UNIVERZITET U BEOGRADU

FAKULTET ORGANIZACIONIH NAUKA

Milan M. Dobrota

STATISTIČKI PRISTUP DEFINISANJU ZONE OSETLJIVOSTI U METODAMA DALJINSKOG UZORKOVANJA

Doktorska disertacija

Beograd, 2018

UNIVERSITY OF BELGRADE

FACULTY OF ORGANIZATIONAL SCIENCES

Milan M. Dobrota

A STATISTICAL APPROACH TO SENSITIVITY ZONE DEFINITION IN REMOTE SENSING METHODS

Doctoral Dissertation

Belgrade, 2018

Mentor:

_____

Prof. dr Zoran Radojičić, redovni profesor
Univerziteta u Beogradu, Fakulteta organizacionih nauka

Članovi komisije:

_____

Prof. dr Dušan Starčević, redovni profesor u penziji
Univerziteta u Beogradu, Fakulteta organizacionih nauka

_____

Prof. dr Boris Delibašić, redovni profesor
Univerziteta u Beogradu, Fakulteta organizacionih nauka

_____

Doc. dr Aleksandar Đoković, docent
Univerziteta u Beogradu, Fakulteta organizacionih nauka

_____

Prof. dr Dušan Surla, profesor emeritus
Univerziteta u Novom Sadu, Prirodno-matematičkog fakulteta

Datum odbrane:

_____

STATISTIČKI PRISTUP DEFINISANJU ZONE OSETLJIVOSTI U METODAMA DALJINSKOG UZORKOVANJA

Rezime:

Osnovni predmet istraživanja u disertaciji je analiza metodologije daljinskog uzorkovanja s aspekta prikupljanja i pripreme podataka u formu pogodnu za obradu, i posebno obrade podataka statističkim metodama u svrhu klasifikacije, sa ciljem identifikacije određenih pojava. Poseban akcenat istraživanja je predlaganje metodologije za definisanje zone osetljivosti pri klasifikaciji pojava, primarmo pomoću statističkih metoda multivarijacione analize.

Daljinsko uzorkovanje se najčešće opisuje kao naučna oblast i tehnika za prikupljanje informacija o objektu (najčešće Zemljinoj površini) bez dolaženja u kontakt sa njim. Sprovodi se uzorkovanjem (očitavanjem) putem beleženja reflektovane ili emitovane energije (elektromagnetno zračenje, akustičnost, itd.) objekta, procesiranjem, analiziranjem i primenom informacija. Danas, daljinsko uzorkovanje je prepoznata interdisciplinarna oblast širom sveta. Često je uparena sa disciplinama obrade slika i geografskog informacionog sistema (GIS) za široku oblast geospatijalne nauke i tehnologije, te se u disertaciji čini osvrt na povezanost GIS-a i daljinskog uzorkovanja, gde je GIS nezaobilazan alat u analizi prostornih podataka. Svaka digitalna slika je sastavljena od piksela, koji čine najsitnije komponente jedne slike i koji imaju svoju osvetljenost i zabeležni spektar boja, a koje možemo smatrati pojedinačnim entitetima statističkog uzorka koji predstavlja sama slika. U disertaciji je od posebnog značaja klasifikacija daljinski uzorkovanih podataka i koristi se da identifikuje i klasifikuje delove ili piksele slike. Klasifikacija se izvodi na višestrukim skupovima podataka a cilj je dodeljivanje svakog piksela slike određenoj klasi na osnovu statističkih karakteristika intenziteta i obojenosti piksela.

Termin multivarijaciona analiza se koristi da predstavi višedimenzionalni aspekt analize podataka. Mnogobrojne pojave i fenomeni opisani su većim brojem različitih promenljivih, a to se svakako odnosi i na podatke dobijene daljinskim uzorkovanjem gde je svako piksel tipično predstavljen u tri ili više različitih opsega svetlosnog spektra. Multivarijaciona analiza se opisuje kao skup statističkih metoda koje simultano analiziraju višedimenziona merenja dobijena za svaku jedinicu posmatranja iz skupa objekata koji ispitujemo. U disertaciji su naročito opisane tehnike Analize grupisanja, koja je od manjeg značaja za kasnije predloženi metod, i Diskriminacione analize, koja je u srži predloženog metoda. Posebno je opisana Bajesova teorija odlučivanja kao fundamentalan statistički pristup problemu klasifikacije. Pristup bazira na kvantifikaciji kompromisa između različitih odluka klasifikacije pomoću verovatnoće i cene ili napora koji se javljaju tokom odlučivanja. Bajesova teorija odlučivanja pretpostavlja da je problem odlučivanja postavljen u probabilističkom kontekstu. Diskriminaciona analiza se bavi problemom razdvajanja grupa i alokacijom opservacija u ranije definisane grupe

U disertaciji je od posebnog značaja cilj diskriminacione analize koji se naziva klasifikacija a koji se odnosi na utvrđivanje postupka za klasifikaciju opservacija na osnovu vrednosti nekoliko promenljivih u dve ili više razdvojenih, unapred definisanih grupa.

U disertaciji je izvršen pregled tehnika koje se koriste za klasifikaciju u daljniskom uzorkovanju. Napravljen je uvod u pojam digitalnih slika, sa teoretskim osnovama, kao i u osnovne tehnike njihove obrade u daljinskom uzorkovanju a zatim i pregled standardnih procedura klasifikacije koje se grupišu u dve grupe. Prva je Klasifikacija sa nadzorom ili nadgledana klasifikacija, gde analitičar identifikuje delove slike kao reprezentativne uzorke tipova površine (klase informacija), koji se nazivaju oblastima učenja, a na osnovu čijih se spektralnih svojstva piksela algoritam „trenira“ da prepoznaje spektralno slične oblasti za svaku od klasa. Druga, Klasifikacija bez nadzora ili nenadgledana klasifikacija, u kom slučaju su spektralne klase prvo grupisane, jedino na osnovu svojih numeričkih podatka, a zatim se analitičar bavi njihovim validiranjem i mapiranjem u klase informacija (ukoliko je moguće). Kada je izlaz daljinskog uzorkovanja klasifikovana mapa, od suštinskog je značaja proceniti njenu tačnost. Najprihvaćeniji metod za procenu tačnosti u mapama proisteklim iz daljinskog uzorkovanja je pomoću poređenja sa referentnim podacima (poznatim kao „istina sa tla“, eng. ground truth). Bez obzira na statistički metod klasifikacije piksela jasno je da klasifikaciju treba razmatrati uz određeni nivo tačnosti klasifikovanja, te je iz tog razloga dat pregled tehnika merenja tačnosti klasifikacije. Jedan od ciljeva istraživanja opisanog u disertaciji je upravo razvoj metodologije koja će (korišćenjem zone osetljivosti) pospešiti ukupnu tačnost klasifikovanja, pri čemu će potrebno vreme i resursi za sprovođenje klasifikacije biti smanjeni.

U šestom poglavlju je prvo detaljnije diskutovan problem klasifikacije, određivanje zone osetljivosti i stepena preklapanja, a potom je dat predlog nove metode računanja i primene zone osetljivosti u daljinskom uzorkovanju, kao rezultat istraživačkog rada sprovedenog tokom izrade disertacije. Problem klasifikacije nastaje kada istraživač napravi potreban broj merenja i želi da klasifikuje individue u jednu od nekoliko kategorija na bazi ovih mera. Osnovno pitanje je iz koje populacije potiče navedeni entitet, u našem slučaju piksel, sa određenim merama. U konstrukciji procedure klasifikacije potrebno je minimizirati verovatnoću pogrešne klasifikacije, ili konkretnije, poželjno je minimizirati rezultate loših efekata pogrešne klasifikacije. Zona osetljivosti predstavlja granične vrednosti kategorija a definiše se u slučajevima kada nismo u mogućnosti da utvrdimo kategoriju entiteta. Osim određivanja kategorija pojava, bitno je odrediti i granice tih pojava, tj. odrediti one oblasti u kojima možemo sa sigurnošću pretpostaviti da entiteti pripadaju definisanoj kategoriji, i one kada to ne možemo pretpostaviti sa sigurnošću. Definisanjem metodologije određivanja zone osetljivosti rešava se problem na način da se identifikuju elementi koji se nalaze u graničnoj oblasti, između dve kategorije, te je jedan od ciljeva istraživanja opisanog u disertaciji upravo iznalaženje pogodne metodologije za određivanje zone osetljivosti

kod daljinskog uzorkovanja, koja će na unapređen način piksele klasifikovati u odgovarajuće klase. U disertaciji se detaljno, po koracima, opisuje metoda za određivanje zone osetljivosti, i istaknute su njene prednosti u odnosu na dosadašnje metode klasifikacije daljinski uzorkovanih podataka. Prednosti se prvenstveno ogledaju u unapređenju tačnosti klasifikacije pri smanjenju ukupnog vremena i računarskih resurasa potrebnih za računanje. U predloženoj metodi se klasifikacija sprovodi u dve faze, u prvoj se primenjuje veoma brza metoda određivanja (hiper)ravni podele pomoću diskriminacione analize, zatim se utvrđuju entiteti u zoni osetljivosti i na njih primenjuje u drugom koraku računski zahtevna klasifikacija pomoću metode *k*-najbližih suseda (kNN). Takođe se definiše i način merenja tačnosti klasifikacije.

