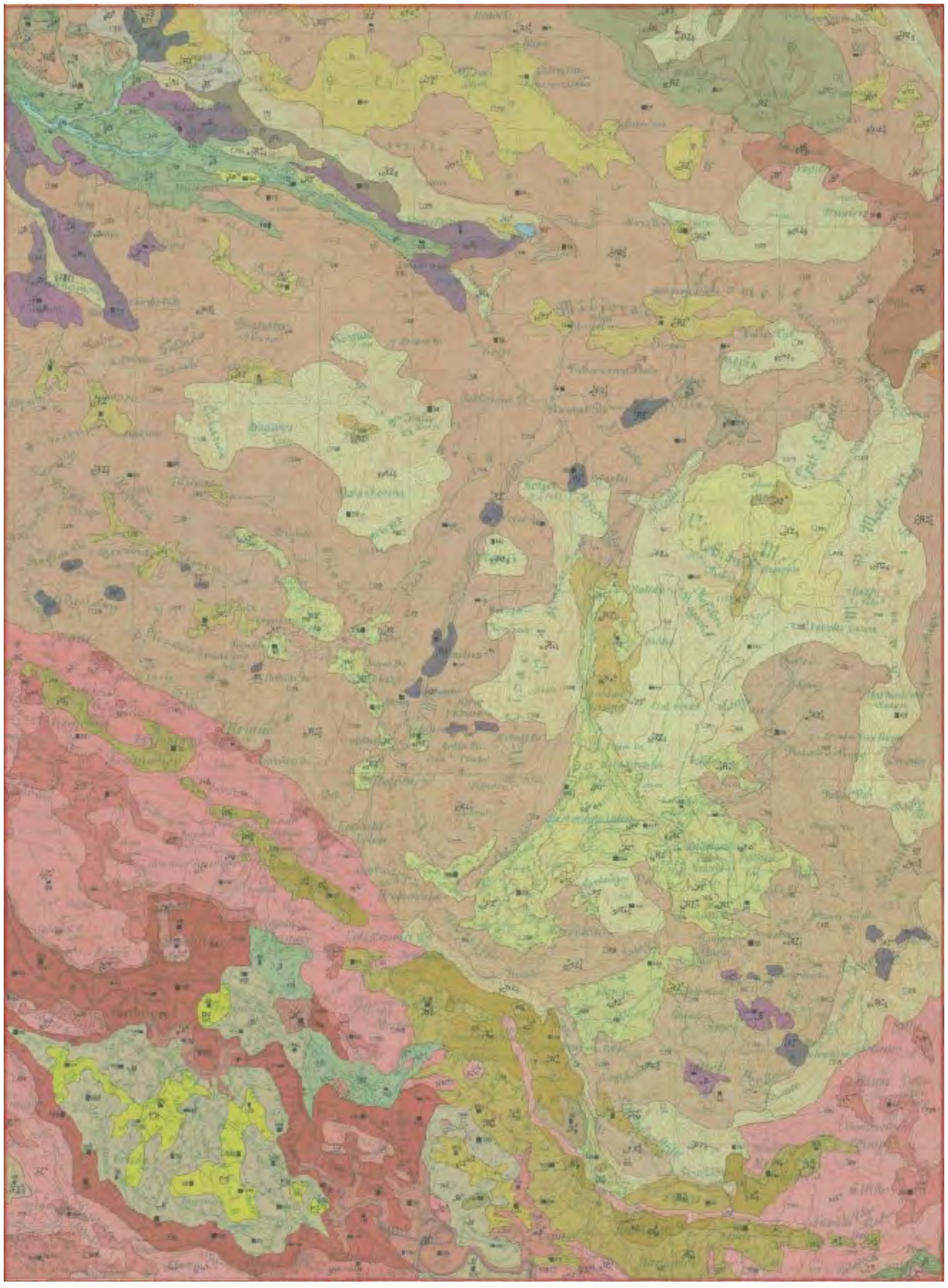
Jedan od ključnih rezultata pedoloških istraživanja iz skore prošlosti je produkcija niza ručno rađenih ekspertskih pedoloških karata razmjere 1:50000 koje prikazuju tipove zemljišta u Crnoj Gori. Karte pokrivaju cijelu površinu Crne Gore (vidjeti sliku 3.1), a na njima je obilježen i veliki broj profila koji su planski otvarani radi izrade ovih karata. Te pojedinačne pedološke karte (tzv. listovi ili sekcije; vidjeti, npr. sliku 3.2 za sekciju „Nikšić 4“) su u novije doba skenirane i

[1] Preciznije rečeno tabela u bazi, ali radi jednostavnosti, biće korišten izraz baza čak i kada se misli na tabelu.

sačuvane u elektronskom georeferenciranom rasterskom formatu, a čak su i koordinate obilježenih profila sačuvane u elektronskom tekstualnom tabelarnom formatu, sa sačuvanom originalnom numeracijom profila. Autor ovog rada nije učestvovao u procesu skeniranja karte i georeferenciranja profila, ali je učestvovao u kasnijoj digitalizaciji numeričkih podataka skladištenih u sveskama, kao i povezivanju tih podataka sa odgovarajućim georeferenciranim profilima.


Slika 3.1: Ručno rađena ekspertska pedološka karta Crne Gore, dobijena skeniranjem pojedinačnih listova, a zatim njihovim spajanjem u jedinstvenu kartu

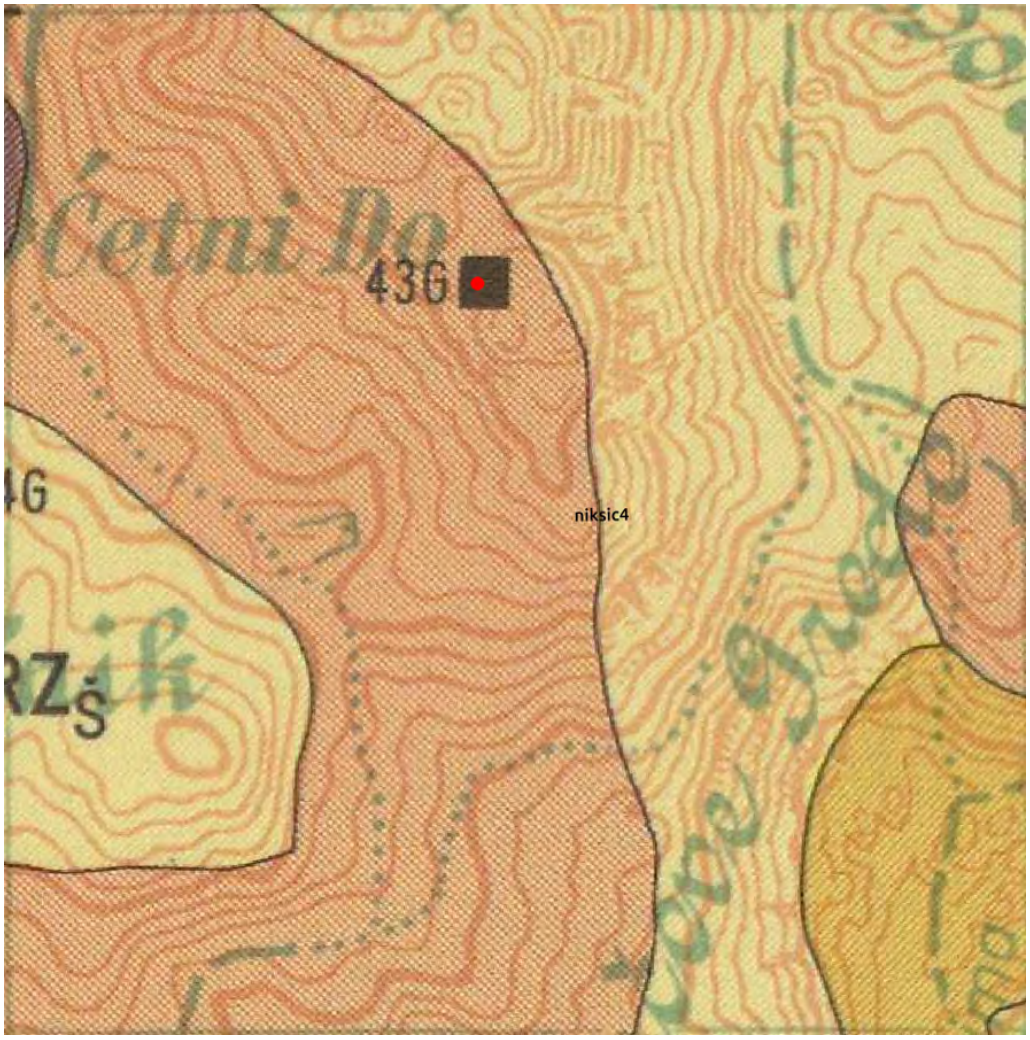
Na karti su prisutne sekcije: Višegrad 3, Žabljak 1, Žabljak 2, Žabljak 3, Žabljak



Slika 3.2: Sekcija „Nikšić 4“ pedološke karte Crne Gore

4, Pljevlja 1, Pljevlja 3, Pljevlja 4, Gacko 2, Gacko 3, Gacko 4, Sjenica 3, Trebinje 1, Trebinje 2, Trebinje 3, Trebinje 4, Nikšić 1, Nikšić 2, Kolašin 1, Kolašin 2, Peć 1, Peć 2, Peć 3, Nikšić 3, Nikšić 4, Kolašin 3, Kolašin 4, Kotor 1, Kotor 2, Cetinje 1, Cetinje

2, Cetinje 3, Cetinje 4, Skadar 1, Skadar 3 i Ulcinj-Lješ. U okviru svake sekcije imamo više, tzv. kvadrata. Na slici 3.3 je prikazan, kao primjer, kvadrat „G7“ koji se nalazi u okviru sekcije „Nikšić 4“, i u kome se nalazi profil oznake „43G“, blizu lokaliteta „Ćetni Do“.




Slika 3.3: Kvadrat „G7“ u okviru sekcije „Nikšić 4“, u kome se nalazi profil „43G“, blizu lokaliteta „Ćetni Do“

Bitno je napomenuti da je značajan problem korištenja ovako skenirane karte nepreciznost. Npr. na slici 3.4 se može vidjeti rezolucija preciznosti obilježavanja profila, dok se na slici 3.5 vidi jedan primjer (ne)poklapanja *Google maps* satelitskog snimka sa skeniranom kartom. Nepreciznost podataka dobijenih putem ovako skenirane karte utiče na kvalitet ma kojeg interpolacionog algoritma.


Slika 3.4: Primjer obilježenog profila na skeniranoj mapi koja je postavljena kao providni overlay iznad *Google maps* satelitskog snimka

## 3.3 Digitalizacija numeričkih podataka

Pedološki podaci na teritoriji Crne Gore su prvobitno prikupljani od 1959. do 1984. godine, a kao rezultat tog projekta nastale su ručno pisane sveske (primjerak stranice je dat na slici 3.6) od kojih 3 sadrže hemijske, a 3 sadrže mehaničko-fizičke osobine zemljišta. Postoje još i podaci koje je prikupljao tim hrvatskih pedologa, a koji se nisu nalazili u prethodno navedenim sveskama, kao i podaci koji su objavljeni u (Fuštić i Đuretić 2000), čiji se dobar dio već nalazi u 6 svesaka, kao i u hrvatskim podacima. Ovo je stoga jer su podaci za knjigu (Fuštić i Đuretić 2000) birani tako da budu reprezentativni za cijelu Crnu Goru, a navedene sveske su bile glavni izvor autora.

Podaci su iz svesaka prekucavani u *Excel* fajlove, a u tome je učestvovalo oko 100 unosilaca, unoseći sve zajedno oko 1000 stranica (tj. dvostranica kada su u pitanju sveske). Jedan unosilac je sam unio oko 300 stranica, u knjizi (Fuštić i Đuretić 2000) je bilo oko 200 stranica (već otkucanih), a ostali unosioci su unosili u rasponu od par stranica do 12 stranica, sa prosjekom oko 6. Određeni broj stranica

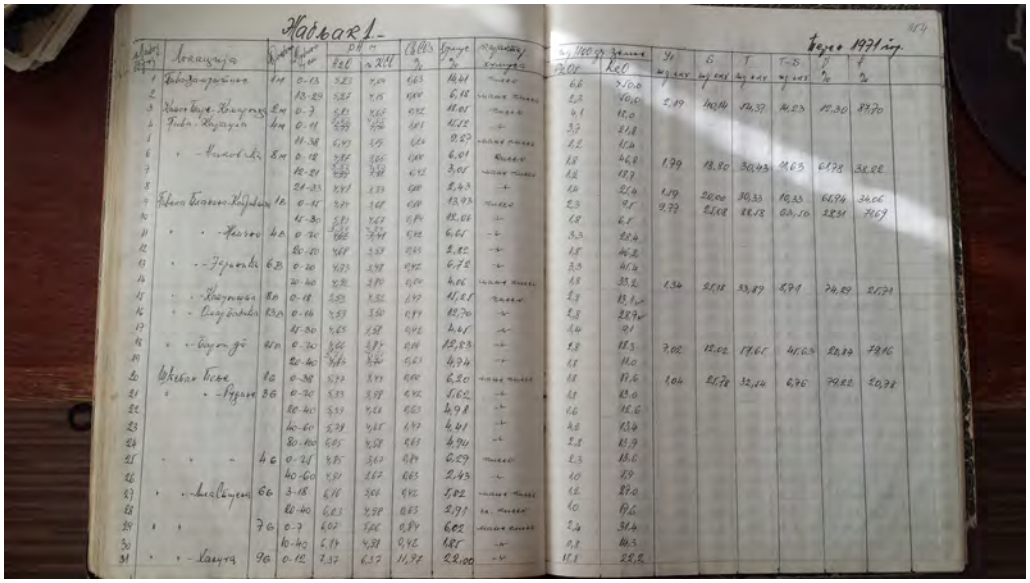

Slika 3.5: Primjer odstupanja pedološke karte od satelitskog snimka

je dva puta kucan, radi preciznosti. Konačan[2] broj unešenih redova (što MP, što C) je oko 26000.

U početku, prilikom unosa, usljed nedovoljne obučenosti unosilaca i nepostojanja standardizovanog *Excel* fajla kojeg bi koristili svi unosioci, a i zbog kompleksnosti samog sadržaja svesaka, napravljeno je dosta grešaka koje su otežale objedinjavanje svih fajlova u jedinstvenu bazu. Takođe, usljed *Excel*-ovih „pametnih“ opcija ispravljanja podataka i prepoznavanja tipa podatka,[3] ogroman broj čisto

[2]Konačnost treba uzeti sa rezervom, jer su naravno moguće greške u obradi postojećih podataka, kao i pronalazak novih.

[3]U ovom poglavlju će, umjesto tekstualnih tabela, biti dati snimci (*screen snapshot*-ovi) direktno iz *Excel*-a, radi bolje ilustracije.