U sedmom poglavlju prikazane su studije slučaja i mogućnosti primene zone osetljivosti u daljinskom uzorkovanju u preciznoj poljoprivredi. Opisan je problem izdvajanja useva od zemljišta (ili korova ili drugih objekata, odnosno svega što ne predstavlja usev – biljke od interesa). Predložena metoda je sprovedena na nekoliko različitih vrsta useva i rezultati su upoređeni sa ostalim metodama klasifikacije u upotrebi, SVM, kNN (koja je manje u upotrebi zbog performansi na velikim količinama podataka ali je zbog postignute tačnosti korišćena kao referentna) i LDA/QDA bez računanja i upotrebe zone osetljivosti. S obzirom na prirodu podataka nad kojim je vršeno testiranje, naročito je dat doprinos u pogledu mogućnosti analiza slika prikupljenih pomoću bespilotnih letelica, i to pre svega u poljoprivredi u nadzoru jednogodišnjih i višegodišnjih useva sa specifičnostima koje takva primena nosi.

Na kraju je dat zaključak sa odgovorima na pitanja u vezi sa postavljenim ciljem i hipotezama, a u prilogu je ukratko opisano realizovano softversko rešenje pomoću kojeg je sprovođeno istraživanje i analiza podataka.

Ključne reči:
daljinsko uzorkovanje, diskriminaciona analiza, Bajesova procedura, klasifikacija, zona osetljivosti

Naučna oblast:
Tehničke nauke

Uža naučna oblast:
Računarska statistika

UDK broj:
519.21/.24

A STATISTICAL APPROACH TO SENSITIVITY ZONE DEFINITION IN REMOTE SENSING METHODS

Abstract:

The main topic of the research in this dissertation is the analysis of remote sensing methodology from the aspect of collecting and preparing data in a form suitable for processing, and in particular data processing by the means of statistical methods for the purpose of classification, to identify certain phenomena. A particular emphasis of the research is to propose a methodology for defining the sensitivity zone in the classification of phenomena, primarily using statistical methods of multivariate analysis.

Remote sensing is typically described as a scientific field and technique for collecting information about an object (usually the Earth's surface) without being in direct contact with the object. It is carried out by sampling through recording reflected or emitted energy (electromagnetic radiation, acoustics, and others) of the object, processing, analysing, and applying the information. Today, remote sensing is recognized as an interdisciplinary field all over the world. It is often paired with image processing disciplines and a geographic information system (GIS) for a vast area of geospatial science and technology, hence, in this dissertation a link between GIS and remote sensing is described, where GIS is a necessary tool in spatial data analysis. Each digital image is composed of pixels, which are the smallest components of a single image and have their brightness and a color spectrum recorded, which we can consider as individual entities of a statistical sample representing the image. In this dissertation, the classification of remotely-sensed data is of particular importance and is used to identify and classify parts or the pixels of the image. The classification is performed on multiple data sets, and the goal is to assign each pixel of an image to a particular class based on the statistical characteristics of the intensity and color of the pixels.

The term multivariate analysis is used to present a multidimensional aspect of data analysis. Numerous phenomena are described by many different variables, and this certainly applies to data obtained by remote sensing where each pixel is typically represented in three or more different light spectrum bandwidths. Multivariate analysis is described as a set of statistical methods that simultaneously analyze multidimensional measurements obtained for each observation unit from a set of objects that we are examining. This dissertation describes the Data clustering techniques, which are, though, of less importance to the proposed method, and the Discriminant Analysis, which is at the core of the proposed method. Bayesian decision theory is described as a fundamental statistical approach to the problem of classification. The approach is based on the quantification of the compromise between different classification decisions using the probability and cost or effort that arise during decision-making. Bayesian decision theory assumes that the decision-making problem is set in a probabilistic context. Discriminant analysis deals with the problem of group separation and the allocation of

observations in previously defined groups. In this dissertation, it is of particular importance the objective of discriminant analysis called classification, which refers to the establishment of a procedure for classifying observations based on the value of several variables in two or more separated, predefined a group.

The dissertation reviewed the techniques used for the classification in the remote sensing. An introduction is made to the notion of digital images with theoretical foundations, as well as the basic techniques of their processing in remote sensing, and after a review of widely used classification methods which are grouped into two groups. The first is the Supervised classification, where the analyst identifies the parts of the image as representative samples of the surface types (class of information), which are called training sets, based on whose pixel's spectral properties the algorithm "trains" to recognize spectrally similar areas for each of the class. Second, Unsupervised classification, in which case the spectral classes are grouped only based on their numerical data, and then the analyst deals with their validation and mapping into the information classes (when possible). When the remote sensing output is a classified map, it is essential to evaluate its accuracy. The most accepted method for estimating the accuracy in the maps generated by remote sensing is by comparison with reference data (known as "ground truth"). Regardless of the statistical method applied for pixel classification, it is clear that the classification should be considered with a certain level of classification accuracy, and for this reason, an overview of the techniques for measuring the accuracy of the classification is given. One of the aims of the research described in this dissertation is development of a methodology that will (by using the sensitivity zone) improve the overall classification accuracy, with the time and resources needed for running the classification will be reduced.

In sixth chapter, the problem of classification was first discussed in more detail, with determining the sensitivity zone and overlapping level, and then the proposal of a new method to calculate and apply the sensitivity zone in remote sensing, as a result of the research work carried out during this dissertation, is given. The classification problem arises when the researcher makes the required number of measurements and wants to classify individuals into one of several categories based on these measures. The fundamental question is from which population is the given entity, in our case, the pixel, with specific measures. In the construction of the classification procedure, it is necessary to minimize the probability of the wrong classification, or more specifically, it is desirable to minimize the results of the effects of the wrong classification. The sensitivity zone represents the boundary values of categories and is defined in cases where we are unable to determine the category of entities. Therefore, in addition to determining categories of phenomena, it is also essential to determine the boundaries of these phenomena, i.e., identify those areas in which we can safely assume that entities belong to a defined category and those that cannot be presumed with certainty. By defining the methodology of determining the sensitivity zone, the problem is solved by identifying the elements that are in the border region, between the two categories, and

one of the aims of the research described in this dissertation is exactly finding the appropriate method for determining the sensitivity zone in remote sensing, with the aim to improve the way in which the pixels are classified in the appropriate classes. In this dissertation, in detailed steps, the method for determining the sensitivity zone is described, and its advantages are highlighted in reference to other methods of classification of remotely sensed data. The benefits are primarily reflected in improving the accuracy of the classification and in reducing the total time and computing resources required. In the proposed method, the classification is carried out in two phases: in the first phase, a very fast method of determining the (hyper)plane of discrimination by the means of discriminant analysis is used, then the entities in the sensitivity zone are identified and in the next step pixels are classified using computationaly more complex classification using the *k*-nearest neighbors method (kNN). It also defines the method of measuring the accuracy of the classification.

The seventh chapter presents the case studies and the possibilities of applying the sensitivity zone in the remote sensing for precision agriculture. The problem of separating the crops from the soli (or weeds or other objects, or anything that does not represent crops of interest) is described. The proposed method was tested in several different types of crops, and the results were compared with other classification methods in use, SVM, kNN (which is less used because of the performance issues on big amounts of data, but because of the accuracy was used as a reference) and LDA/QDA without calculating using the sensitivity zone. Considering the nature of the data tested, a contribution was made in particular to the possibility of analyzing images collected by unmanned aircraft, primarily in agriculture in the control of seasonal and perennial crops with the specificities that such an application carries.

Finally, a conclusion was given with answers to questions related to the goala and hypotheses set, and the appendix briefly describes the developed software solution through which the research and data analysis was conducted.

Keywords:
remote sensing, discriminant analysis, Bayesian procedure, classification, sensitivity zone

Scientific Area:
Technical Sciences

Specific Scientific Area:
Computational Statistics

UDK Number:
519.21/.24

SADRŽAJ

1 UVOD

Predmet istraživanja u doktorskoj disertaciji je analiza metodologije daljinskog uzorkovanja (eng. *Remote Sensing*) sa aspekta prikupljanja i pripreme podataka u formu pogodnu za obradu, i posebno obrade podataka statističkim metodama u svrhu klasifikacije, sa ciljem identifikacije određenih pojava, koje će unaprediti donošenje odluka od interesa u oblasti primene. Poseban akcenat je stavljen na predlaganje metoda za definisanje zone osetljivosti pri klasifikaciji pojava, primarmo pomoću statističkih metoda multivarijacione analize.