Slika 3.6: Stranica ručno pisane sveske sa pedološkim podacima

numeričkih unosa je označen kao datum (slika 3.7), a usljed razlike u interpretaciji zareza[4] ponekad je jednocifreni broj sa decimalnim ostatkom prepoznavan kao broj reda veličine hiljade. Ovo su samo neki primjeri problema sa podacima koje su ukucavali unosioci.

| Lab. br. | Lokacija | Profil | Dubina u cm |
|---|---|---|---|
| 252 | Međurečje Cerovica | 28p | 0-19 |
| 253 | Međurečje | 29p | Feb-20 |
| 254 | Međurečje | 30p | 0-18 |
| 255 | Sretenska Gora dolovi | 31p/a | 0-18 |
| 256 | Sretenska Gora ? | 33p | Jan-15 |
| 257 | Pelev Brijeg Seoštica | 34p | 0-17 |
| 258 | Pelev Brijeg Seoštica | | 17-32 |
| 259 | Pelev Brijeg Seoštica | 35p | 0-17 |

Slika 3.7: *Excel*-ovo „pametno“ prepoznavanje unosa kao datum

Radi gore navedenog, napisan je kompleksan parser[5] koji je analizirao sve dostupne tabelarne fajlove, i imao zadatak da detaljno opiše svaki red prilikom unosa

[4]U engleskom jeziku se zarez koristi za označavanje hiljada, miliona itd., a ne kao decimalni zarez.

[5]Ovdje će se pod pojmom parser podrazumijevati računarski program koji „čita“ *Excel* fajlove i pretvara pojedinačne redove svakog fajla u redove koji su pogodni za skladištenje u relacionu bazu podataka.

u bazu podataka, čime bi bilo olakšano pronalaženje greške u kucanim podacima, kao i pronalaženje odgovarajućeg podatka u originalnim sveskama. Npr. podatak o tipu sveske (MP ili C), rednom broju sveske, broju/rasponu stranica, i imenu osobe koja je unosila podatke je „kodiran" u sami naziv fajla, npr. „C_1, str 079-081, Petar Petrović.xlsx"[6], što je omogućilo parseru da te podatke jednostavno „izvuče" iz naziva fajla. Takođe, u folderu svake sveske je smješten podfolder „images", u koji su smještene sve skenirane slike pojedinačnih stranica te sveske. Konačno, unutar samih *Excel* fajlova je naznačavano, na parseru razumljiv način, kojoj stranici određeni redovi pripadaju (slika 3.8). Sve navedeno je omogućilo parseru da automatski zabilježi sve potrebne detalje o nekom unosu, a zatim, recimo prilikom nailaska na grešku, da parser automatski otvori u *Excel*-u problematični fajl, prikaže na ekranu skeniranu sliku dotične stranice, pa čak i precizno prikaže broj problematičnog reda i u *Excel* fajlu, i na skeniranoj stranici.

| Lab.br. | Lokacija | Profil | Dubina u cm | pH u $H_2O$ |
|---|---|---|---|---|
| 1779 | | | 30-50 | 5.86 |
| Str 65 | | | | |
| 1780 | Bijelo Polje (Rasovo) | 11 | 70-90 | 5.8 |
| 1781 | | | 100-130 | 5.95 |
| 1782 | | 12 | 0-20 | 6.47 |
| 1783 | | | 30-60 | 6.1 |
| 1784 | | | 70-100 | 6.44 |
| 1785 | | 13 | 0-30 | 5.54 |
| 1786 | | | 60-90 | 5.54 |
| 1787 | | | 100-130 | 6.09 |
| Str 66 | | | | |
| 1788 | Jezero Malo Blato (Vodoprivreda-Titograd) | 1 | 0-0.40 | 6.57 |
| 1789 | | | 0.40-1.30 | 6.88 |
| 1790 | | | 1.30-1.60 | 6.48 |
| 1791 | | | 1.30-2.30 | 6.8 |

Slika 3.8: Stranice u sveskama su eksplicitno naznačavane u *Excel* fajlovima

Kod skladištenja redova u bazu, najbitnije je bilo utvrditi validnost određenog reda, tj. da li red sadrži korisne MP i C podatke, ili pak neki komentar i sl., što bi ga činilo nevalidnim. Pošto je prva kolona u svesci „laboratorijski broj", tj. redni broj uzorka zemlje koji je originalno ispitivan u laboratoriji, a on je prisutan u svakom validnom redu, provjera validnosti *Excel* reda se svela na provjeru validnosti rednog broja[7].

[6]Fajl sadrži podatke sa stranica 79–81 sveske C1, a podatke je unosio Petar Petrović.

[7]Postojao je određeni broj slučajeva gdje je red i bez laboratorijskog broja ipak bio validan, i koje je parser prepoznavao.

Naravno, nije bilo dovoljno provjeriti validnost samo laboratorijskog broja, već i validnost većeg dijela MP i C podataka, i to je rađeno pomoću *regularnih izraza* (*regular expressions*) i drugih vidova prepoznavanja stringova. Primjera radi, *Dorsijev koeficijent*[8], je bilježen u nekoliko varijanti eksponencijalne notacije:

- $1,05 * 10 - 4$
- $1,05 * 10 - 4$
- 1.9 x 10 na $-5$

što je parser uzimao u obzir.

Ono što je važno napomenuti je da je prilikom nailaženja na greške u podacima, greška uvijek ispravljana u originalnim *Excel* fajlovima unosilaca, a zatim bi parser iznova generisao konačnu bazu podataka. Iako se ovaj način može na prvi pogled činiti bespotrebno repetitivnim[9], ipak se njime omogućava jednostavnije ispravljanje grešaka, a i *Excel* fajlovi bolje odslikavaju originalne sveske. Takođe, repetitivnost se mogla i izbjeći malo složenijim parsiranjem, ali usljed brzog izvršavanja čitavog procesa, nije bilo potrebe za time.

### 3.3.1 Napredna provjera koherentnosti podataka

Pošto su svi podaci objedinjeni u bazi, moguće je bilo izvršiti neke naprednije provjere koherentnosti podataka, a nakon toga i ispravku ili promjena tih podataka.[10] Pod naprednijim provjerama se ovdje smatraju provjere više međuzavisnih redova i kolona odjednom i sl. Neke od ovih provjera je bilo moguće izvršiti i prilikom parsinga[11], ali je preglednije kada se vrše nad cjelokupnom bazom, jer je tada moguće lančano povezati više međusobno nezavisnih provjera. Lančano povezivanje provjera je bilo dodatno olakšano i time što je SQLite baza samo jedan fajl, čija veličina u ovom slučaju nije prolazila nekoliko MB, pa je nakon svake promjene mogao biti generisan novi SQLite fajl, čime je omogućena jednostavna inspekcija svakog koraka u lancu SQL upitima i sl. Na primjer, evo nekoliko vezanih naprednih provjera pomoću skripti u programskom jeziku *Python*:

[8] MP karakteristika.

[9] Jer parser svaki put mora da iznova učitava sve fajlove.

[10] U nastavku pojam provjera podrazumijeva i ispravku, odnosno promjenu.

[11] Neke i jesu vršene tako.

```
1 python check_lab_no.py C_1.sqlite
2 python fix_profile.py C_1.sqlite
3 python categories2numbers.py C_1.fix_profile.sqlite
4 python norm.py C_1.fix_profile.sqlite
```

gdje prva skripta provjerava validnost laboratorijskog broja baze hemijskih karakteristika za knjigu C1, druga „sređuje“ numeraciju profila, treća pretvara kategoričke vrijednosti u brojčane, a četvrta normalizuje sve numeričke parametre. Neke od gore navedenih skripti generišu nove SQLite fajlove, koji se zatim koriste u narednim skriptama.

Ovdje ćemo kao primjer jedne konkretne provjere opisati provjeru numeracije profila. Prvobitni unosioci podatka u sveske su svakom profilu dodjeljivali broj, ali bez preciznog redosljeda, tako da je prisutan veliki broj profila sa brojem 1, 2, 3 itd., uz eventualni sufiks, najčešće inicijal unosioca, a često su brojevi i preskakani. Međutim, sami profil sadrži nekoliko horizonata zemljišta[12], tako da svaki red numerisan laboratorijskim brojem odgovara jednom horizontu nekog profila (slika 3.9).

| Lab. br. | Lokacija | Profil | dubina u cm | Granulometrijski sastav u % | | |
|---|---|---|---|---|---|---|
| 1998 | Jezero-Malo Blato | 26 | 9.9--11.3 | 0 | 0.38 | 50.84 |
| 1999 | | | 11.3--12.1 | 0 | 4.27 | 51.38 |
| 2000 | | | 12.1--15 | 0 | 9.58 | 41.62 |
| 2001 | | 27 | 0--1.6 | 0 | 0.14 | 20.03 |
| 2002 | | | 3.35--4.35 | | | |
| 2003 | | | 6.6--8.4 | | | |
| 2004 | | | 8.4--10.2 | 0 | 1.29 | 36.81 |
| 2005 | | | 10.2--12.1 | 0 | 3.28 | 61.12 |
| 2006 | | | 12.1--14.15 | 0 | 14.02 | 29.33 |
| 2007 | | | 14.55--16.9 | 0 | 20.21 | 48.79 |

Slika 3.9: Primjer numeracije profila i horizonata

Horizonti su unošeni od plićeg ka dubljem, ali je podatak o broju profila bilježen samo kod prvog, najplićeg horizonta, a identično tome su postupili i unosioci u *Excel*, tako da je bilo potrebno provjeriti redove sa neunešenim brojem profila da li su poređani od plićeg ka dubljem, i ako jesu, dodijeliti im profilni broj prvog horizonta u nizu. Paralelno sa ovim je svakom profilu dodijeljivan i jedinstveni

[12]Npr. horizont od 9.9 do 11.3 cm, horizont od 11.3 do 12.1 cm itd.

redni broj u okviru cijele baze podataka, a slično je rađeno i sa rednim brojem reda.

Još jedan primjer napredne provjere je prepoznavanje oznake „-II-“[13], gdje je bilo potrebno prekopirati sadržaj analognog unosa iz prethodnog reda. Ono što još uvijek nije riješeno usljed niskog stepena važnosti jesu unosi[14] gdje ima više oznaka „-II-“, ili gdje je oznaka „-II-“ miješana sa tekstom, npr. „-II- Dudovo brdo“, a gdje treba prepoznati da se oznaka „-II-“ odnosi na riječ „Nikšić“ iz prethodnog unosa “Nikšić-Litije”.

U primjere napredne provjere podataka se može ubrojati i obrada kolona koje sadrže kategoričke podatke. Npr. C kolona „Karakter humusa“, gdje su kategorije bile „kiseo“, „slabo kiseo“, „neutralan“, itd. Ovakve kolone je za potrebe geostatičkih algoritama neophodno pretvoriti u redne brojeve-indekse[15]. Dosta softverskih alata može samostalno da odradi ovu konverziju, ali, usljed prirode podataka, ponekad je bolje odraditi „ručno“ konverziju. Recimo, kategorija „slabo kiseo“ je značenjski bliža kategoriji „kiseo“ negoli kategoriji „vrlo slabo kiseo“ i to bi dodijeljeni indeks trebao da oslikava, tako da, npr. „kiseo“ dobije indeks 1, „slabo kiseo“ indeks 2, a „vrlo slabo kiseo“ indeks 3. Još jedan „manji“ problem koji se javio kod kategoričkih podataka je bilo i objedinjavanje međusobno ekvivalentnih unosa, npr. pretvaranje „sl. kiseo“ u „slabo kiseo“ i sl.

### 3.3.2 Objedinjavanje svih podataka u jedinstvenu bazu

Ono što je bio vjerovatno najteži zadatak prilikom digitalizacije je bilo objedinjavanje MP i C podataka u jedinstvenu bazu. Naime, MP i C podaci su uglavnom unošeni u odvojene sveske, ali to i ne bi bio toliki problem da su laboratorijski brojevi unosa u MP i C sveskama bili isti, ili da je bar redosljed unosa bio isti. Međutim, u MP i C sveske 1, 2 i 3 su podaci najčešće unošeni odvojeno, tj. bez međusobno zajedničkog redosljeda unosa[16], tako da je bilo potrebno naknadno povezati odgovarajuće redove iz jednog i drugog skupa. Sa druge strane, hrvatski podaci, i podaci iz (Fuštić i Đuretić 2000) su najčešće upisivani po istom redosljedu, što je umnogome olakšalo povezivanje, koje se uglavnom svodilo na

[13]Sa značenjem: „isto kao prethodno“.
[14]Najčešće prisutni u koloni „lokacija“.
[15]Npr. od 1 do 10.
[16]Laboratorijski broj ne bi bio isti.

jednostavnu konkatenaciju podataka.

Ispostavilo se da je povezivanje redova ne tako jednostavan problem, pa *Python* skripta zadužena za to, u vrijeme pisanja ovog rada, sadrži preko 500 linija programskog koda. Prije nego su podaci georeferencirani jedini ključ za povezivanje su bili redni broj profila[17], koji je bio identičan i kod MP i kod C unosa, i raspored, odnosno dubine horizonata, koje naravno moraju biti identične i kod MP i kod C unosa. Problem je pokušan da se riješi na dva načina:

- **pronalaženjem jedinstvenih profila i u MP i u C bazi sa istim brojem profila i istim rasporedom horizonata**: iako se na prvi pogled može činiti da je ovaj metod najbolji, postojao je veliki broj profila koji imaju isti redni broj i samo jedan horizont[18] iste dubine. Ipak, zanemarivanjem tih duplih profila, bilo je dovoljno samo naći presjek istih unosa u obije baze, što je računski brza operacija.

- **pronalaženjem identičnih nizova više uzastopnih profila**[19]: ovim načinom je pokušano da se profili koji bi po prethodnom metodu bili označeni kao dupli ipak upare, ali ovaj put zato što se nalaze u uzastopnom nizu od N profila[20] koji postoji i u MP i u C bazi. Tj. pretpostavljeno je da ako postoji N identičnih uzastopnih profila, onda je velika vjerovatnoća da su svi ti profili zaista identični u obije baze. Ovaj metod je uspijevao da upari veći broj profila od prethodnog, ali po cijenu mnogo sporijeg izvršavanja[21] koje je iznosilo i više desetina minuta.

Ipak, usljed nedostatka vremena, a ujedno i da bi podaci bili što tačniji radi kvalitetne prostorne prognoze parametara, uparivanje je za potrebe ovog rada vršeno isključivo po osnovu georeference, tj. uparivani su samo oni MP i C profili koji su bili jednoznačno georeferencirani.