Cilj istraživanja je unapređenje rezultata daljinskog uzorkovanja razvojem metodoloških elemenata za povećanje tačnost klasifikacije daljinskih uzoraka, te bolje razumevanje i strukturu zaključaka kada veću tačnost nije moguće postići.

U disertaciji se prvo opisuje *Geografski informacioni sistem* (GIS) kao jedan od osnovnih alata u metodama daljinskog uzorkovanja. GIS se može definisati kao kompjuterski informacioni sistem koji prikuplja, skladišti, analizira i prikazuje prostorne entitete i njihove atribute, za rešavanje kompleksnih istraživačkih, projektantskih i problema upravljanja (Fischer & Nijkamp, 1992). Iz pogleda korisnika GIS je sistem zasnovan na računaru koji pruža sledeće mogućnosti za upravljanje georefenciranim podacima: prikupljanje i priprema podataka, upravljanje podacima (uključujući skladištenje i održavanje), manipulacije nad podacima i analize i na kraju prezentovanje podataka. Sve ovo implicira da GIS korisnici mogu očekivati podršku od strane sistema za unos (georeferenciranih) podataka, da ih analiziraju na razne načine i da proizvedu prezentacije podataka (uključujući mape i druge načine) (Huisman & de By, 2009). Podaci su činjenice, mere i karakteristike o nečemu od interesa a prostorni podaci (eng. *spatial data*) se odnose na geografske objekte realnog sveta od značaja, poput ulica, zgrada, jezera, zemalja, sa svojim respektivnim lokacijama. Pored lokacije, svaki od ovih objekata takođe poseduje određene osobine od značaja, ili atribute, poput naziva, broj spratova, dubina ili populacija (Campbell & Shin, 2011). GIS softver vodi računa o oba tipa podataka, odnosno i o prostornim podacima i o atributima i dozvoljava povezivanje ova dva tipa podatka kako bi se dobile potrebne informacije i olakšale analize. Prostorni podaci opisuju geografske prostorne aspekte fenomena, dok atributski podaci opisuju kvalitete i karakteristike datog fenomena, a njihova integracija je jedna od osnovnih karakteristika GIS-a (Campbell & Shin, 2011; Sutton, Dassau, & Sutton, 2009).

Daljinsko uzorkovanje je naučna oblast i tehnika za prikupljanje informacija o objektu (najčešće Zemljinoj površini) bez dolaženja u kontakt sa njim (dakle na daljinu, npr. iz letelice, satelita, itd.). Sprovodi se uzorkovanjem (očitavanjem) putem beleženja reflektovane ili emitovane energije (elektromagnetno zračenje, akustičnost, itd.) objekta, procesiranjem, analiziranjem i primenom informacija (Khorram, Wiele, Koch, Nelson,

& Potts, 2016). Oblasti primene tehnika daljinskog uzorkovanja su brojne, a ovde navodimo neke (Levin, 1999): poljoprivreda, šumarstvo, geologija, hidrologija, morski led, zemljišni pokrivač i upotreba zemljišta, mapiranje, nadzor okeana i obala, itd. Danas, daljinsko uzorkovanje je prepoznata interdisciplinarna oblast širom sveta. Često je uparena sa disciplinama obrade slika i geografskog informacionog sistema (GIS) za široku oblast geospatijalne nauke i tehnologije.

Elementi neophodni za izvođenje daljinskog uzrokovanja su: 1) Izvor energije ili osveljenost; 2) Zračenje i uticaj atmosfere, 3) Interakcija energije (svetlosti) sa objektom, 4) Snimanje energije putem senzora, 5) Emitovanje, prijem i procesiranje energije, 6) Interpretacija i analiza, 7) Primena (Canada Centre for Remote Sensing, 2007). Za nas je u ovoj disertaciji od posebnog značaja praktično samo tačka 6, koja se odnosi na interpretaciju i analizu podataka zabeleženih u formi digitalnih slika, i to njihova interpretacija računarskim putem, odnosno statističkim metodama. Ovde je ključna činjenica da je svaka digitalna slika sastavljena od piksela, koji čine najsitnije komponente jedne slike i koji imaju svako svoju osvetljenost i zabeležni spektar boja, a koje mi u daljem razmatranju možemo smatrati pojedinačnim entitetima statističkog uzorka koji predstavlja sama slika.

Ključni deo procesa je ekstrakcija smislenih informacija iz slike putem njene interpretacije i analize. To uključuje identifikaciju i/ili merenje različitih ciljnih objekata na slici kako bi se dobile korisne informacije o njima. Ciljni objekti mogu biti tačke, linije ili površine i moraju imati svojstvo da se mogu razlikovati od ostalih oblika na slici. Dakle, automatsko procesiranje i analiza digitalnih slika se sprovodi za automatsku identifikaciju ciljnih objekata i dobijanje informacija, kada je moguće bez ručnih intervencija analitičara. Ipak, procesiranje i analiza se u praksi retko sprovode u potpunosti bez ljduskih intervencija i nadzora, već kao pomoć analizi koju sprovodi analitičar. Međutim, kada govorimo o simultanoj analizi većeg broja spektara, koji podrazumevaju i veću količinu podataka, nesporan je značaj uloge automatskog procesiranja za ukupnu efikasnost procesa. U tome se takođe ogleda i jedan od doprinosa ove disertacije – povećanje nivoa automatizacije procesa automatskog uzorkovanja. U tom smislu prepoznavanje ciljnih objekata je ključ interpretacije podataka i dobijanja informacija, što podrazumeva posmatranje razlika između ciljnog objekta i njegove okoline poređenjem nekih, ili svih, njegovih vizuelnih elemenata, poput boje, oblika, veličine, obrasca, teksture, senke i asocijativnosti.

Većina standardnih funkcija obrade i analize slika se mogu kategorizovati u sledeće četiri kategorije: *predprocesiranje*, uključuje operacije koje su zahtevane pre glavne analize podataka i ekstrakcije informacija, *poboljšanja*, uključuje prikaz slike, najčešće od značaja samo prilikom vizuelne interpretacije i analiza slika, *transformacije*, za razliku od poboljšanja slike, transformacija obično uključuje kombinovano procesiranje podataka iz više spektralnih opsega, i *klasifikacija i analiza*, koristi se da identifikuje i klasifikuje delove ili piksele slike, odnosno podatke. Obično se izvodi na višestrukim skupovima podataka a cilj je dodeljivanje svakog piksela slike

određenoj klasi (npr. voda, vrsta šume, kukuruz, pšenica), na osnovu statističkih karakteristika intenziteta i obojenosti piksela (Canada Centre for Remote Sensing, 2007).

Koraci koji se odnose na predprocesiranje, poboljšanja i transformacije slika su od značaja za ovo istraživanje jedino u smislu da se bave pripremom ulaznih podataka, te ćemo se samo na tom nivou njima i baviti, dok ih nećemo dublje razrađivati. U osnovi, ovo će se istraživanje baviti tehnikama klasifikacije i analize podataka. Standardne procedure klasifikacije se obično grupišu u dve grupe na osnovu korišćenih metoda:

- *Klasifikacija s nadzorom* ili *nadgledana klasifikacija* (eng. *Supervised classification*): gde analitičar identifikuje delove slike kao reprezentativne uzorke tipova površine (klase informacija), koji se nazivaju oblastima učenja, a na osnovu čijih se spektralnih svojstva piksela algoritam „trenira“ da prepoznaje spektralno slične oblasti za svaku od klasa. Jedna od oblasti multivarijacione analize koja se bavi ovakvom klasifikacijom je Diskriminaciona analiza.
- *Klasifikacija bez nadzora* ili *nenadgledana klasifikacija* (eng. *Unsupervised classification*): u ovom slučaju su spektralne klase prvo grupisane, jedino na osnovu svojih numeričkih podatka, a zatim se analitičar bavi njihovim validiranjem i mapiranjem u klase informacija (ukoliko je moguće). Jedna od oblasti multuvarijacione analize koja se bavi ovom vrstom klasifikacije je Analiza grupisanja (Klaster analiza).

Kod daljinskog uzorkovanja, pored tehnika klasifikacije, se takođe koristi i broj drugih statističkih metoda, kao što su regresija, kanonočka korelaciona analiza, Bajesove mreže uslovnih verovatnoća, itd. (Evans, 1998; Deekshatulu, i drugi, 1995) ali one nisu bile u fokusu istraživanja, već su date pregledno u ovoj disertaciji. Sveobuhvatan uvod u tehnike daljinskog uzorkovanja, sa svim elementima, može se naći u (Khorram, Wiele, Koch, Nelson, & Potts, 2016; Canada Centre for Remote Sensing, 2007; Levin, 1999).