Georeferenciranje je vršeno tako što je prvo u *Excel* fajlovima, a na osnovu sadržaja sveski, označavano, na parseru shvatljiv način, odakle počinju profili koji pripadaju nekoj sekciji originalne pedološke karte (slika 3.10). Potom je podatak

[17] Ali koji nije bio jedinstven, kako je ranije navedeno.
[18] Tj. nemaju drugi horizont.
[19] Koji imaju identičan redni broj i identičan broj, odnosno dubinu horizonata.
[20] N je u kodu iznosilo 10.
[21] Nije vršena analiza vremenske kompleksnosti ova dva algoritma.

o sekciji unošen u posebnu kolonu u bazi za svaki profil koji pripada nekoj sekciji. Određeni broj profila u sveskama je imao naznačen i kvadrat u okviru sekcije gdje se nalazi, što je dodatno olakšavalo georeferenciranje. Konačno, kada bi podaci o sekciji, i eventualno kvadratu, bili uneseni u bazu, rađeno je uparivanje profila sa posebnom bazom u kojoj su se nalazile precizne koordinate profila, a čija se struktura može vidjeti na slici 3.11. Međutim, prilikom uparivanja se javio novi problem: postojanje duplih profila[22] u okviru iste sekcije (vidjeti primjer na slici 3.12), tako da je bilo potrebno dodatno unaprijediti proces tako da se izbjegnu dupli profili[23]. Nakon eliminisanja duplih profila, parser je mogao da jednostavno upari koordinate sa odgovarajućim profilima, a nakon toga su se i georeferencirani MP i C podaci mogli vrlo jednostavno upariti.

| **Labor.** | | | | | **GRANULC** |
|---|---|---|---|---|---|
| **broj** | | **Lokacija** | **Profil** | **Dubina** | **Skelet** |
| | | | | | **>2mm** |
| 4064 | 13 | | | 60-90 | 69.7 |
| 4065 | 14 | | 5 | 0-10 | 3.84 |
| 4066 | 15 | | | 30-60 | 70.5 |
| 4067 | 16 | - vocnjak | 1 | 0-18 | 8.26 |
| 4068 | 17 | | | 20-50 | 15.1 |
| 4069 | 18 | | | 70-100 | 0 |
| 4070 | 19 | | | 110-130 | 16.56 |
| | | | | | |
| | | | | | |
| Sekcija Kolašin 1 | | | | | |
| | | | | | |
| **Labor.** | | | | | **GRANULC** |
| **broj** | | **Lokacija** | **Profil** | **Dubina** | **Skelet** |
| | | | | **u cm** | **>2cm** |
| **Str 152** | | | | | |
| | 1 | Mojkovac - Bistrica | 1 | 0-10 | 3.5 |
| | 2 | | | 10,-30 | 35.81 |
| | 3 | | | 30-70 | 63.34 |
| | 4 | | | 80-120 | 72.12 |
| | 5 | | 2 | 0-10 | 47.56 |

Slika 3.10: Označavanje početka sekcije u *Excel* fajlovima

[22]Ukupno je bilo oko 50-tak duplih profila na čitavoj karti.

[23]Ovo je rađeno tako što su koordinate za problematične profile unošene direktno u *Excel* fajlove.

| Profile | Section | x | y |
|---|---|---|---|
| 8G | pec2 | 6690629.16 | 4752875.52 |
| 6M | pec2 | 6686266.53 | 4755738.23 |
| 5G | pec2 | 6687276.45 | 4755915.8 |
| 7M | pec2 | 6686703.05 | 4755619.86 |
| 4G | pec2 | 6688574.91 | 4754942.89 |
| 8M | pec2 | 6685959.49 | 4754121.63 |

Slika 3.11: Tabela preciznih koordinata profila

### 3.3.3 Koraci kod parsiranja

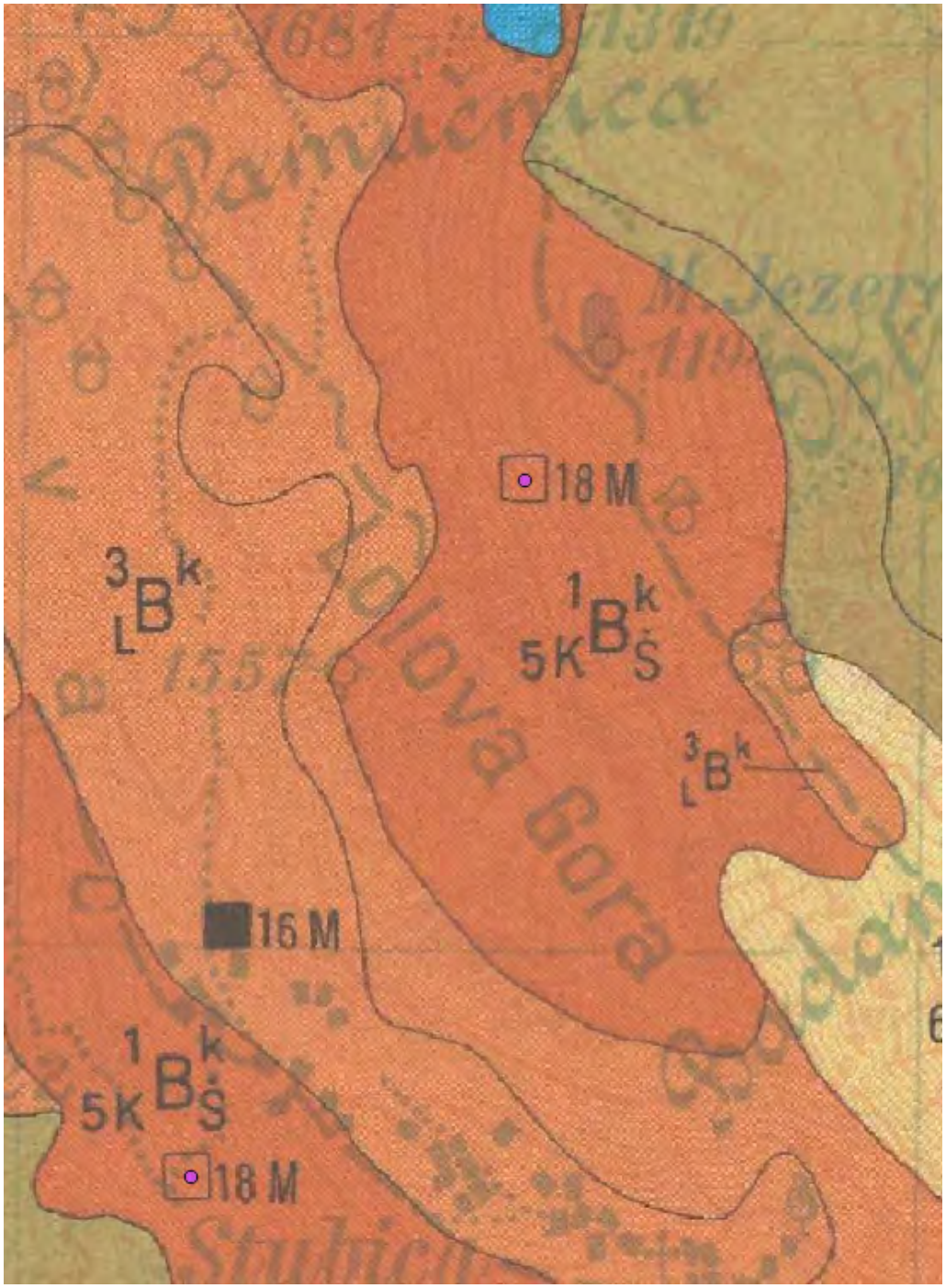
Proces parsiranja se ugrubo može svesti na sljedeće korake:

1. skladištenje svih koordinata[24] u posebnu SQLite bazu i provjera njihove validnosti;
2. generisanje SQLite baze za prve tri MP sveske i prve tri C sveske i provjera validnosti podataka;
3. uparivanje koordinata sa profilima iz ovih 6 svesaka, i snimanje neuparenih koordinata u posebnu bazu;
4. uparivanje profila iz prve tri MP sveske sa profilima iz prve tri C sveske u jedinstvenu MPC bazu[25];
5. prosto uparivanje MP i C parametara za hrvatske podatke;
6. uparivanje hrvatskih MPC podataka sa preostalim koordinatama, i snimanje novo-preostalih neuparenih koordinata u posebnu bazu;
7. prosto uparivanje MP i C parametara za podatke iz knjige (Fuštić i Đuretić 2000);
8. uparivanje MPC podataka knjige (Fuštić i Đuretić 2000) sa preostalim koordinatama[26];

[24]Iz *Excel* fajla sa strukturom prikazanom na slici 3.11.

[25]Ključ za uparivanje su bile koordinate dodijeljene u prethodnom koraku.

[26]Snimanje preostalih koordinata pri svakom koraku uparivanja koordinata je bilo potrebno upravo radi ovog koraka, jer kako je već rečeno, mnogi od profila iz knjige (Fuštić i Đuretić 2000) se nalaze u ostalim sveskama, pa je ovime spriječen njihov ponovni unos u konačnu bazu, što je moglo da negativno utiče na algoritme prostorne prognoze.




Slika 3.12: Primjer duplog profila u okviru iste sekcije

9. objedinjavanje sve tri skupine georeferenciranih MPC podataka u jednu bazu.

Kao što je već rečeno, kod svakog nailaska na grešku tokom bilo kog od gore navednih koraka, parser bi automatski naznačio na kojoj stranici i u kom redu na

| | |
|---|---|
| Ukupan broj profila | 13636 |
| Ukupan broj horizonata | 28778 |
| Ukupan broj georeferenciranih profila | 3921 |
| Ukupan broj georeferenciranih horizonata | 6482 |
| Ukupan broj profila sa MP karakteristikama | 7038 |
| Ukupan broj horizonata sa MP karakteristikama | 15717 |
| Ukupan broj georeferenciranih profila sa MP karakteristikama | 2197 |
| Ukupan broj georeferenciranih horizonata sa MP karakteristikama | 4527 |
| Ukupan broj profila sa C karakteristikama | 6679 |
| Ukupan broj horizonata sa C karakteristikama | 17451 |
| Ukupan broj georeferenciranih profila sa C karakteristikama | 3829 |
| Ukupan broj georeferenciranih horizonata sa C karakteristikama | 6345 |
| Ukupan broj georef. profila i sa MP i sa C karakteristikama | 2110 |
| Ukupan broj georef. horizonata i sa MP i sa C karakteristikama | 4390 |

Tabela 3.1: Sumarni prikaz broja profila i horizonata u finalnoj bazi

stranici, odnosno u kom *Excel* fajlu i redu se nalazi greška, što je u mnogome olakšalo proces ispravljanja podataka. Takođe, svi koraci odgovaraju *Python* skriptama koje je, kako je već rečeno, moguće modularno uklapati.

Tabela 3.1 daje sumarni prikaz profila i horizonata u finalnoj bazi.

# Poglavlje 4

# Generisanje tematske karte fosfora

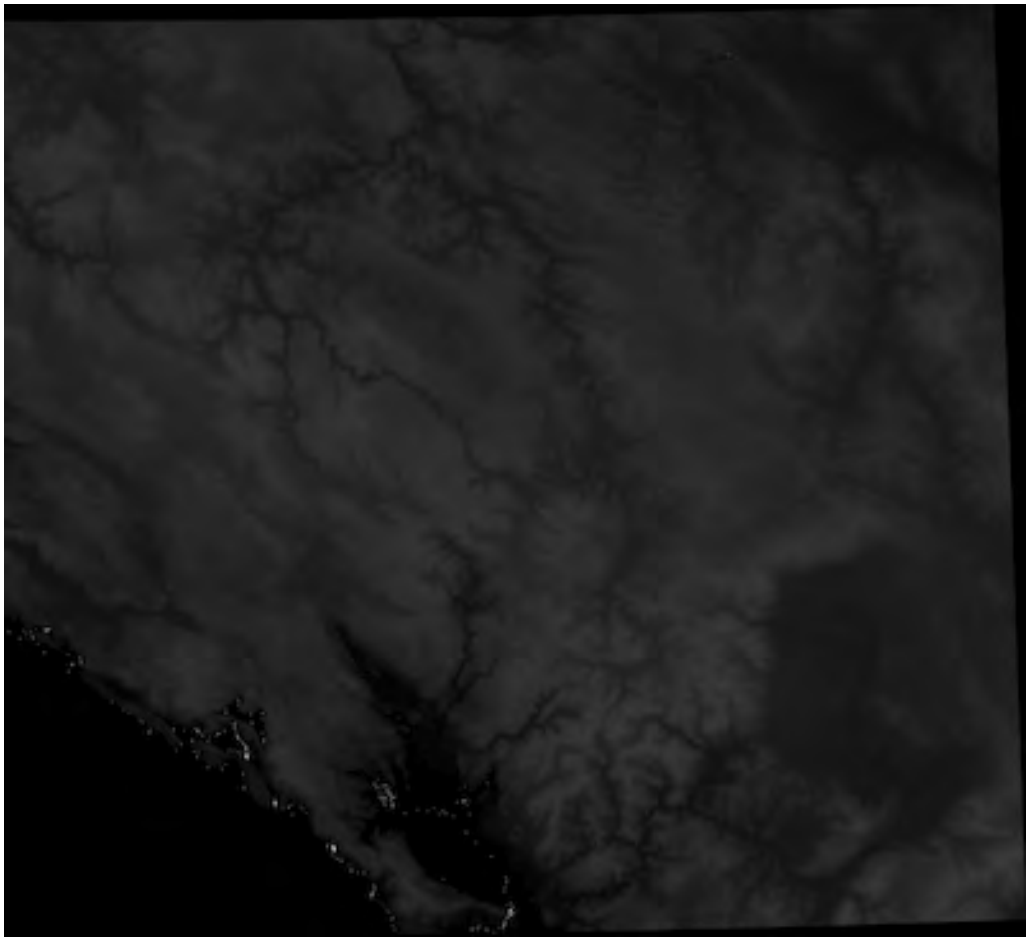
U ovom poglavlju će biti dat prikaz generisanja tematske karte za lako pristupačni fosfor[1] ($P_2O_5$ u $mg/100g$) na osnovu baze opisane u poglavlju 3 pomoću nekoliko interpolacionih algoritama. Odabran je fosfor, jer su podaci stari već nekoliko decenija, a količina fosfora se relativno sporo mijenja u zemljištu koje nije izloženo uticaju čovjeka.

## 4.1 Ulazni parametri

Kao osnova za generisanje karte fosfora korištene su smjernice istraživanja navedene u (Sarmadian i dr. 2014) i (Keshavarzi i dr. 2015), gdje su kao ulazni parametri korišteni topografski podaci (*DEM*) i satelitski snimci za prognozu fosfora.

Topografski podaci (samo elevacija) su preuzeti u HGT formatu iz baze (de Ferranti 2014), sa rezolucijom od 3″ (90 m). HGT fajl je pretvoren u GeoTIFF, iz kojeg je zatim izdvojena samo regija Crne Gore (slika 4.1), a potom su na osnovu tog fajla, pomoću programa SAGA GIS, generisani GeoTIFF fajlovi za *nagib* (*slope*), *aspekt* (*aspect*), i *površinsku zakrivljenost* (*plan_curvature*). Ostali ulazni podaci su satelitski podaci iz skupova podataka *Global Land Survey*, tj. MSS opsezi 1, 2, 3 i 4 za godinu 1975 (USGS 2008a) i TM opsezi 1, 2, …7 za godinu 1990 (USGS 2008b). Odabrani su upravo ovi GLS podaci, jer su među najstarijim satelitskim podacima dostupnim za oblast Crne Gore, tj. najbliži su vremenu u kome su se prikupljali podaci koji se interpoliraju.

[1] U nastavku samo fosfor.



Slika 4.1: Raster nadmorske visine za oblast Crne Gore

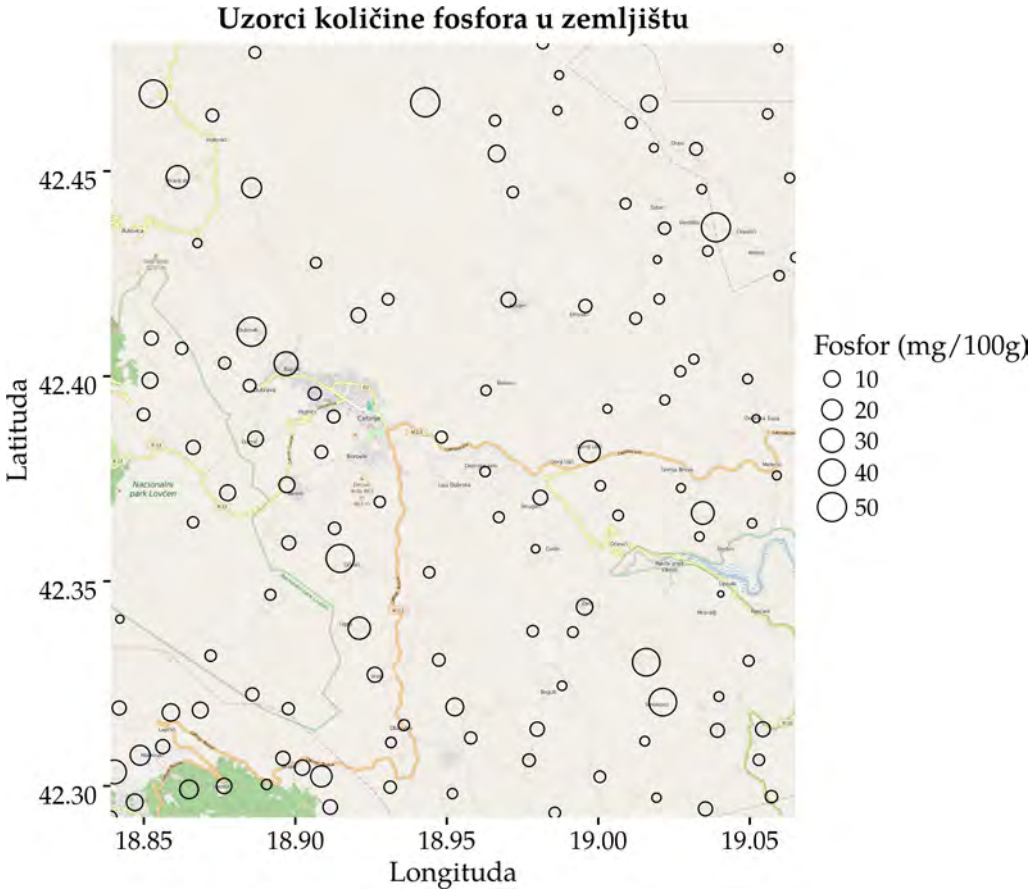
## 4.2 Ogledna površ sa lokacijama uzoraka

Tematske karte su generisane za pravouganu oblast u okviru pedološke karte „Cetinje 1“ (slika 4.2, blizu Cetinja). Odabrana je relativno mala površina radi sljedećeg:

- da bi bila u okviru samo jedne pedološke karte, jer je jako teško precizno „poklopiti“ pedološke karte (posebno više njih od jednom) sa satelitskim snimkom;
- da bi srednja vrijednost parametra bila barem približno stacionarna;
- da bi se algoritmi što brže izvršavali.

A takođe je ova oblast odabrana jer nema puno vodenih površina, što bi dodatno

moglo da utiče na interpolacione algoritme.[2] Nakon „ručnog“ podešavanja skenirane karte sa satelitskom, zabilježen je geografski koordinatni referentni sistem[3] (CRS), a isti CRS je korišten svugdje prilikom obrade podataka[4].




Slika 4.2: Uzorci količine fosfora na oglednoj površi

Koordinate ogledne površi su:

$$
\begin{aligned}
x_{min} &= 6569639 \\
x_{max} &= 6588163 \\
y_{min} &= 4683447 \\
y_{max} &= 4704582
\end{aligned}
$$

[2]Prisustvo vode značajno utiče na difuziju supstanci kroz zemljište.

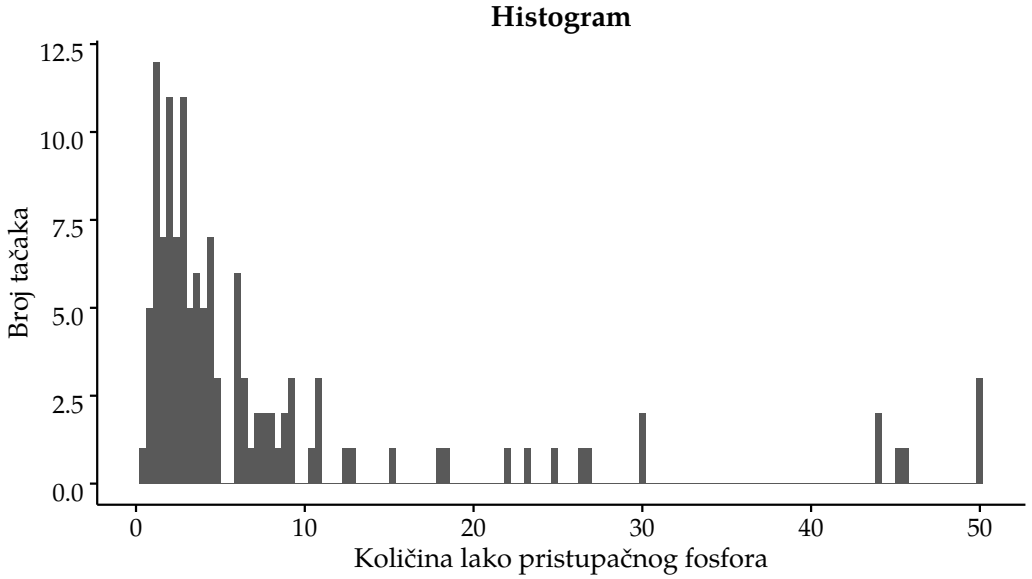
[3]Koji približno odgovara CRS-u *MGI/Balkans zone 6*, odnosno *EPSG:31276*.

[4]Prvobitno su greškom DEM i GLS podaci učitani sa malo drugačijim CRS-om, što je mjerljivo uticalo na rezultate interpolacije.

Na ovoj površi je bilo $N = 125$ georeferenciranih profila za koje je vršena analiza količine lako pristupačnog fosfora.

## 4.3 Priprema podataka

Iz baze podataka su izdvojeni samo površinski horizonti profila[5] koji pripadaju ovoj oblasti. Nakon uvida u histogram (slika 4.3), uzet je logaritam vrijednosti fosfora, jer su originalne vrijednosti uglavnom koncentrisane blizu nule, tj. nemaju normalnu distribuciju, a ona je uslov za dobre rezultate interpolacionih algoritama. Potom je i ta vrijednost skalirana (slika 4.4), a skalirane su i vrijednosti ulaznih parametara, opet da bi se dobili što bolji rezultati prilikom interpolacije.[6]



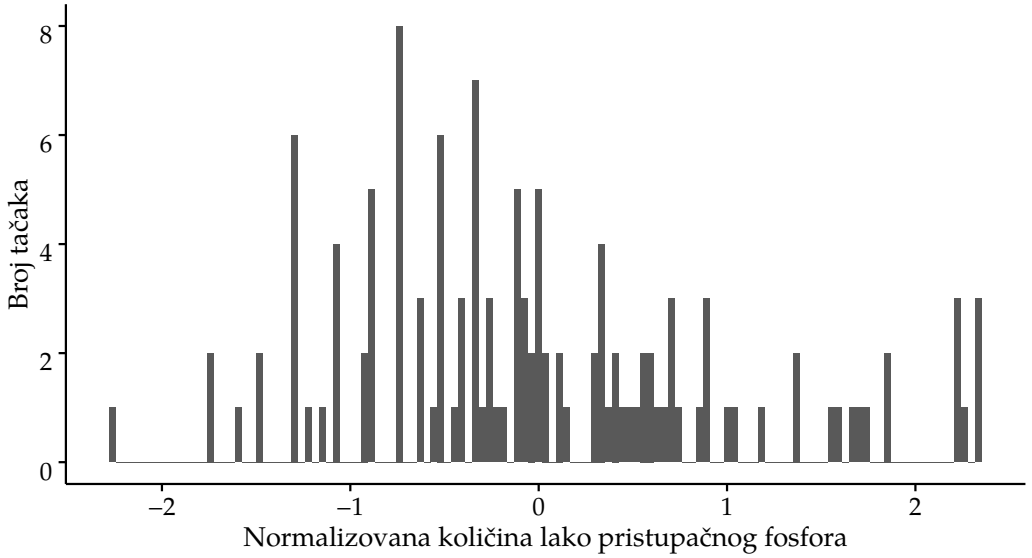

Slika 4.3: Histogram prvobitne količine fosfora

Potom je nad ulaznim parametrima[7] sprovedena analiza glavnih komponenti (*Principal Component Analysis — PCA*) pomoću biblioteke *GSIF*, a za regresionu analizu je odabrano onoliko glavnih komponenti koliko „objašnjava“ 80% varijacije. I ovdje je bilo moguće dobiti bolje rezultate regresije boljim odabirom kompo-

[5] Debljine od 0 do $x$ cm.

[6] Ovakvom pripremom podataka su slijeđene preporuke iz (Hengl 2009).

[7] Tj. nad ambijentalnim satelitski dobijenim podacima. Prilikom PCA analize je korišten kompletan raster svakog od ulaznih parametara za oglednu površ.




Slika 4.4: Histogram logaritmovane, a zatim skalirane količine fosfora

nenti, a takođe je moguće implementirati kod koji na osnovu svih tačaka automatski bira optimalne ulazne komponente za dati prostorni parametar, ali bi onda rezultati validacije bili pristrasni[8].

## 4.4 Rezultati

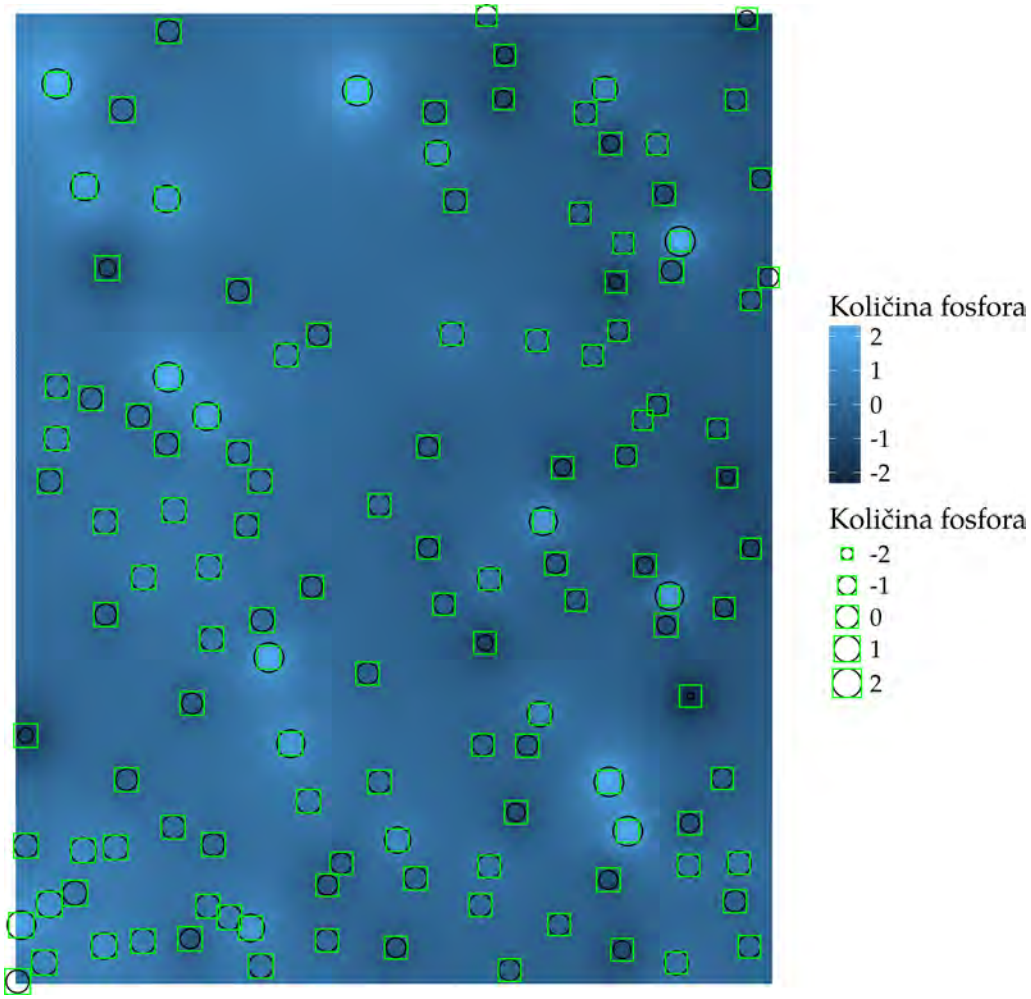
Interpolacioni algoritmi su primijenjeni u programskom jeziku *R*, pomoću koda navedenog u dodatku A. Na prikazanim kartama uzorkovane vrijednosti za validacione tačke su prikazane kružićima, a prognozirane kvadratićima, pri čemu njihove veličine odgovaraju (skaliranim) brojčanim vrijednostima, pa je moguće vizuelno uporediti tačnost prognoze. Korištenjem svih dostupnih uzoraka, dobijene su karte za sljedeće algoritme:

- model inverzne udaljenosti (slika 4.5), pomoću koda:

```
krige(p2o5~1, mydata.train, mydata.grid)
```

- obični kriging (slika 4.6), pomoću koda:

[8]Kod sa takvom funkcionalnošću i postoji u dodatku A, a implementiran je u prokomentarisanim linijama koda pomoću fukcije *step*, ali je često prilikom validacije davao gore rezultate negoli jednostavni kod.