Multivarijaciona analiza predstavlja skup statističkih metoda koje simultano analiziraju višedimenziona merenja dobijena za svaku jedinicu posmatranja iz skupa objekata koji ispitujemo. Ovim metodama istovremeno postižemo pojednostavljivanje složene strukture posmatranog fenomena u cilju njegove lakše interpretacije. Pored ovog, pre svega deskriptivnog zadatka, metode multivarijacione analize koristimo u procesu zaključivanja, tako što ocenjujemo, na primer stepen međuzavisnosti promenljivih i/ili testiramo njihovu statističku značajnost (Kovačić, 1994). Takođe, neke od metoda multivarijacione analize su istraživačkog karaktera, što će reći da se koriste ne za testiranje apriori definisanih hipoteza, nego za njihovo generisanje, odnosno konstruisanje. (Kovačić, 1994) Tokom istraživanja, posebno smo se bavili tehnikama Diskriminacione analize i u manjoj meri Analizom grupisanja.

Diskriminaciona analiza se bavi problemom razdvajanja grupa i alokacijom opservacija u ranije definisane grupe. Primena diskriminacione analize omogućava

identifikaciju promenljive koja je najviše doprinela razdvajanju grupa kao i predviđanje verovatnoće da će objekat pripasti jednoj od grupa, na osnovu vrednosti skupa nezavisnih promenljivih (Kovačić, 1994). Ona ima dva osnovna cilja: Prvi, da utvrdi da li postoji statistički značajna razlika u sredinama dve ili više grupa, a zatim da odredi koja od promenljivih daje najveći doprinos utvrđenoj razlici. Ovaj cilj analize nazivamo diskriminacija ili razdvajanje između grupa. Drugi cilj odnosi se na utvrđivanje postupka za klasifikaciju opservacija na osnovu vrednosti nekoliko promenljivih u dve ili više razdvojenih, unapred definisanih grupa. Ovaj cilj analize nazivamo klasifikacija ili alokacija opservacija (Kovačić, 1994). Strateški pristup u diskriminacionoj analizi podređen je nalaženju sredstava za razdvajanje grupa, odnosno pouzdanu klasifikaciju opservacija. S obzirom da klasifikacija pomoću samo jedne od zavisnih promenljivih često dovodi do suviše velikog broja grešaka klasifikacije, želimo da odredimo linearnu kombinaciju nezavisnih promenljivih tako da se minimizira verovatnoća pogrešne klasifikacije opservacija – kombinacija dve ili više promenljivih može doprineti boljoj klasifikaciji. Više detalja se može naći u (Kovačić, 1994; Duda, Hart, & Stork, 2000; Radojičić, 2001; Webb & Copsey, 2011).

Analiza grupisanja je metoda za redukciju podataka koja je orijentisana ka entitetima (objektima) matrice podataka. Ovom analizom kombinujemo objekte u grupe relativno homogenih objekata. Zadatak u mnogim istraživanjima upravo je identifikovanje manjeg broja grupa, tako da su elementi koji pripadaju nekoj grupi u izvesnom smislu sličniji jedan drugom, nego što su to elementi koji pripadaju drugim grupama (Kovačić, 1994). Dakle, osnovni zadatak analize grupisanja jeste podela skupa objekata na grupe, tako da su varijacije između grupa znatno veće od varijacija unutar grupa. Dok su kod diskriminacione analize grupe unapred poznate, to kod analize grupisanja nije slučaj. Ovde samo pretpostavljamo da objekti pripadaju jednoj od "prirodnih" grupa ili jednostavno želimo izvršiti grupisanje objekata u izvestan manji broj grupa (Kovačić, 1994). U prvom koraku vrši se izbor izbog odgovarajuće mere sličnosti ili odstojanja, a zatim se formira matrica sličnosti (ili odstojanja) između objekata. Nakon toga vrši se izbor metode grupisanja. Metode grupisanja se mogu podeliti na Hijerarhijske i Nehijerarhijske metode. Primenom određene metode grupisanja na matricu sličnosti objekata, može se identifikovati određeni broj grupa, koji može ili ne mora biti zadat unapred. Više detalja se može naći u (Kovačić, 1994; Radojičić, 2001; Duda, Hart, & Stork, 2000; Rogerson, 2001; Webb & Copsey, 2011).

S obzirom na opisanu dvostepenu prirodu klasterovanja, metodološki pristup koji će biti predložen kao rezultat ovog istraživanja uvažavaće eventualnu potrebu grupisanja piksela u dva koraka: piksel=>Spektralna klasa=>Klasa informacija. Dva nivoa grupisanja nisu nov pristup rešavanja pojedinih problema, ali je činjenica da takav pristup zahteva eksplorativni pristup koji se odnosi na način kako pristupiti grupisanju formiranih klastera u prvom koraku. Sličan pristup je korišćen u (Đoković, 2013) i (Dobrota, Delibašić, & Delias, 2016). U ovom istraživanju kao deo predložene

metodologije je razmatran i analiziran dvostepeni (dvofazni) pristup grupisanju (klsterovanju), odnosno klasifikaciji.

Kada je izlaz daljinskog uzorkovanja klasifikovana mapa, od suštinskog je značaja proceniti njenu tačnost. Mapa je zaista nesavršena reprezentacija fenomena kojeg oslikava. Drugim rečima, svaka mapa sadrži greške, i odgovornost je analitičara da okarakteriše te greške pre nego se mapa nađe u daljoj upotrebi. Najprihvaćeniji metod za procenu tačnosti u mapama proisteklim iz daljinskog uzorkovanja je pomoću poređenja sa referentnim podacima (poznatim kao „istina sa tla“, eng. *ground truth*) prikupljenim obilaskom i adekvatnim uzorkovanjem lokacija na licu mesta (Khorram, Wiele, Koch, Nelson, & Potts, 2016). Bez obzira na statistički metod klasifikacije piksela u Spektralne klase i potom u Klase informacija, jasno je da klasifikaciju treba razmatrati uz određeni nivo preciznosti klasifikovanja. U literaturi se mogu naći teoretske osnove i radovi koji se bave tačnošću klasifikovanja kod daljinskog uzorkovanja i svakako su uzeti u obzir tokom istraživanja. Kako se u istraživanju posebna pažnja posvećuje tačnosti klasifikovanja, jasno je da je od posebnog značaja izučavanje zone osetljivosti kod klasifikovanja, koja se odnosi na upravo na onaj opseg vrednosti u kom se objekat ne može sa zadovoljavajućom sigurnošću klasifikovati u jednu od populacija, ili u našm slučaju piksel klasifikovati u jednu od Spektralnih ili Klasa informacija. Jedan od ciljeva istraživanja je razvoj metodologije za definisanje takve zone osetljivosti koja će pospešiti ukupnu tačnost klasifikovanja.

Problem klasifikacije nastaje kada istraživač napravi potreban broj merenja i želi da klasifikuje individue u jednu od nekoliko kategorija na bazi ovih mera. Istraživač ne može da identifikuje individue sa kategorijama direktno već mora da koristi dobijene rezultate merenja. U mnogim slučajevima može se pretpostaviti da postoji konačan broj ovih kategorija ili populacija od kojih su ove individue potekle i svaka populacija je kategorizovana verovatnoćom raspodele tih mera. Stoga se individue podrazumevaju kao slučajne opservacije ovih populacija. Osnovno pitanje je iz koje populacije potiče navedena individua (entitet) sa određenim merama. Problem klasifikacije može se posmatrati kao problem statističke funkcije odluke. Imamo određen broj hipoteza, a svaka hipoteza je definisana raspodelom opservacija. Mi moramo da prihvatimo jednu od ovih hipoteza i odbacimo ostale. Ukoliko su dve populacije poznate mi imamo elementaran problem testiranja jedne hipoteze specifične raspodele u odnosu na drugu. U nekim okolnostima kategorije su definisane unapred u smislu da je verovatnoća raspodele u potpunosti poznata. U ostalim slučajevima forma svake raspodele može biti poznata, ali parametri raspodele su ocenjeni kao primeri iz te populacije. Dakle, u konstrukciji procedure klasifikacije potrebno je minimizirati verovatnoću pogrešne klasifikacije, ili konkretnije, poželjno je minimizirati rezultate loših efekata pogrešne klasifikacije. (Radojičić, 2001)

Mi ćemo podrazumevati načine definisanja minimalnih gubitaka u dva slučaja. Pretpostavimo da imamo prioritetne verovatnoće kategorije za ove dve populacije. Neka verovatnoća opservacije koja dolazi iz populacije P_1 bude q_1, a verovatnoća opservacije

koja dolazi iz populacije P_2 bude q_2. Verovatnoća mogućih opcija populacije P_1 su definisane funkcijama raspodele. Mi ćemo tretirati samo slučaj kada raspodela ima određenu gustinu, iako slučaj diskretne verovatnoće dozvoljava sebi isti tretman. Neka gustina populacije P_1 bude $p_1(x)$, a gustina populacije P_2 bude $p_2(x)$. Ako imamo region R_1 klasifikacije P_1 verovatnoća tačne klasifikacije opservacije koja u stvari potiče iz populacije P_1 je:

$$P(1/1, R) = \int_{R_1} p_1(x)dx$$

gde je $dx=dx_1,...,dx_p$, a verovatnoća pogrešne klasifikacije opservacije koja dolazi iz populacije P_1 je:

$$P(2/1, R) = \int_{R_2} p_1(x)dx$$

Verovatnoća tačne klasifikacije za opservaciju koja dolazi iz populacije P_2 je:

$$P(2/2, R) = \int_{R_2} p_2(x)dx$$

dok je verovatnoća pogrešne klasifikacije za opservaciju koja dolazi iz populacije P_2 je:

$$P(1/2, R) = \int_{R_1} p_2(x)dx$$

Kako je verovatnoća izvlačenja opservacija iz populacije P_1 jednaka q_1, onda je verovatnoća tačne klasifikacije je $q_1P(1/1, R)$ (entitet je klasifikovan ispravno u populaciju P_1). Analogno ovome su i ostala tri slučaja, koji se odnose na ispravnu ili neispravnu klasifikaciju. Prosek očekivanih gubitaka pogrešne klasifikacije je suma proizvoda uzroka svake pogrešne klasifikacije pomnožena sa verovatnoćom njihovog pojavljivanja i ona je

$$C(2/1)*P(2/1, R)*q_1 + C(1/2)*P(1/2, R)*q_2$$

To je prosek gubitaka koji želimo da minimiziramo tj. hoćemo da podelimo naš prostor na regione R_1 i R_2 tako da očekivani gubitak bude što je moguće manji. Procedura koja minimizira prethodni izraz za dato q_1 i q_2 se zove Bajesova procedura.

Zona osetljivosti, koja predstavlja granične vrednosti kategorija, se definiše u slučajevima kada nismo u mogućnosti da utvrdimo stanje entiteta. Dakle, osim određivanja kategorija pojava, bitno je odrediti i granice tih pojava, tj. odrediti one oblasti u kojima možemo sa sigurnošću pretpostaviti da entiteti pripadaju definisanoj kategoriji, i one kada to ne možemo pretpostaviti sa sigurnošću (Radojičić, 2001). Problem nastaje kada su verovatnoće pripadanja entiteta jednoj ili drugoj kategoriji dosta male i ne zadovoljavaju nivo poverenja koji želimo da zadržimo. Definisanjem metodologije određivanja oblasti osetljivosti rešavamo problem na način da identifikujemo elemente koji se nalaze u graničnoj oblasti, između dve kategorije, a koju nazivamo oblast osetljivosti (Radojičić, 2001). Jedan od glavnih ciljeva ove doktorske disertacije je upravo iznalaženje pogodnog metoda za određivanje zone

osetljivosti kod daljinskog uzorkovanja, koja će na prihvatljiv i unapređen način klasifikovati piksele u odgovarajćue klase.

S obzirom da smo ušli u eru opservacija Zemljine površine sa visokim rezolucijama, podaci prikupljeni daljinskim uzorkovanjem doživljavaju eksplozivan rast (Ma, i drugi, 2015). Brz rast količine podataka takođe utiče na povećanje kompleksnosti podataka daljinskog uzorkovanja, kao što su raznovrsnost i visoka dimenzionalnost. Time takvi podaci sa pravom dobijaju epitet *Big Data* i s tim u vezi zahtevaju dodatnu dimenziju razmatranja koje se odnosi na performanse obrade podataka i generalno tehnike neophodne sa prikupljanje, čuvanje i obradu tako velikih količina podataka. Ova činjenica je značajno uzeta u obzir tokom ovog istraživanja i predloženi metodološki pristupi uvažavaju činjenicu o obimu podataka i sve performansne aspekte koje takva činjenica nosi. Dalje, veoma je važna povezanost daljinskog uzorkovanja i Geografskih informacionih sistema (GIS). Podaci dobijeni daljinskim uzorkovanjem su geospatijalni po svojoj prirodi, što znači da su opažene oblasti i objekti referencirani prema svojoj lokaciji u geografskom koordinatnom sistemu, tako da mogu biti locirani na mapi. Ovo omogućuje da takvi podaci budu analizirani u sprezi sa ostalim geospatijalnim podacima, npr. poput onih koji oslikavaju mreže puteva, gustinu naseljenosti, itd. Podaci daljinskog uzorkovanja sa dovoljno detalja mogu biti korišćeni da okarakterišu stvari koje ne mogu drugačije biti efikasno opažene na mapama kreiranim pomoću terenskih opservacija. Ova činjenica ilustruje jedinstven značaj daljinskog uzorkovanja kao izvora podataka GIS, koji su organizovani skupovi hardvera, softvera, geografskih podataka i osoblja namenjenih efikasnom beleženju, čuvanju, ažuriranju, manipulaciji i analizi svih oblika geografski referenciranih podataka. (Khorram, Wiele, Koch, Nelson, & Potts, 2016)

U daljem razvoju tehnika i primena daljinskog uzorkovanja, pored inovacija koje se konstantno dešavaju kod tradicionalne satelitske tehnologije, razvija se i spektar novih platformi za prikupljanje podataka putem daljinskog uzorkovanja. Ovo uključuje nanosatelite, mikrosatelite i posebno bespilotne letelice (eng. *Unmanned Aerial Vehicle*, UAV). Rastuća popularnost bespilotnih letelica i pomeranje granica njihove komercijalne upotrebe od strane velikih svetskih kompanija ukazuju na dalji rast njihove primene. Paralelno sa napretkom hardverskih platformi, očekuje se dalji napredak algoritama za procesiranje podataka, gde je jedna od ključnih oblasti inovacije razvoj algoritama paralelnog procesiranja podataka daljinskog uzorkovanja. Generalno, napredovanje oblasti daljinskog uzorkovanja i vezanih tehnologija već omogućuju usmereniji nadzor poljoprivrednih i prirodnih resurasa, brže i efikasnije odgovore na hitne slučajeve, proizvodnju preciznijih mapa, poboljšanu navigaciju i bolje geospatijalne informacije dostupne javnosti i profesionalcima na različitim poljima. Primena daljinskog uzorkovanja takođe dobija nove dimenzije svojim mogućnostima u realnom vremenu (ili vremenu bliskom realnom, eng. *near real-time*). Dakle, trendovi koji će uticati i produbljivati primenu daljinskog uzorkovanja u budućnosti su dalja minijaturizacija i integracija elektronike, razvoj prikupljanja podataka pomoću

bespilotnih letelica, rast računarske moći pomoću paralelizacije, *cloud computing*-a, itd., razvoj novih i moćnijih senzora, rast transmisijske moći aktivnih sistema, minijaturizacija optike, napredak tehnologije skladištenja podataka, razvoj malih satelita, napredak mobilnog računarstva, napredak tehnika za obradu *Big Data*, itd. (Khorram, Wiele, Koch, Nelson, & Potts, 2016)

Zbog svega gore navedenog, oblast kojim se bavi istraživanje opisano u ovoj doktorskoj disertaciji sadrži mnogo prostora za dalje unapređenje, pri čemu se očekuje dalja popularizacija daljinskog uzorkovanja kao naučne discipline. U tom smislu možemo izdvojiti još dva doprinosa i rezultata koji se postižu ovim istraživanjem, a to su definisanje metode za primene koje se javljaju upotrebom bespilotnih letelica, i to pre svega u poljoprivredi (ili konkretnije „preciznoj poljoprivredi“) u nadzoru useva sa specifičnostima koje takva primena nosi.

1.1 Polazne hipoteze

Osnovna hipoteza u ovoj doktorskoj disertaciji je:

1. Moguće je ustanoviti Bajesovu proceduru koja prilikom klasifikacije podataka daljinskog uzorkovanja minimizira mogućnost greške klasifikacije, kao i zonu osetljivosti klasifikacije koja maksimizira broj tačnih klasifikacija;

Ostale hipoteze su:

2. Klasifikacija na osnovu gornjih elemenata može povećati tačnost klasifikacije;
3. Primena takve metode može povećati nivo automatizacije obrade podataka (obrada podataka uz manji broj zadatih ulaza od strane analitičara);
4. Predloženi model se može primeniti na različite primene daljinskog uzorkovanja.