Slika 4.5: Model inverzne udaljenosti; ○: uzorkovane vrijednosti, □: prognozirane vrijednosti

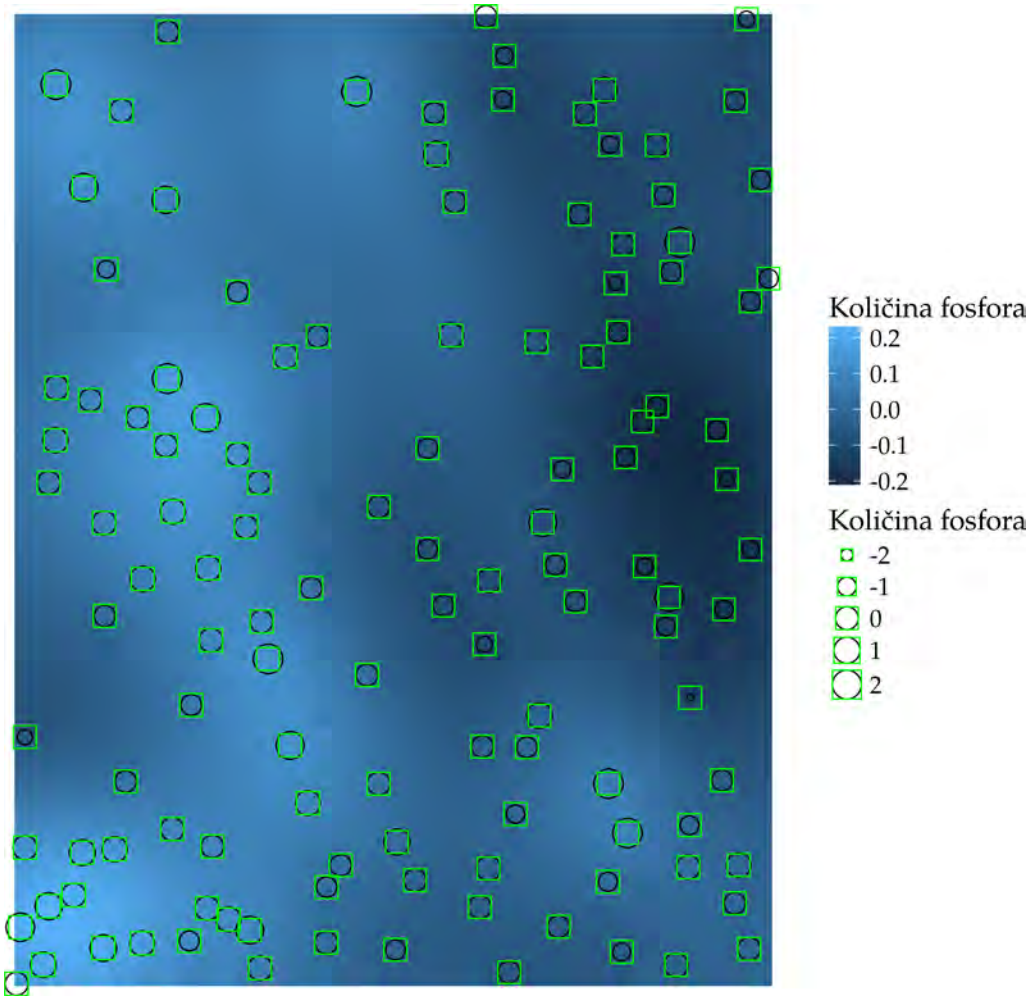
```
krige(p2o5~1, mydata.train, mydata.grid, model=p2o5.ok.vgm.fit)
```

- linearna regresija (slika 4.7), pomoću koda:

```
krige(myformula, mydata.train, mydata.grid)
```

- univerzalni kriging (slika 4.8), pomoću koda:

```
krige(myformula, mydata.train, mydata.grid,model=p2o5.uk.vgm.fit)
```

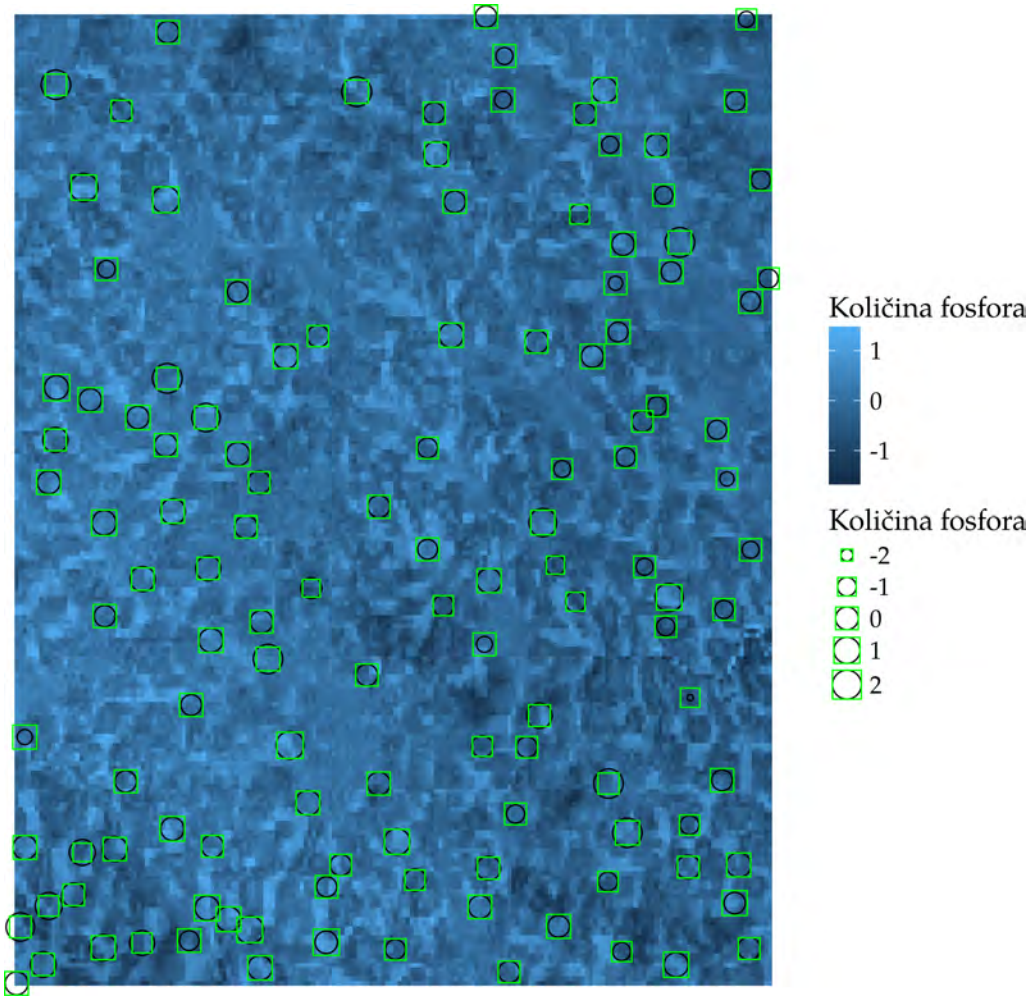



Slika 4.6: Obični kriging; ○: uzorkovane vrijednosti, □: prognozirane vrijednosti

Bitno je napomenuti da je kod običnog kriginga korišten variogram dat na slici 4.9, a kod univerzalnog kriginga variogram dat na slici 4.10.

Pri samom procesu konstrukcije modela variograma na osnovu eksperimentalnog variograma, korištena je R biblioteka automap (Hiemstra i dr. 2008), koja automatski konstruiše model variograma na osnovu ulaznih podataka. Ručnim podešavanjem su dobijani bolji rezultati, ali je ovako kod jednostavniji, i moguće ga je lakše prilagoditi novim podacima, pa čak i novom prostornom parametru.

Što se tiče validacije i međusobnog poređenja algoritama, korištene su funkcije iz *R* biblioteke gstat (Pebesma 2004). Pošto je ukupan broj tačaka relativno mali, jako teško je dati pouzdano poređenje algoritama. Kvantitativni parametri koji su korišteni za iskazivanje kvaliteta prognoze su sljedeći:

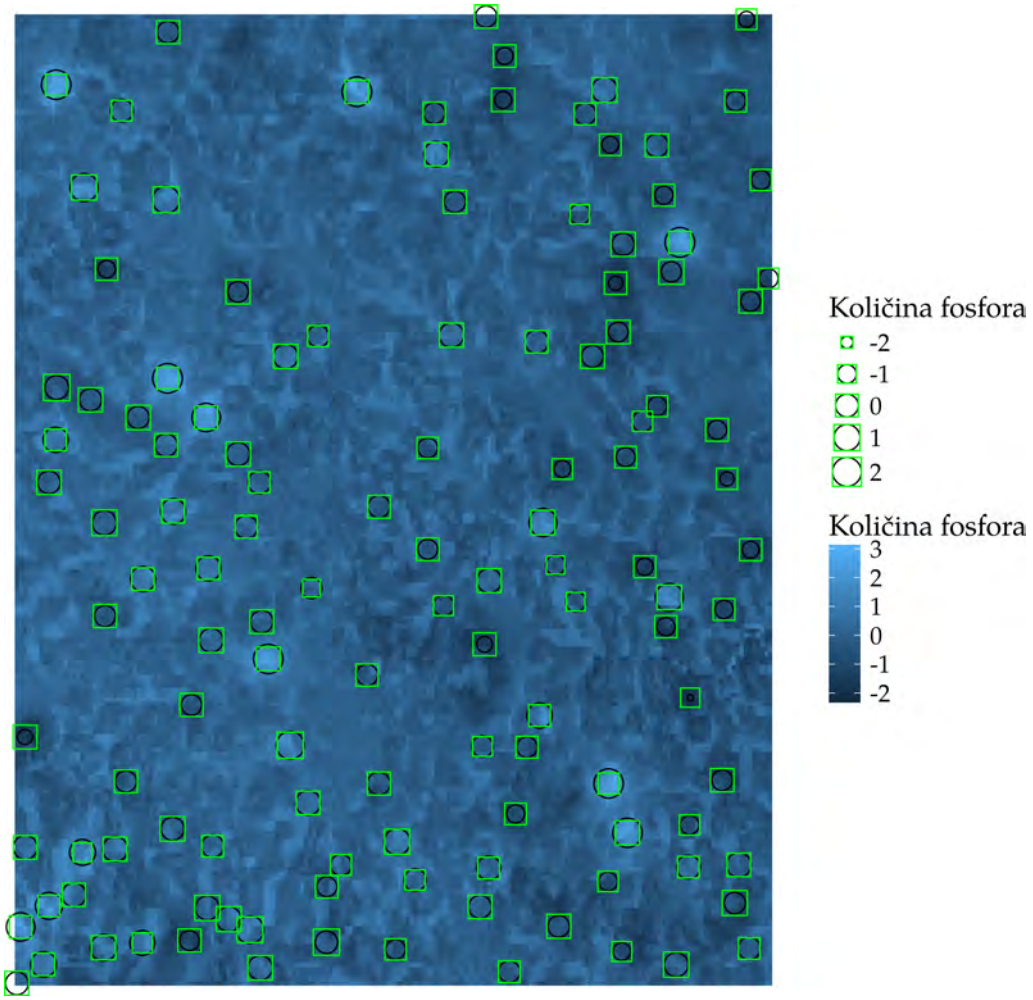



Slika 4.7: Linearna regresija; ○: uzorkovane vrijednosti, □: prognozirane vrijednosti

- $RMSE$: korijen srednje kvadratne greške (razlike između prognozirane i uzorkovane vrijednosti), idealno 0;
- $Cor$: korelacija između prognozirane i uzorkovane vrijednosti, idealno 1;
- $R^2$: koeficijent determinacije, idealno što bliži 1.

Takođe je, kao reper ostalim algoritmima, korišten pseudo-slučajni „prognozer“ (označen u tabelama sa RND), kod koga sve „prognoze“ pripadaju pseudo-slučajnom nizu sa (normalnom) distribucijom koju ima originalni niz izmjerenih podataka.[9]

[9]Prilikom zvanične prijave teme ovog rada postavljena je inicijalna hipoteza da će se korištenjem

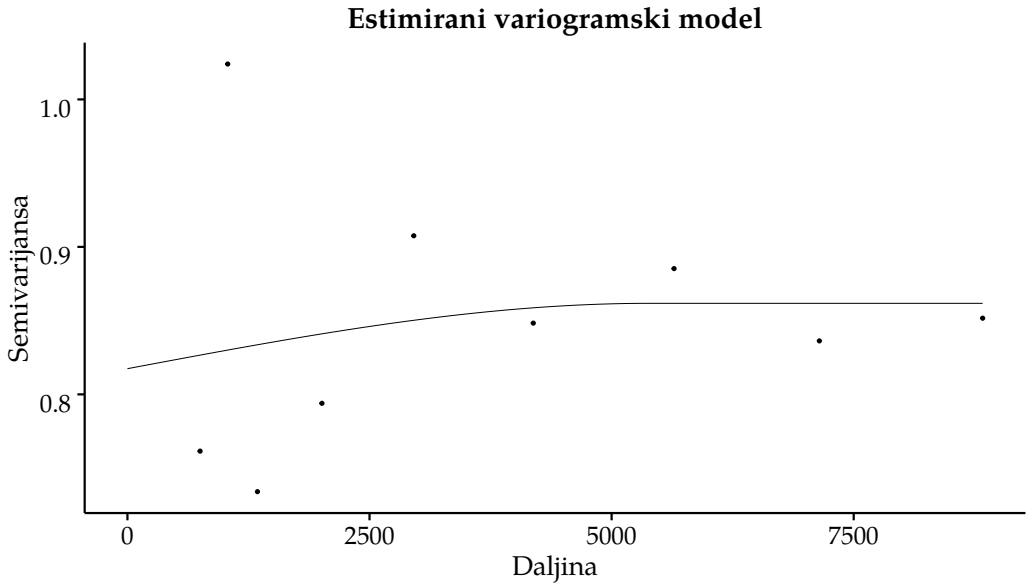



Slika 4.8: Univerzalni kriging; ○: uzorkovane vrijednosti, □: prognozirane vrijednosti

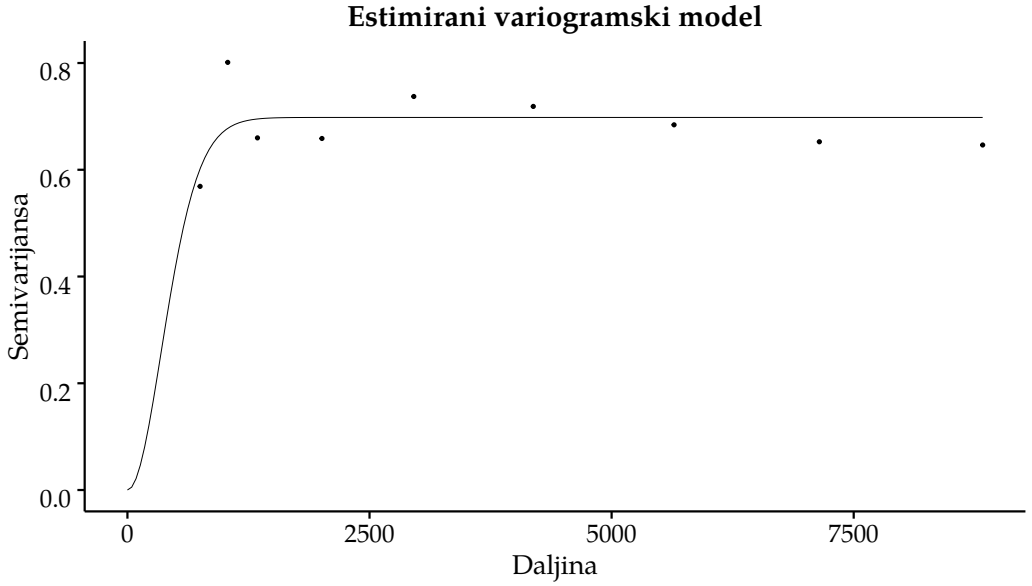
Nakon isprobavanja više načina validacije, ovdje će biti prikazani validacioni rezultati dobijeni na sljedeće načine:

- korištenjem svih tačaka za konstrukciju modela i *k*-segmentne validacije sa 10 segmenata nad svim tačkama (tabela 4.1);
- korištenjem svih tačaka za konstrukciju modela i *leave-one-out* validacije nad svim tačkama (tabela 4.2);
- korištenjem jednog podskupa tačaka za konstrukciju modela i *k*-segmentne validacije sa 10 segmenata nad preostalim podskupom tačaka (tabela 4.3);

interpolacionih algoritama dobiti karta bolja od slučajno generisane.