1.2 Metode istraživanja

Osnovni metod istraživanja u ovoj doktorskoj disertaciji je sakupljanje i proučavanje dostupne literature, njena analiza i sistematizacija, kao i izvođenje eksperimenata pomoću prikupljenih daljinski uzorkovanih podataka, u cilju pokazivanja opravdanost i korisnost definisanja novog metoda definisanja zone osetljivosti i klasifikacije podataka daljinskog uzorkovanja koja na njoj bazira. Rad će se zasnivati na primeni:

- Metoda za prikupljanje podataka pomoću daljinskog uzorkovanja,
- Metoda za predprocesuiranje digitalnih slika (spajanje, tagovanje, filtriranje, itd.),
- Metoda i tehnika eksplorativne analize podataka,
- Projektovanja i implementacije algoritamskih struktura
- Metoda za statističku analizu i obradu podataka,
- Multivarijacionih statističkih analiza (primarno diskriminaciona analiza),
- Metoda komparativne analize.

Disertacija će sadržati detaljnu analizu postojećih metoda klasifikaciju u daljinskom uzorkovanju, i njihovu komparaciju sa predloženim modelom u okviru studija slučajeva (primera). U okviru istraživanja, biće sprovedeno pretraživanje i analiza relevantne literature u oblasti od interesa. Doktorska disertacija će obuhvatati:

- Analizu koncepta klasifikacije podataka daljinskog uzorkovanja,
- Analizu postojećih metoda u oblasti klasifikacije podataka daljinskog uzorkovanja,
- Prikupljanje podataka u svrhe izvođenje eksperimenata (analizu i testiranje).
- Ocena tačnosti klasifikovanja pomoću predloženog rešenja.
- Analizu postojećih rešenja klasifikacije i poređenje sa predloženim rešenjem,
- Analizu osetljivosti predloženog rešenja.

2 GEOGRAFSKI INFORMACIONI SISTEMI (GIS)

Pojam Geografski Informacioni Sistem prvi put uvodi (Tomlinson, 1968) kada je kreirao prvi geografski informacioni sistem za Kanadsku poljoprivrednu agenciju, koji je za cilj imao da prikupi tačan inventar prirodnih resursa i potencijala države. Osnovno svojstvo GIS-a jeste da prihvata i skladišti sve vrste lokacijski-specifičnih (geografskih, georeferenciranih) informacija, odnosno bilo koje informacije koje mogu biti vezane za region, liniju ili tačku na mapi. Informacije vezane za zemljišne resurse su najčešće lokacijski-specifične informacije. Tako se sistem može najbolje opisati pomoću dve celine: kao skladište podataka i skup procedura i metoda za skladištenje i kasniju manipulaciju tim podacima (Tomlinson, 1968).

Termin „geografska informacija“ označava fenomen direktno ili indirektno vezan za lokaciju na Zemljinoj površi. U istom značenju se koristi termin prostorne informacije. One se mogu odnositi na relativno male površi kao što su zgrade, pojedinačna stabla drveća, do velikih površina i globalnih pojava poput vulkanskih pojaseva, klimatskih zona i ekonomskog razvoja kontinentalnih razmera. Osnovna karakteristika prostornih informacija je znanje o lokaciji određenog fenomena u odnosu na ostale objekte i pojave u okruženju, što je karakteristika koja doprinosi tome da se GIS razlikuje od ostalih informacionih sistema (Jovanović, Đurđev, Srdić, & Stankov, 2012).

Prve definicije geografskih informacionih sistema su polazile od informatičkih principa. Postojećim definicijama informacionih sistema dodata su teorijska saznanja o prostoru i vremenu. Tako (Fischer & Nijkamp, 1992) definišu geografski informacioni sistem kao kompjuterski informacioni sistem koji prikuplja, skladišti, analizira i prikazuje prostorne entitete i njihove atribute, za rešavanje kompleksnih istraživačkih, projektantskih i problema upravljanja. Najveći broj definicija GIS-a se mogu svrstati u grupu definicija koje su zasnovane na GIS-u kao sredstvu za rad (npr. kao skup sredstava za prikupljanje, memorisanje, rukovanje i analizu prostornih podataka), zatim u grupu definicija koje su zasnovane na bazama podataka (npr. sistem baza podataka u kojem je većina podataka prostorno indeksirana) i u definicije koje se zasnivaju na organizaciji (npr. kao skup funkcija i sistem za podršku odlučivanju sa svojom ulogom u organizaciji). Uzimajući prethodne definicije u obzir, izvedena definicija opšteg smisla bi mogla da glasi: Geografski informacioni sistem je organizovan skup računarskog hardvera, softvera, podataka, osoblja i mreža radi efikasnog prikupljanja, skladištenja, ažuriranja, rukovanja, analize, modelovanja, prenosa i prikaza svih oblika prostornih informacija (Jovanović, Đurđev, Srdić, & Stankov, 2012).

Iz pogleda korisnika GIS je sistem zasnovan na računaru koji pruža sledeće mogućnosti za upravljanje georefenciranim podacima: prikupljanje i priprema podataka, upravljanje podacima (uključujući skladištenje i održavanje), manipulacije nad podacima i analize i na kraju prezentovanje podataka. Sve ovo implicira da GIS

korisnici mogu očekivati podršku od strane sistema za unos (georeferenciranih) podataka, da ih analiziraju na razne načine i da proizvedu prezentacije podataka (uključujući mape i druge načine) (Huisman & de By, 2009). Geografski informacioni sistemi su u stanju da prikupe prostorno indeksirane podatke iz različitih izvora, menjajući ih u korisne formate, skladište podatke, povrate ih i manipulišu njima za različite analize, i na kraju generišu izlaze koje zahteva korisnik. Njihova snaga se zasniva na mogućnosti da manipulišu velikim, višeslojnim i heterogenim bazama podataka i da ispituju postojanje, lokaciju i osobine širokog spektra prostornih objekata na interaktivan način (Fischer & Nijkamp, 1992). Danas je GIS najčešće kombinacija računarskog hardvera koji koristi GIS softver sa grafičkim korisničkim interfejsom, dok pristupa podacima smeštenim lokalno, na centralnom serveru ili u *cloud*-u, a ono što ga takođe čini moćnim je što može uvezati prostorne vektorske podatke sa neprostornim bazama podataka uz laku mogućnost njihove vizuelizacije (Aber & Aber, 2017). Međutim, GIS je mnogo više od samo softvera, on se odnosi na sve aspekte upravljanja i korišćenja digitalnih geografskih podataka (Sutton, Dassau, & Sutton, 2009).

2.1 Podaci u GIS-u

Podaci su činjenice, mere i karakteristike o nečemu od interesa a prostorni podaci (eng. *spatial data*) se odnose na geografske objekte realnog sveta od značaja, poput ulica, zgrada, jezera, zemalja, sa svojim respektivnim lokacijama. Pored lokacije, svaki od ovih objekata takođe poseduje određene osobine od značaja, ili atribute, poput naziva, broj spratova, dubina ili populacija (Campbell & Shin, 2011). GIS softver vodi računa o oba tipa podataka, odnosno i o prostornim podacima i o atributima i dozvoljava povezivanje ova dva tipa podatka kako bi se dobile potrebne informacije i olakšale analize. Prostorni podaci opisuju geografske prostorne aspekte fenomena, dok atributski podaci opisuju kvalitete i karakteristike datog fenomena, a njihova integracija je jedna od osnovnih karakteristika GIS-a (Campbell & Shin, 2011; Sutton, Dassau, & Sutton, 2009).

Prostorni podaci ukazuju gde se stvari ili pojave nalaze, ili možda gde su se nalazili ili gde će se nalaziti. Brojne su potrebe za analizama koje uključuju pitanja vezana za geografski prostor, koji se definiše kao prisustvo pozicionih podataka koji su relativni u odnosu na Zemljinu površinu (Huisman & de By, 2009). Pozicioni podaci koji nemaju geografsku prirodu takođe postoje, što su podaci koji ukazuju na poziciju nečega ali ne u vezi sa Zemljinom površinom (npr. pozicija organa u ljudskom telu ili delova u automobilu), ali takvi pozicioni podaci nam nisu od značaja za dalje razmatranje u ovoj disertaciji.

Pod podacima ovde ćemo smatrati reprezentacije kojima se može manipulisati pomoću računara. Konkretnije, prostornim podacima ćemo smatrati podatke koji imaju pozicione vrednosti, poput (x, y) koordinata. Za dalje refinisanje pojma, često se koristi

naziv geoprostorni (eng. *geospatial*) podaci, što se odnosi na prostorne podatke koji su georeferencirani, a nadalje pojmove prostorni i georeferencirani podaci možemo koristiti kao sinonime. Pod informacijom mislimo na podatke koji su interpretirani od strane čoveka, dok je pojam geoinformacija specifičan tip informacije koji nastaje interpretacijom prostornih podataka (Huisman & de By, 2009).