Slika 4.9: Variogram korišten prilikom običnog kriginga




Slika 4.10: Variogram korišten prilikom univerzalnog kriginga

- korištenjem jednog podskupa tačaka za konstrukciju modela i *leave-one-out* validacije nad preostalim podskupom tačaka (tabela 4.4);
- korištenjem svih tačaka, uz uklonjene outlier-e, za konstrukciju modela i *leave-one-out* validacije nad svim tačkama (tabela 4.5).

| | IDW | OK | LR | UK | RND |
|---|---|---|---|---|---|
| RMSE | 0.9720 | 0.9886 | 0.9549 | 0.9403 | 1.4450 |
| Cor | 0.2598 | 0.1294 | 0.3273 | 0.3537 | -0.0526 |
| $R^2$ | 0.0491 | 0.0149 | 0.0807 | 0.1086 | -1.1051 |

Tabela 4.1: Parametri pouzdanosti interpolacije za 10-to segmentnu validaciju nad svim tačkama

| | IDW | OK | LR | UK | RND |
|---|---|---|---|---|---|
| RMSE | 0.9720 | 0.9850 | 0.9434 | 0.9338 | 1.4450 |
| Cor | 0.2535 | 0.1680 | 0.3446 | 0.3657 | -0.0526 |
| $R^2$ | 0.0490 | 0.0219 | 0.1028 | 0.1210 | -1.1051 |

Tabela 4.2: Parametri pouzdanosti interpolacije za *LOO* validaciju nad svim tačkama

| | IDW | OK | LR | UK | RND |
|---|---|---|---|---|---|
| RMSE | 0.9675 | 0.9173 | 0.8882 | 0.8277 | 1.4450 |
| Cor | -0.0049 | 0.0886 | 0.3273 | 0.4255 | -0.1290 |
| $R^2$ | -0.1765 | -0.0613 | 0.0051 | 0.1368 | -1.6198 |

Tabela 4.3: Parametri pouzdanosti interpolacije za 10-to segmentnu validaciju nad podskupom svih tačaka

Bitno je pomenuti, da je kod podjele na trening i validacione podskupove korišten odnos 45:55, tj. 45% tačaka je korišteno za treniranje, tj. određivanje variograma za obični i univerzalni kriging, a 55% tačaka je korišteno prilikom *LOO*, odnosno 10-to segmentne validacije. Pokušani su i neki drugi odnosi, ali, pošto je ukupan broj tačaka mali, ovaj odnos daje koliko-toliko smislene rezultate. Koliko je, sa aspekta validacije, mali ukupan broj tačaka govori i podatak da je čak i izbor „sjemena“

| | IDW | OK | LR | UK | RND |
|---|---|---|---|---|---|
| RMSE | 0.8796 | 0.8958 | 0.8214 | 0.7732 | 1.4750 |
| Cor | 0.2549 | 0.1621 | 0.4508 | 0.5307 | -0.1031 |
| $R^2$ | 0.0547 | 0.0198 | 0.1757 | 0.2697 | -1.6539 |

Tabela 4.4: Parametri pouzdanosti interpolacije za *LOO* validaciju nad podskupom svih tačaka

| | IDW | OK | LR | RK | RND |
|---|---|---|---|---|---|
| RMSE | 0.9519 | 0.9009 | 0.8870 | 0.8744 | 1.3140 |
| Cor | 0.3247 | 0.4258 | 0.4701 | 0.4890 | 0.1288 |
| $R^2$ | 0.0894 | 0.1813 | 0.2064 | 0.2287 | -0.7425 |

Tabela 4.5: Parametri pouzdanosti interpolacije za *LOO* validaciju nad svim tačkama, isključujući outlier-e

(*seed*) korištenog kod pseudo-slučajne podjele na trening i validacione skupove značajno uticao na rezultate validacije. Stoga rezultate prikazane za slučaj podjele na trening i validacioni skup treba uzeti sa rezervom.

Radi ilustracije, u tabeli 4.6 su navedene konkretne numeričke skalirane vrijednosti prognoze za tri tačke:

- za tačku sa najvećom greškom (MAX),
- za tačku sa središnjom (median) greškom (MED),
- za tačku sa najmanjom greškom (MIN).

Ono što se može jednostavno zaključiti je da svi savremeni algoritmi generisanja karata daju bolje karte od pseudo-slučajne interpolacije.

Dalje se da primijetiti da je IDW redovno bolji od OK. Ovo je usljed problematičnog automatski dobijenog modela OK variograma, na koji je posebno uticao jedan *outlier* eksperimentalnog variograma. Dokaz toga je i to što je OK bolji od IDW, i to drastično, za slučaj kada su *outlier*-i uklonjeni.

| | MAX | MED | MIN |
|---|---|---|---|
| Uzorak | 2.2134 | -0.6412 | -0.2554 |
| Prognoza | -0.4113 | -0.0839 | -0.2881 |
| Greška | 2.6247 | 0.5573 | 0.0326 |

Tabela 4.6: Nekoliko konkretnih numeričkih primjera greške u prognozi

Kompleksniji algoritmi LR i (najkompleksniji) UK su redovno bolji od jednostavnijih algoritama, a UK je malo bolji od LR. LR i UK zato i daju karte sa najviše detalja.

Konačno, na osnovu parametra $R^2$ za koji se u (Hengl 2009, str. 133) navodi da treba biti oko 0.80, može se zaključiti da ni najbolji model, uz najbolje pripremljene podatke, ne daje rezultate ni blizu toga. To ukazuje da je gustina uzoraka mala, kao i da su ulazni parametri nedovoljno korelisani sa izlaznim parametrom, a moguće je i da podaci nijesu dovoljno prečišćeni.

U zaključku je dat rezime rada, i moguće dalje smjernice za istraživanje na ovom polju.

# Poglavlje 5

# Zaključak

Nakon uvodnih poglavlja, u centralnim poglavljima je prikazana digitalizacija pedoloških podataka, koja predstavlja modernu sistematizaciju kvantitativnih saznanja o zemljištima Crne Gore, i primjena savremenih interpolacionih algoritama u svrhe generisanja tematskih pedoloških karata, konkretno tematske karte lako pristupačnog fosfora.

Što se tiče digitalizacije, iako je zasigurno moguće još dorađivati podatke i povećati im tačnost, čini se da je glavni dio ovog procesa završen, i da se postojeća baza već može koristiti kao pouzdan izvor informacija o zemljištu Crne Gore.

Iako generisane karte zaista povećavaju količinu informacija o lako pristupačnom fosforu za oglednu površ, u odnosu na pseudo-slučajni interpolator, upitna je, sa aspekta pedologije, pouzdanost i korisnost generisanih karata. Ono što se može reći je da je gustina uzorkovanja originalnih podataka nedovoljna, kao i da izabrani ulazni parametri nisu dovoljni da pouzdano „objasne“ varijaciju izlaznog parametra.

Ono u čemu bi interpolacioni algoritmi, posebno regresija, sigurno mogli da pomognu je još preciznije „podešavanje“ CRS-a originalne skenirane karte, tj. pojedinačnih listova, tako što bi se ključni parametri CRS-a iterativno mijenjali dok se ne dobije najpreciziniji mogući CRS za postojeće tačke, odnosno za svaku od pojedinačnih karata. Ovo bi, povratno, povećalo tačnost digitalne baze podataka, a zatim i generisanih karata, jer bi koordinate profila bile još pouzdanije određene.

Takođe, dubljom geostatističkom analizom bi se mogli utvrditi posebno problematični profili, i onda eventualno eliminisati iz baze, ili biti ponovo laboratorijski ispitani. Ovakva analiza bi pomogla i u preporučivanju dovoljne preciznosti

uzorkovanja za eventualna buduća terenska istraživanja zemljišta.

Zatim, treba još istražiti javno dostupne skupove podataka koji mogu poslužiti kao ulazni parametri, kao što su podaci o zemljišnom pokrivaču (*land cover — landcov*), temperaturi zemlje, i slično. Valja napomenuti da, iako su ulazni parametri korišteni u radu korelisani sa izlaznim parametrom, ipak su na kraju konfigurisani automatizovano, i ekspert-pedolog bi sigurno mogao još bolje da preporuči odabir i kombinaciju parametara koji bi bili optimalni za generisanje pedoloških karata, i to ne samo za jedan, nego za razne izlazne pedološke parametre koji su već digitalizovani.

To dalje vodi do mogućnosti generisanja karata za više digitalizovanih pedoloških parametara, što bi u konačnici moglo da se iskoristi za automatizovano generisanje tipske karte zemljišta za oblast, ili čak čitavu Crnu Goru, tako što bi se nekoliko pedoloških parametara odabralo da budu ključni, i onda bi nekim klasifikacionim algoritmom bili iskorišteni za izradu tipske karte.

Takođe, postoji i poligonska karta tipova zemljišta koja je dobijena iz rasterske skenirane pedološke karte, tj. predstavlja digitalnu formu originalne ekspertske pedološke tipske karte, i ona bi se sigurno mogla iskoristiti za još bolju interpolaciju, recimo sugerisanjem (pod)oblasti interpolacije koja bi obuhvatila samo jedan tip zemljišta, odnosno profile koji pripadaju samo jednom tipu zemljišta.

Što se tiče samih interpolacionih algoritama, korisno bi bilo pokušati generisanje karata i pomoću neke varijante regresija-kriging algoritma gdje bi model determinističke komponente bio baziran na neuronskim mrežama i sl.