S obzirom da je svrha ovakvih (i drugih) informacija smanjenje neizvesnosti u procesu donošenja odluka, sve greške i neizvesnosti u prostornim informacijama mogu imati praktične, finansijske, pa čak i pravne implikacije za korisnika. Iz tog razloga je ključno za one koji su uključeni u prikupljanje i obradu prostornih podataka, da su u stanju proceniti kvalitet osnovnih podataka i izvedenih informacija (Huisman & de By, 2009). Ključne komponente kvaliteta prostornih podataka su poziciona preciznost (horinzontalna i vertikalna), vremenska preciznost (da su podaci ažurni), preciznost atributa (npr. u označavanju osobina ili karakteristika ili u njihovoj klasifikaciji), poreklo (istorija podataka i njihovih izvora), kompletnost (da li skup podataka predstavlja sve karakteristike stvarnosti) i logička konzistentnost (da su podaci logički struktuirani) (Huisman & de By, 2009).

S obzirom na karakter podatka, strukturu zapisa, odnosno na njihovu organizaciju, geografski informacioni sistemi integrišu rasterske, vektorske i alfa-numeričke podatke i digitalne modele visina, a multimedija u GIS-u je omogućila korišćenje i skladištenje čak i novih tipova podataka kao što su ton (zvuk), animacije i video zapisi (Jovanović, Đurđev, Srdić, & Stankov, 2012). Svojstva podataka u GIS-u još mogu biti kategorizovana kao diskretna ili kontinualna. Diskretna su ona koja su dobro definisana, koja je lako locirati, meriti, brojati ili čije granice su jasno određene (npr. zgrade, putevi, saobraćajni znaci, itd.) dok su kontinualna ona koja su manje definisana i postoje „širom prostora“ (npr. temperatura ili elevacija), tj. koja se postepeno menjaju preko relativno velikih površina (Campbell & Shin, 2011). Pored navedenih osobina GIS podataka, (Kapetsky & Aguilar-Manjarrez, 2007) još govore o podacima sa aspekta njihove raspoloživosti i kvaliteta. Tako navode da je važna distinkcija među podacima koji su javno dostupni i mogu se besplatno skinuti sa Interneta i komercijalno pripremljenim podacima koji moraju biti kupljeni. Takođe, uvode i razliku između podataka koji imaju globalni obuhvat (pokrivenost) i nacionalnih podataka (Kapetsky & Aguilar-Manjarrez, 2007).

Za potpuno razumevanje podataka koji se čuvaju u GIS-u, krenućemo od modelovanja istih. Modelovanje je proces kreiranja apstrakcije prikaza realnog sveta kako bi se nekim njegovim delovima moglo lakše manipulisati. U GIS okruženju, najpoznatiji model je mapa, koja je minijaturna reprezentacija dela realnog sveta, koja je uvek grafička reprezentacija na određenom nivou detalja (Huisman & de By, 2009). Polje, ili geografsko polje, je geografski fenomen za koji, za svaku tačku u posmatranoj oblasti, posmatrana vrednost može biti određena. Ova polja mogu biti kontinualna (temperatura vazduha, pritisak, elevacija) ili diskretna po svojoj prirodi (upotrebe zemljišta i klasifikacija tla), u kom slučaju svakoj lokaciji razmatrane oblasti se može

dodeliti jedinstvena klasa upotrebe zemljišta ili klasa tla (Huisman & de By, 2009). S obzirom na diferencijaciju između kontinualnih i diskretnih polja, postoje i različiti tipovi vrednosti podataka koje možemo koristiti za prezentaciju različitih „fenomena“. Ovde je takođe važno napomenuti da neki od tipova podataka ograničavaju vrste analiza koje nad njima možemo sprovesti. Čak i fenomeni sa kontinualnim i/ili beskonačnim skupom karakteristika moraju biti predstavljeni ograničenim sredstvima (kao što je računarska memorija) za manipulaciju na računaru, a svaka takva konačna reprezentacija je podložna greškama interpretacije. Imajući prethodno u vidu, polja se obično predstavljaju pristupom mozaika, a objekti vektorskim pristupom (Huisman & de By, 2009).

Možemo zaključiti da se digitalni geoprostorni podaci čuvaju u dve osnovne forme: rasterskoj i vektorskoj (Aber & Aber, 2017; Jovanović, Đurđev, Srdić, & Stankov, 2012; Huisman & de By, 2009). U narednim poglavljima daćemo više detalja o obe forme, ali za ovu disertaciju je od naročitog značaja rastersko predstavljanje podataka, te ćemo se time pozabaviti u značajno više detalja.

Vredi još napomenuti da pored ove dve osnovne forme čuvanja podataka u GIS-u, (Jovanović, Đurđev, Srdić, & Stankov, 2012) navode kao posebne forme i atributske podatke, kojima se izražavaju negeometrijske karakteristike entiteta, najčešće u alfa-numeričkom obliku i koji mogu biti tabelarni podaci kao deo analitičkih procedura ili izlaza iz GIS-a, i digitalne modele visina (eng. *Digital Elevation Model*), koji su organizovani skup podataka o visinama terena zapisan u digitalnom obliku, gde za svaku tačku model sadrži podatke o položaju tačke u prostoru, na površi, i podatke o njenoj visini (Jovanović, Đurđev, Srdić, & Stankov, 2012).

2.1.1 Vektorski podaci

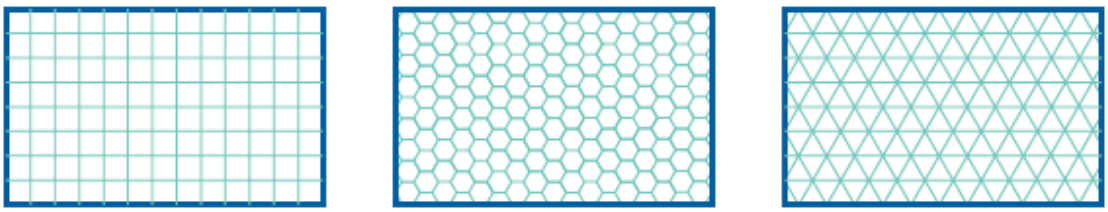
Vektorski podaci su poseban tip podataka čiju strukturu čine osnovne geometrijske primitive: tačka, linija i poligon (region). Položaj primitiva definiše se koordinatama, odnosno vektorom položaja, a u računarskoj memoriji se čuvaju kao serija *x*,*y* koordinatnih parova (Sutton, Dassau, & Sutton, 2009; Jovanović, Đurđev, Srdić, & Stankov, 2012). Drugim rečima, oblik vektorskog objekta je predstavljen pomoću geometrijskih oblika koje čine jedno ili više međusobno povezanih temena, čija je poziciju u prostoru predstavljena pomoću *x*, *y* i opciono *z* ose. Geometrije koje imaju temena opisana i sa *z* osom se često nazivaju i 2.5D, budući da su im opisana ili visina ili dubina svakog temena, ali ne oboje (Sutton, Dassau, & Sutton, 2009).

Osnovni element vektorskog sadržaja je tačka. Položaj tačke definisan je njenim koordinatama (linijskim, uglovnim ili linijsko-uglovnim veličinama), odnosno njenim vektorom položaja. Tačkom su prikazani entiteti veoma malih dimenzija koji se, zbog razmere prikaza, ne mogu prikazati pomoću linije ili poligona. Izvedeni elementi vektorskog sadržaja su linija i poligon. Linija je organizovani skup povezanih tačaka i predstavlja jednodimenzionalnu geometrijsku primitivu. Linijom su prikazani entiteti

koji se zbog malih dimenzija ne mogu prikazati pomoću poligona. Poligon je organizovani skup linija kojima se definiše neka oblast, a kod kojih se prva i poslednja tačka poklapaju. Poligon predstavlja dvodimenzionalnu geometrijsku primitivu (Jovanović, Đurđev, Srdić, & Stankov, 2012). Vektorski podaci uključuju dva koncepta, geometriju i atribute. Geometrija vektorskog svojstva ili objekta opisuje njegov oblik i poziciju, dok atributi opisuju njegova svojstva (npr. boju, veličinu, starost, itd.) (Sutton, Dassau, & Sutton, 2009).

2.1.2 Rasterski podaci

Ranije pomenuti pristup mozaika za predstavljanje geografskih fenomena, odnosno polja, podrazumeva deljenje prostora u međusobno isključive ćelije koje zajedno čine kompletnu posmatranu oblast. Svakoj ćeliji je dodeljena određena (tematska) vrednost koja karakteriše taj deo oblasti. Postoje neregularni načini mozaika, koji su kompleksniji od regularnih ali su više adaptivni i mogu doneti uštede u utrošenoj memoriji, ali ćemo se dalje u ovoj disertaciji baviti samo regularnim mozaikom, i to sa kvadradnim ćelijama (pored toga ćelije mogu biti npr. heksaugaone, trouglaste, itd.) (Slika 2.1.1) (Huisman & de By, 2009).