# Dodaci

# Dodatak A

# Programski kod

Ovdje je naveden kod koji je korišten za generisanje i validaciju interpolacionih karata lako pristupačnog fosfora za oglednu oblast blizu Cetinja. Pošto je kod generisan, sadržavao je dosta ponovljenih djelova, pa su izostavljeni neki od djelova koji se ponavljaju.

```
library(gstat)
library(sp)
library(rgdal)
library(spatstat)
library(DAAG)
library(automap)
library(GSIF)
library(ggplot2)
library(xtable)
library(ggmap)

source('./extrafont_my.r')

## set the seed to make the experiment reproductible
set.seed(200)

mycrs <- paste("+proj=tmerc +lat_0=0 +lon_0=18 +k=0.9999 +x
    ↪ _0=6500000 +y_0=0
```

```
+ellps=bessel␣+towgs84=750,-230,580,0,0,0,0␣+units=m␣+no_
    ↪ defs")
mycrs <- gsub(pattern='\\s',replacement="␣",x=mycrs)
DIR = "/media/Evo850-1TB-1/edin/afh/me_soil_tables/"

get_cutoff <- function(myarray) {
  return(abs(median(myarray)) + 2*sd(myarray))
}

remove_outliers_from_original <- function(mydata, mymax,
    ↪ mymin) {
  mydata = mydata[mydata$p2o5<mymax,]
  mydata = mydata[mydata$p2o5>mymin,]
  return(mydata)
}

remove_outliers_from_prediction <- function(mydata, mymax,
    ↪ mymin) {
  # mydata = mydata[mydata$var1.pred<mymax,]
  # mydata = mydata[mydata$var1.pred>mymin,]
  mydata$var1.pred[mydata$var1.pred>mymax] = mymax
  mydata$var1.pred[mydata$var1.pred<mymin] = mymin
  return(mydata)
}

myrnorm = function(list_) { mean(list_)+sd(list_)*scale(
    ↪ rnorm(length(list_))) }

insertRow <- function(existingDF, newrow, r) {
  existingDF[seq(r+1,nrow(existingDF)+1),] <- existingDF[
    ↪ seq(r,nrow(existingDF)),]
  existingDF[r,] <- newrow
  existingDF
}
```

```
# Nonscaled data
mydata.nonscaled <- read.table(paste(DIR, "p2o5_xy.data.csv
    ↪ ", sep=""), header=TRUE, sep=",")
#mydata.nonscaled = remove.duplicates(mydata.nonscaled,
    ↪ zero = 0.0, remove.second = TRUE, memcmp = TRUE)
cutoff = get_cutoff(mydata.nonscaled$p2o5)
coordinates(mydata.nonscaled) = ~x+y
proj4string(mydata.nonscaled) <- CRS(mycrs)
qplot(x=mydata.nonscaled$p2o5, geom="histogram", bins=
    ↪ length(mydata.nonscaled$p2o5))+xlab("Kolicina␣lako␣
    ↪ pristupacnog␣fosfora")+ylab("Broj␣tacaka") + ggtitle
    ↪ (sprintf("Histogram")) + mytheme
#qplot(x=mydata.nonscaled$p2o5, geom="histogram", bins=
    ↪ length(mydata.nonscaled$p2o5))+xlab("Kolicina lako
    ↪ pristupacnog fosfora")+ylab("Broj tacaka") + ggtitle(
    ↪ sprintf("Histogram")) + mytheme
save_plot_pdf("p2o5.histogram.non-filtered.pdf")

mydata.grid.nonscaled <- read.table(paste(DIR, "p2o5_xy.
    ↪ grid.csv", sep=""), header=TRUE, sep=",")
coordinates(mydata.grid.nonscaled) <- ~ x+y
proj4string(mydata.grid.nonscaled) <- CRS(mycrs)
gridded(mydata.grid.nonscaled) <- TRUE

# Scaled data
mydata <- read.table(paste(DIR, "p2o5_xy.data.csv", sep="")
    ↪ , header=TRUE, sep=",")
# mydata = remove.duplicates(mydata, zero = 0.0, remove.
    ↪ second = TRUE, memcmp = TRUE)
coordinates(mydata) = ~x+y
proj4string(mydata) <- CRS(mycrs)

mydata2 = spTransform(mydata, CRS("+init=epsg:4326"))
```

```
# Needed when saving map
# mydata_map = get_map(mydata2@bbox, maptype="roadmap",
    ↪ source="osm", zoom = 13)
# save(mydata_map, file="mydata_map.rda")
load("mydata_map.rda")
ggmap(mydata_map) + geom_point(data=data.frame(mydata2),
    ↪ aes(x=x,y=y,size=p2o5), colour="black", shape=1) +
    ↪ labs(x="Longituda", y="Latituda", size="Fosfor␣(mg/
    ↪ 100g)") + ggtitle(sprintf("Uzorci␣kolicine␣fosfora␣u␣
    ↪ zemljistu")) + mytheme_map
save_plot_png("p2o5-bubble.png")

mydata.grid <- read.table(paste(DIR, "p2o5_xy.grid.csv",
    ↪ sep=""), header=TRUE, sep=",")
coordinates(mydata.grid) <- ~ x+y
proj4string(mydata.grid) <- CRS(mycrs)
gridded(mydata.grid) <- TRUE

# Scaling
# mydata = remove_outliers_from_original(mydata, cutoff, -
    ↪ cutoff)
mydata$p2o5 = log(mydata$p2o5)
mydata$p2o5 = scale(mydata$p2o5)
new_max = max(mydata$p2o5)
new_min = min(mydata$p2o5)
qplot(x=mydata$p2o5, geom="histogram", bins=length(mydata.
    ↪ nonscaled$p2o5))+xlab("Normalizovana␣kolicina␣lako␣
    ↪ pristupacnog␣fosfora")+ylab("Broj␣tacaka") + mytheme
save_plot_pdf("p2o5.histogram.filtered.pdf")

mydata$DEM_elevation = scale(mydata$DEM_elevation)
mydata$DEM_slope = scale(mydata$DEM_slope)
mydata$DEM_aspect = scale(mydata$DEM_aspect)
mydata$DEM_plan_curvature = scale(mydata$DEM_plan_curvature
```

```
    ↪ )
mydata$GLS1975_band1 = scale(mydata$GLS1975_band1)
mydata$GLS1975_band2 = scale(mydata$GLS1975_band2)
mydata$GLS1975_band3 = scale(mydata$GLS1975_band3)
mydata$GLS1975_band4 = scale(mydata$GLS1975_band4)
mydata$GLS1990_band1 = scale(mydata$GLS1990_band1)
mydata$GLS1990_band2 = scale(mydata$GLS1990_band2)
mydata$GLS1990_band3 = scale(mydata$GLS1990_band3)
mydata$GLS1990_band4 = scale(mydata$GLS1990_band4)
mydata$GLS1990_band5 = scale(mydata$GLS1990_band5)
mydata$GLS1990_band6 = scale(mydata$GLS1990_band6)
mydata$GLS1990_band7 = scale(mydata$GLS1990_band7)

mydata.grid$DEM_elevation = scale(mydata.grid$DEM_elevation
    ↪ )
mydata.grid$DEM_slope = scale(mydata.grid$DEM_slope)
mydata.grid$DEM_aspect = scale(mydata.grid$DEM_aspect)
mydata.grid$DEM_plan_curvature = scale(mydata.grid$DEM_plan
    ↪ _curvature)
mydata.grid$GLS1975_band1 = scale(mydata.grid$GLS1975_band1
    ↪ )
mydata.grid$GLS1975_band2 = scale(mydata.grid$GLS1975_band2
    ↪ )
mydata.grid$GLS1975_band3 = scale(mydata.grid$GLS1975_band3
    ↪ )
mydata.grid$GLS1975_band4 = scale(mydata.grid$GLS1975_band4
    ↪ )
mydata.grid$GLS1990_band1 = scale(mydata.grid$GLS1990_band1
    ↪ )
mydata.grid$GLS1990_band2 = scale(mydata.grid$GLS1990_band2
    ↪ )
mydata.grid$GLS1990_band3 = scale(mydata.grid$GLS1990_band3
    ↪ )
mydata.grid$GLS1990_band4 = scale(mydata.grid$GLS1990_band4
```

```
    ↪ )
mydata.grid$GLS1990_band5 = scale(mydata.grid$GLS1990_band5
    ↪ )
mydata.grid$GLS1990_band6 = scale(mydata.grid$GLS1990_band6
    ↪ )
mydata.grid$GLS1990_band7 = scale(mydata.grid$GLS1990_band7
    ↪ )


# Calculate important spatial principal components
auxiliary = ~DEM_elevation+DEM_slope+DEM_aspect+DEM_plan_
    ↪ curvature+GLS1975_band1+GLS1975_band2+GLS1975_band3+
    ↪ GLS1975_band4+GLS1990_band1+GLS1990_band2+GLS1990_
    ↪ band3+GLS1990_band4+GLS1990_band5+GLS1990_band6+
    ↪ GLS1990_band7
myspc <- spc(mydata.grid, auxiliary)
mydata.grid.spc = myspc@predicted
mydata.grid@data <- cbind(mydata.grid.spc@data, mydata.
    ↪ grid@data)
mydata.spc = over(mydata, mydata.grid)
mydata@data <- cbind(mydata.spc, mydata@data)

for(i in 1:length(myspc@pca$sdev))
{
  if (sum(myspc@pca$sdev[1:i])/sum(myspc@pca$sdev) > 0.80){
    PCA.NUM = i
    break
  }
}
# PCA.NUM = length(myspc@pca$sdev)
auxiliary = as.formula(paste("~", paste(names(mydata.grid.
    ↪ spc)[1:PCA.NUM], collapse = "+")))
myformula = update(auxiliary, p2o5 ~ .)
```

```
#tmp <- bubble(mydata, "p2o5", col=c("#00ff0088", "#00
    ↪ ff0088"), main = "p2o5")
#png(filename="p2o5.bubble.png")
#print(tmp)
#dev.off()

# Random "interpolated" data. Used to asses real
    ↪ interpolation algorithms
mydata.rnd = mydata
mydata.rnd$p2o5 = myrnorm(mydata$p2o5)

# List for storing cross validation data
CV.list = list()

# Split into training and CV sets
train.percent = 100/100
train_size <- floor(train.percent * nrow(mydata))

if (0 <= train.percent && train.percent < 1) {
    train_ind <- sample(seq_len(nrow(mydata)), size = train_
        ↪ size)

    mydata.train = mydata[train_ind, ]
    mydata.cv = mydata[-train_ind, ]
    mydata.rnd.train = mydata.rnd[train_ind, ]
    mydata.rnd.cv = mydata.rnd[-train_ind, ]
} else if (train.percent == 1) {
    mydata.train = mydata
    mydata.cv = mydata
    mydata.rnd.train = mydata.rnd
    mydata.rnd.cv = mydata.rnd
} else {
    stop("The train_size should be in the (0, 1] range ")
}
```

```
# Calculate important auxiliary variables with linear
   ↪ regression
p2o5.lr = lm(myformula, mydata.train)
p2o5.lrs <- step(p2o5.lr)
# summary(p2o5.lrs)
# myformula = p2o5.lrs$call$formula

# Ordinary kriging variogram
p2o5.ok.autofitVariogram = autofitVariogram(p2o5~1, mydata.
   ↪ train)
p2o5.ok.vgm = p2o5.ok.autofitVariogram$exp_var
p2o5.ok.vgm.fit = p2o5.ok.autofitVariogram$var_model
#p2o5.ok.vgm = variogram(p2o5~1, mydata.train)
#model=vgm(0, "Exp", 3000)
#p2o5.ok.vgm.fit = fit.variogram(p2o5.ok.vgm, model=model)
ggplot(rbind(cbind(variogramLine(p2o5.ok.vgm.fit, maxdist =
   ↪  max(p2o5.ok.vgm$dist)))), aes(x = dist, y = gamma))
   ↪ + geom_line(size=I(0.2)) + geom_point(data=p2o5.ok.
   ↪ vgm, aes(x = dist, y = gamma), size=I(0.5)) + labs(x=
   ↪ "Daljina", y="Semivarijansa") + ggtitle("Estimirani␣
   ↪ variogramski␣model") + mytheme
save_plot_pdf("p2o5.ok.variogram.pdf")

# Universal kriging variogram
p2o5.uk.autofitVariogram = autofitVariogram(myformula,
   ↪ mydata.train)
p2o5.uk.vgm = p2o5.uk.autofitVariogram$exp_var
p2o5.uk.vgm.fit = p2o5.uk.autofitVariogram$var_model
#p2o5.uk.vgm = variogram(myformula, mydata.train)
#model=vgm(0, "Exp", 3000)
#p2o5.uk.vgm.fit = fit.variogram(p2o5.uk.vgm, model=model)
ggplot(rbind(cbind(variogramLine(p2o5.uk.vgm.fit, maxdist =
   ↪  max(p2o5.uk.vgm$dist)))), aes(x = dist, y = gamma))
```

```
    ↪ + geom_line(size=I(0.2)) + geom_point(data=p2o5.uk.
    ↪ vgm, aes(x = dist, y = gamma), size=I(0.5)) + labs(x=
    ↪ "Daljina", y="Semivarijansa") + ggtitle("Estimirani␣
    ↪ variogramski␣model") + mytheme
save_plot_pdf("p2o5.uk.variogram.pdf")


p2o5.idw = krige(p2o5~1, mydata.train, mydata.grid)
#p2o5.idw = remove_outliers_from_prediction(p2o5.idw, new_
    ↪ max, new_min)
p2o5.idw.cv = krige.cv(p2o5~1, mydata.cv)

error.abs = abs(p2o5.idw.cv$var1.pred - p2o5.idw.cv$
    ↪ observed)
mymax = max(error.abs)
myindex = which(error.abs==mymax)
MAX = c(p2o5.idw.cv$observed[myindex], p2o5.idw.cv$var1.
    ↪ pred[myindex], mymax)
mymed = median(error.abs)
myindex = which(error.abs==mymed)
MED = c(p2o5.idw.cv$observed[myindex], p2o5.idw.cv$var1.
    ↪ pred[myindex], mymed)
mymin = min(error.abs)
myindex = which(error.abs==mymin)
MIN = c(p2o5.idw.cv$observed[myindex], p2o5.idw.cv$var1.
    ↪ pred[myindex], mymin)
CV.example = data.frame(MAX, MED, MIN)
rownames(CV.example) = c("Uzorak", "Prognoza", "Greska")

CV.example.t = xtable(CV.example, caption="Nekoliko␣
    ↪ primjera␣validiranih␣tacaka", label="tab:CV.example",
    ↪  digits=4, align="lrrr")
print(CV.example.t, type="latex", file="CV.example.tex",
    ↪ floating=F, hline.after=NULL, add.to.row=list(pos=
```

```
    ↪ list(-1,0, nrow(CV.example.t)), command=c('\\toprule\
    ↪ n', '\\midrule\n', '\\bottomrule\n')), sanitize.text.
    ↪ function = function(x){x})

CV.list$p2o5.idw = p2o5.idw.cv
# tmp <- spplot(p2o5.idw["var1.pred"])
# png(filename="p2o5.idw.tiff.png")
# print(tmp)
# dev.off()
# ggplot(data.frame(p2o5.idw)) + geom_raster(aes(x, y, fill
    ↪ =var1.pred))
tmp1 = data.frame(p2o5.idw.cv)
tmp2 = data.frame(p2o5.idw.cv)
tmp1$Kolicina = tmp1$observed
tmp1$Tip = "Uzorak"
tmp2$Kolicina = tmp1$var1.pred
tmp2$Tip = "Prognoza"
df = do.call(rbind, list(tmp1, tmp2))
ggplot(data.frame(p2o5.idw)) + geom_tile(aes(x, y, fill=
    ↪ var1.pred)) + labs(fill="Kolicina␣fosfora") + geom_
    ↪ point(data=tmp1, aes(x=x,y=y,size=observed), shape
    ↪ =21, colour="black") + geom_point(data=tmp1, aes(x=x,
    ↪ y=y,size=var1.pred), shape=0, colour="green") + labs(
    ↪ size="Kolicina␣fosfora") + mytheme_map + theme(axis.
    ↪ text.x = element_blank(),
  axis.text.y = element_blank(),
  axis.ticks = element_blank(),
  axis.title.x = element_blank(),
  axis.title.y = element_blank())
save_plot_png("generated-map.p2o5.idw.png")

p2o5.ok = krige(p2o5~1, mydata.train, mydata.grid, model=
    ↪ p2o5.ok.vgm.fit)
#p2o5.ok = remove_outliers_from_prediction(p2o5.ok, new_max
```

```
    ↪ , new_min)
p2o5.ok.cv = krige.cv(p2o5~1, mydata.cv, model=p2o5.ok.vgm.
    ↪ fit)

error.abs = abs(p2o5.ok.cv$var1.pred - p2o5.ok.cv$observed)
mymax = max(error.abs)
myindex = which(error.abs==mymax)
MAX = c(p2o5.ok.cv$observed[myindex], p2o5.ok.cv$var1.pred[
    ↪ myindex], mymax)
mymed = median(error.abs)
myindex = which(error.abs==mymed)
MED = c(p2o5.ok.cv$observed[myindex], p2o5.ok.cv$var1.pred[
    ↪ myindex], mymed)
mymin = min(error.abs)
myindex = which(error.abs==mymin)
MIN = c(p2o5.ok.cv$observed[myindex], p2o5.ok.cv$var1.pred[
    ↪ myindex], mymin)
CV.example = data.frame(MAX, MED, MIN)
rownames(CV.example) = c("Uzorak", "Prognoza", "Greska")

CV.example.t = xtable(CV.example, caption="Nekoliko␣
    ↪ primjera␣validiranih␣tacaka", label="tab:CV.example",
    ↪  digits=4, align="lrrr")
print(CV.example.t, type="latex", file="CV.example.tex",
    ↪ floating=F, hline.after=NULL, add.to.row=list(pos=
    ↪ list(-1,0, nrow(CV.example.t)), command=c('\\toprule\
    ↪ n', '\\midrule\n', '\\bottomrule\n')), sanitize.text.
    ↪ function = function(x){x})

CV.list$p2o5.ok = p2o5.ok.cv
# tmp <- spplot(p2o5.ok["var1.pred"])
# png(filename="p2o5.ok.tiff.png")
# print(tmp)
# dev.off()
```

```
# ggplot(data.frame(p2o5.ok)) + geom_raster(aes(x, y, fill=
    ↪ var1.pred))
tmp1 = data.frame(p2o5.ok.cv)
tmp2 = data.frame(p2o5.ok.cv)
tmp1$Kolicina = tmp1$observed
tmp1$Tip = "Uzorak"
tmp2$Kolicina = tmp1$var1.pred
tmp2$Tip = "Prognoza"
df = do.call(rbind, list(tmp1, tmp2))
ggplot(data.frame(p2o5.ok)) + geom_tile(aes(x, y, fill=var1
    ↪ .pred)) + labs(fill="Kolicina␣fosfora") + geom_point(
    ↪ data=tmp1, aes(x=x,y=y,size=observed), shape=21,
    ↪ colour="black") + geom_point(data=tmp1, aes(x=x,y=y,
    ↪ size=var1.pred), shape=0, colour="green") + labs(size
    ↪ ="Kolicina␣fosfora") + mytheme_map + theme(axis.text.
    ↪ x = element_blank(),
   axis.text.y = element_blank(),
   axis.ticks = element_blank(),
   axis.title.x = element_blank(),
   axis.title.y = element_blank())
save_plot_png("generated-map.p2o5.ok.png")

p2o5.lrk = krige(myformula, mydata.train, mydata.grid)
#p2o5.lrk = remove_outliers_from_prediction(p2o5.lrk, new_
    ↪ max, new_min)
p2o5.lrk.cv = krige.cv(myformula, mydata.cv)

error.abs = abs(p2o5.lrk.cv$var1.pred - p2o5.lrk.cv$
    ↪ observed)
mymax = max(error.abs)
myindex = which(error.abs==mymax)
MAX = c(p2o5.lrk.cv$observed[myindex], p2o5.lrk.cv$var1.
    ↪ pred[myindex], mymax)
mymed = median(error.abs)
```

```
myindex = which(error.abs==mymed)
MED = c(p2o5.lrk.cv$observed[myindex], p2o5.lrk.cv$var1.
    ↪ pred[myindex], mymed)
mymin = min(error.abs)
myindex = which(error.abs==mymin)
MIN = c(p2o5.lrk.cv$observed[myindex], p2o5.lrk.cv$var1.
    ↪ pred[myindex], mymin)
CV.example = data.frame(MAX, MED, MIN)
rownames(CV.example) = c("Uzorak", "Prognoza", "Greska")

CV.example.t = xtable(CV.example, caption="Nekoliko␣
    ↪ primjera␣validiranih␣tacaka", label="tab:CV.example",
    ↪  digits=4, align="lrrr")
print(CV.example.t, type="latex", file="CV.example.tex",
    ↪ floating=F, hline.after=NULL, add.to.row=list(pos=
    ↪ list(-1,0, nrow(CV.example.t)), command=c('\\toprule\
    ↪ n', '\\midrule\n', '\\bottomrule\n')), sanitize.text.
    ↪ function = function(x){x})

CV.list$p2o5.lrk = p2o5.lrk.cv
# tmp <- spplot(p2o5.lrk["var1.pred"])
# png(filename="p2o5.lrk.tiff.png")
# print(tmp)
# dev.off()
# ggplot(data.frame(p2o5.lrk)) + geom_raster(aes(x, y, fill
    ↪ =var1.pred))
tmp1 = data.frame(p2o5.lrk.cv)
tmp2 = data.frame(p2o5.lrk.cv)
tmp1$Kolicina = tmp1$observed
tmp1$Tip = "Uzorak"
tmp2$Kolicina = tmp1$var1.pred
tmp2$Tip = "Prognoza"
df = do.call(rbind, list(tmp1, tmp2))
ggplot(data.frame(p2o5.lrk)) + geom_tile(aes(x, y, fill=
```

```
    ↪ var1.pred)) + labs(fill="Kolicina␣fosfora") + geom_
    ↪ point(data=tmp1, aes(x=x,y=y,size=observed), shape
    ↪ =21, colour="black") + geom_point(data=tmp1, aes(x=x,
    ↪ y=y,size=var1.pred), shape=0, colour="green") + labs(
    ↪ size="Kolicina␣fosfora") + mytheme_map + theme(axis.
    ↪ text.x = element_blank(),
   axis.text.y = element_blank(),
   axis.ticks = element_blank(),
   axis.title.x = element_blank(),
   axis.title.y = element_blank())
save_plot_png("generated-map.p2o5.lrk.png")

p2o5.uk = krige(myformula, mydata.train, mydata.grid,model=
    ↪ p2o5.uk.vgm.fit)
#p2o5.uk = remove_outliers_from_prediction(p2o5.uk, new_max
    ↪ , new_min)
p2o5.uk.cv = krige.cv(myformula, mydata.cv,model=p2o5.uk.
    ↪ vgm.fit)

error.abs = abs(p2o5.uk.cv$var1.pred - p2o5.uk.cv$observed)
mymax = max(error.abs)
myindex = which(error.abs==mymax)
MAX = c(p2o5.uk.cv$observed[myindex], p2o5.uk.cv$var1.pred[
    ↪ myindex], mymax)
mymed = median(error.abs)
myindex = which(error.abs==mymed)
MED = c(p2o5.uk.cv$observed[myindex], p2o5.uk.cv$var1.pred[
    ↪ myindex], mymed)
mymin = min(error.abs)
myindex = which(error.abs==mymin)
MIN = c(p2o5.uk.cv$observed[myindex], p2o5.uk.cv$var1.pred[
    ↪ myindex], mymin)
CV.example = data.frame(MAX, MED, MIN)
rownames(CV.example) = c("Uzorak", "Prognoza", "Greska")
```

```
CV.example.t = xtable(CV.example, caption="Nekoliko␣
    ↪ primjera␣validiranih␣tacaka", label="tab:CV.example",
    ↪  digits=4, align="lrrr")
print(CV.example.t, type="latex", file="CV.example.tex",
    ↪ floating=F, hline.after=NULL, add.to.row=list(pos=
    ↪ list(-1,0, nrow(CV.example.t)), command=c('\\toprule\
    ↪ n', '\\midrule\n', '\\bottomrule\n')), sanitize.text.
    ↪ function = function(x){x})

CV.list$p2o5.uk = p2o5.uk.cv
# tmp <- spplot(p2o5.uk["var1.pred"])
# png(filename="p2o5.uk.tiff.png")
# print(tmp)
# dev.off()
# ggplot(data.frame(p2o5.uk)) + geom_raster(aes(x, y, fill=
    ↪ var1.pred))
tmp1 = data.frame(p2o5.uk.cv)
tmp2 = data.frame(p2o5.uk.cv)
tmp1$Kolicina = tmp1$observed
tmp1$Tip = "Uzorak"
tmp2$Kolicina = tmp1$var1.pred
tmp2$Tip = "Prognoza"
df = do.call(rbind, list(tmp1, tmp2))
ggplot(data.frame(p2o5.uk)) + geom_tile(aes(x, y, fill=var1
    ↪ .pred)) + labs(fill="Kolicina␣fosfora") + geom_point(
    ↪ data=tmp1, aes(x=x,y=y,size=observed), shape=21,
    ↪ colour="black") + geom_point(data=tmp1, aes(x=x,y=y,
    ↪ size=var1.pred), shape=0, colour="green") + labs(size
    ↪ ="Kolicina␣fosfora") + mytheme_map + theme(axis.text.
    ↪ x = element_blank(),
  axis.text.y = element_blank(),
  axis.ticks = element_blank(),
  axis.title.x = element_blank(),
```

```
  axis.title.y = element_blank())
save_plot_png("generated-map.p2o5.uk.png")

p2o5.rnd.cv = p2o5.uk.cv
p2o5.rnd.cv$var1.pred = mydata.rnd.cv$p2o5
p2o5.rnd.cv$residual = p2o5.rnd.cv$observed - p2o5.rnd.cv$
  ↪ var1.pred
p2o5.rnd.cv$zscore = NA
CV.list$p2o5.rnd = p2o5.rnd.cv

CV.summary = compare.cv(CV.list)
var_expl = c()

var_expl = append(var_expl, 1-var(p2o5.idw.cv$residual)/var
  ↪ (mydata.cv$p2o5))

var_expl = append(var_expl, 1-var(p2o5.ok.cv$residual)/var(
  ↪ mydata.cv$p2o5))

var_expl = append(var_expl, 1-var(p2o5.lrk.cv$residual)/var
  ↪ (mydata.cv$p2o5))

var_expl = append(var_expl, 1-var(p2o5.uk.cv$residual)/var(
  ↪ mydata.cv$p2o5))

var_expl = append(var_expl, 1-var(p2o5.rnd.cv$residual)/var
  ↪ (mydata.cv$p2o5))
CV.summary["var_expl",] = var_expl
CV.summary = CV.summary[c(8, 6, 12), 1:length(CV.summary)]
#CV.summary = CV.summary[-c(2), 1:length(CV.summary)]
rownames(CV.summary) <- c("RMSE", "Cor", "$R^2$")
colnames(CV.summary) <- c("IDW","OK","LR","UK", "RND")
CV.summary.t = xtable(CV.summary, caption="Tacnost␣
  ↪ algoritama", label="tab:CV.summary", digits=4, align=
```

```
    ↪ "lrrrrr")
print(CV.summary.t, type="latex", file="CV.summary.train100
    ↪ .foldLOO.tex", floating=F, hline.after=NULL, add.to.
    ↪ row=list(pos=list(-1,0, nrow(CV.summary.t)), command=
    ↪ c('\\toprule\n', '\\midrule\n', '\\bottomrule\n')),
    ↪ sanitize.text.function = function(x){x})
CV.summary.train100.foldLOO = CV.summary
```

# Bibliografija


Bachmaier, Martin i Matthias Backes (2008). "Variogram or semivariogram? Understanding the variances in a variogram". *Precision Agriculture* 9.3, str. 173–175.

Belić, Milivoj, Ljiljana Nešić i Vladimir Ćirić (2014). *Praktikum iz pedologije*. Poljoprivredni fakultet, Novi Sad.

Cressie, Noel (2015). *Statistics for spatial data*. John Wiley & Sons.

de Ferranti, Jonathan (2014). *Worldwide 3" DEM*. URL: `http://www.viewfinderpanoramas.org/dem3.html` (pogledano 19. 5. 2015).

Foody, G.M. (2004). "Thematic map comparison: evaluating the statistical significance of differences in classification accuracy". *Photogrammetric Engineering and Remote Sensing* 70.5, str. 627–633.

Fuštić, Budimir i Grujica Đuretić (2000). *Zemljišta Crne Gore*. Univerzitet Crne Gore, Biotehnički institut, Podgorica.

Hengl, Tomislav (2009). *A Practical Guide to Geostatistical Mapping*. Lulu.

Hiemstra, P.H. i dr. (2008). "Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network". *Computers & Geosciences*.

Hutchinson, M. F. (1995). "Interpolating mean rainfall using thin plate smoothing splines". *International Journal of Geographical Information Systems* 9.4, str. 385–403.

Jenny, Hans (1994). *Factors of Soil Formation, A System of Quantitative Pedology (Reprint, with Foreword by R. Amundson, of the 1941 McGraw-Hill publication)*. Dover Publications, Inc.

Keshavarzi, Ali i dr. (2015). "A neural network model for estimating soil phosphorus using terrain analysis". *The Egyptian Journal of Remote Sensing and Space Science* 18.2, str. 127–135.

Krige, Danie G. (1951). "A statistical approach to some basic mine valuation problems on the Witwatersrand". Mag. rad. Johannesburg: University of the Witwatersrand.

Li, Jin i Andrew D. Heap (2008). *A Review of Spatial Interpolation Methods for Environmental Scientists*. Geoscience Australia, Record 2008/23.

Lichtenstern, Andreas (2013). "Kriging methods in spatial statistics". Bachelor's Thesis.

Matheron, Georges (1971). *The Theory of Regionalized Variables and its Applications*. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau 5. École Nationale Supérieure des Mines de Paris.

Matheron, Georges i Jean Serra (2002). "The birth of Mathematical Morphology". *Mathematical Morphology VI, Proc. ISMM'02*. Ur. H. Talbot i R. Beare. URL: `http://cmm.ensmp.fr/~serra/communications_pdf/C-72.pdf`.

McBratney, A.B., M.L. Mendonça Santos i B. Minasny (2003). "On digital soil mapping". *Geoderma* 117.1-2, str. 3–52.

Minasny, Budiman i Alex.B. McBratney (2016). "Digital soil mapping: A brief history and some lessons". *Geoderma* 264, Part B. Soil mapping, classification, and modelling: history and future directions, str. 301–311.

Mitasova, Helena i Lubos Mitas (1993). "Interpolation by Regularized Spline with Tension: I. Theory and Implementation". *Mathematical geology* 25.6, str. 641–655.

Mitasova, Helena, Lubos Mitas i Russell S Harmon (2005). "Simultaneous spline approximation and topographic analysis for lidar elevation data in open-source GIS". *Geoscience and Remote Sensing Letters, IEEE* 2.4, str. 375–379.

Oliver, Margaret, Richard Webster i John Gerrard (1989). "Geostatistics in Physical Geography. Part I: Theory". *Transactions of the Institute of British Geographers* 14.3, str. 259–269.

Pebesma, Edzer J. (2004). "Multivariable geostatistics in S: the gstat package". *Computers & Geosciences* 30, str. 683–691.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org`.

Resulović, Husnija i Hamid Čustović (2002). *Pedologija - Opći dio (Knjiga I)*. Univerzitet u Sarajevu.

Salković, Edin (2015). "Digitalizacija pedoloških podataka Crne Gore". *Zbornik radova Informacione Tehnologije - sadašnjost i budućnost*. IT '15. Podgorica: Univerzitet Crne Gore, Elektrotehnički fakultet, str. 64–67.

Sarmadian, Fereydoon i dr. (2014). "Digital mapping of soil phosphorus using multivariate geostatistics and topographic information". *Australian Journal of Crop Science* 8.8, str. 1216–1223.

Shepard, Donald (1968). "A Two-dimensional Interpolation Function for Irregularly-spaced Data". *Proceedings of the 1968 23rd ACM National Conference*. ACM '68. New York, NY, USA: ACM, str. 517–524.

Shiryayev, A.N. (1992). "Interpolation and Extrapolation of Stationary Random Sequences [translation of the original paper in Russian, published in 1941.]" *Selected Works of A. N. Kolmogorov*. Ur. A.N. Shiryayev. Sv. 26. Mathematics and Its Applications (Soviet Series). Springer Netherlands, str. 272–280.

Unser, Michael (1999). "Splines: A Perfect Fit for Signal/Image Processing". *IEEE Signal Processing Magazine* 16, str. 22–38.

USGS (2008a). *Collection Name: Global Land Survey, Epoch: 1975, Sensor name: Landsat MSS, Image Name: 60 meter scene p201r030_3dm19780706*. Sioux Falls, South Dakota.

— (2008b). *Collection Name: Global Land Survey, Epoch: 1990, Sensor name: Landsat TM, Image Name: 60 meter scene p187r031_5dt19870724*. Sioux Falls, South Dakota.

Webster, R. i M. A. Oliver (2001). *Geostatistics for Environmental Scientists*. Statistics in Practice. Chichester: Wiley.

Zhou, F. i dr. (2007). "Scientometric analysis of geostatistics using multivariate methods". *Scientometrics* 73, str. 265–279.