Slika 2.1.1: Tri najčešće korišćena mozaika: kvadratne, heksaugaone i trougaone ćelije.

Raster je skup ravnomerno rapoređenih (i susednih) ćelija sa dodeljenim vrednostima, koje predstavljaju vrednosti ćelija (a ne tačaka). To znači da se vrednost ćelije smatra važećim za sve lokacije unutar te ćelije. Veličina oblasti koju predstavlja svaka rasterska ćelija se naziva rasterska rezolucija (Huisman & de By, 2009).

Dakle, rasterski podaci se čuvaju u formi koordinatne mreže, odnosno matrice. Primera radi, postoje mnogi izvori podataka, kao što su brojni sateliti koji kruže oko Zemlje, koji beleže rasterske podatke koji se kasnije mogu naći u GIS-u (Sutton, Dassau, & Sutton, 2009). Rasterski podaci se koriste u GIS-u kada želimo da prikažemo informacije koje su kontinualne kroz određenu regiju i ne mogu se lako podeliti u vektorske objekte. Naime, neki objekti pejzaža (predela slike) se veoma teško mogu predstaviti pomoću vektorskih objekata. Na primer, prikazani pašnjaci imaju velike varijacije u boji ili gustini prekrivača. Bilo bi jednostavno napraviti jedan vektoriski poligon oko svake oblasti pašnjaka, ali veliki broj informacija o pašnjaku bi bio izgubljen u procesu uprošćenja objekata u jedan poligon (Sutton, Dassau, & Sutton, 2009).

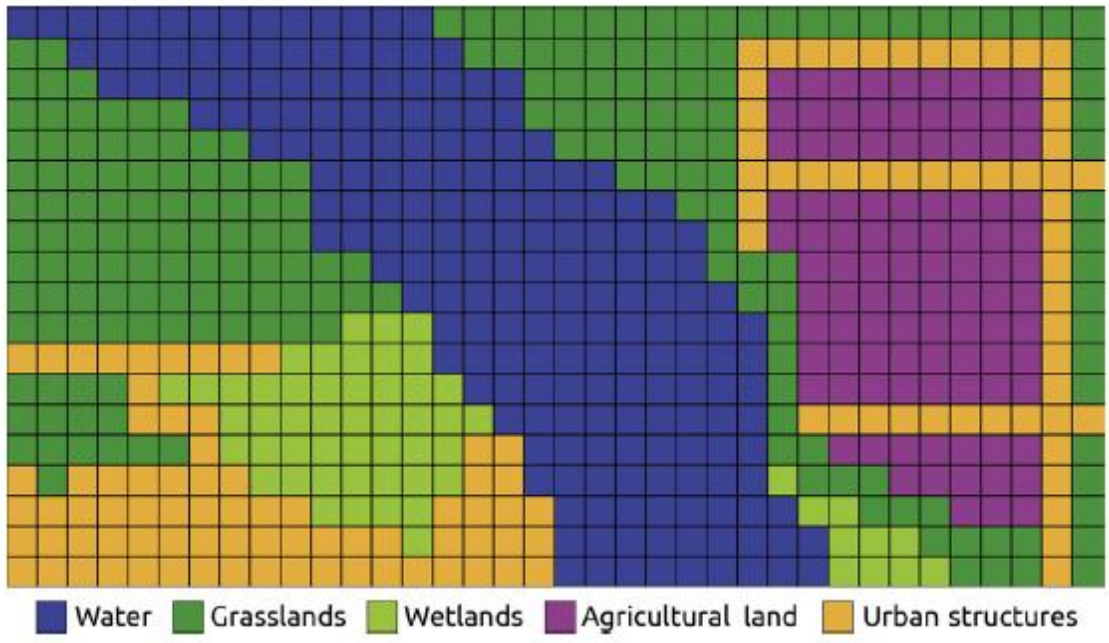
Osnovne karakteristike koje karakterišu rasterske slike su rezolucija, dimenzija, broj boja (dubina) i format zapisa slike. Rezolucija slike predstavlja veličinu piksela izraženu dimenzijom piksela u dužinskim jedinicama ili brojem piksela po jedinici dužne mere. Prostorna rezolucija rasterskog skupa podataka je mera preciznosti ili detalja prikazanih informacija (Campbell & Shin, 2011). Dimenzija rasterske slike definiše se širinom i visinom, odnosno brojem kolona i redova u slici. Broj boja rasterske slike je jedna od veoma važnih karakteristika slike. Često, u praksi se ova karakteristika rasterske slike naziva i dubina slike. U zavisnosti od broja boja na rasterskoj slici u praksi najčešće se sreću: 1-bitne (crno-bele slike), 8-bitne (slike sa 256 boja, odnosno 256 nijansi jedne boje) i 24-bitne (slike sa 1,67 miliona boja) i 32-bitne slike. S obzirom da se radi o podacima koji podrazumevaju digitalni zapis, još jedna veoma bitna karakteristika ovih podataka je format zapisa. Najčešći formati rasterskih slika su bmp, tif, gif, jpg, pcx i drugi. Rasterske slike su podaci u digitalnom obliku zasnovane na matrici. Matrica je predstavljena određenim brojem kolona i redova (Levin, 1999). U preseku jednog reda i jedne kolone je osnovni element matrice, odnosno ćelija matrice, koja se u slučaju rasterske slike naziva piksel. S obzirom da matrica podrazumeva uređen redosled (raspored) kolona i redova, adresa, položaj piksela je definisan brojem reda i brojem kolone u čijem se preseku nalazi piksel, iz čega se može zaključiti da matrica poseduje osnovne karakteristike dvodimenzionalnog koordinatnog sistema. Još jedna veoma bitna karakteristika rasterske slike je da svaki piksel ima svoju tačno definisanu vrednost atributa. Atribut piksela može predstavljati intenzitet reflektovane boje, intenzitet reflektovanog signala, visinu ili intenzitet zagađenosti zemljišta i sl. Na osnovu toga, rastersku sliku možemo definisati i kao trodimenzionalnu ili preciznije 2D+1 matricu. Dve dimenzije se odnose na položaj (koordinatu) piksela u matrici, dok treća izražava vrednost atributa nekog fenomena koji je predmet prikaza. (Jovanović, Đurđev, Srdić, & Stankov, 2012)

Rasterske slike mogu biti *georeferencirane* ili *negeoreferencirane*. Georeferenciranje je proces definisanja gde su tačno na Zemljnoj površini slika i rasterski skup podataka kreirani, pri čemu je ta poziciona informacija sačuvana zajedno sa digitalnom verzijom (vazdušne) fotografije (Sutton, Dassau, & Sutton, 2009). Drugim rečima, g*eoreferencirane* slike su rasterske slike koje, osim već navedenih, imaju još jednu veoma važnu karakteristiku, a to je prostorna definisanost. Naime, koristeći osobine rasterske slike kao matrice, a primenjujući različite metode, moguće je za svaki piksel rasterske slike pridružiti podatak o njegovom položaju u prostoru. Taj proces pridruživanja prostorne karakteristike nekoj rasterskoj slici naziva se georeferenciranje. U tom slučaju se slikovnoj koordinati piksela pridružuje i prostorna koordinata piksela. Osnovne komponente georeferenciranih slika su: sadržaj slike - mape (bit-mape) i set podataka o prostornoj, položajnoj definisanosti (parametri georeferenciranja). Fizička organizacija ovih komponenti je različita kod različitih formata podataka. Neki od dozvoljavaju zapis ovih podataka u jednom fajlu (RLE, RGB, GeoTIFF), dok neki formati zahtevaju zapis u više fajlova. U georeferencirane slike spadaju karte, satelitski snimci, a u nekim slučajevima i digitalni model visina i sl.

(Jovanović, Đurđev, Srdić, & Stankov, 2012). Ovde je od značaja i pojam projekcije koja se koristi pri kreiranju georeferenciranog rasterskog prikaza podataka. Naime, „ravne“ mape imaju samo dve dimenzije, širinu i dužinu, pa je transformacija trodimenzionalne Zemlje na dvodimenzionalnu mapu predmet projekcije mape, odnosno transformacije koordinata. Projekcija mape je matematički opisana tehnika kako predstaviti Zemljinu zakrivljenu površinu na ravnoj mapi (Huisman & de By, 2009).

Negeoreferencirane slike su posebna vrsta rasterskih slika koje se ne mogu, ili ih nema smisla, georeferencirati. Najčešće su rezultat procesa skeniranja različitih dokumenata ili direktnog snimanja optičkim digitalnim kamerama. U ovu vrstu rasterskih slika ubrajamo fotografije, skenirana dokumenta, obrasce, opise i slično (Jovanović, Đurđev, Srdić, & Stankov, 2012).

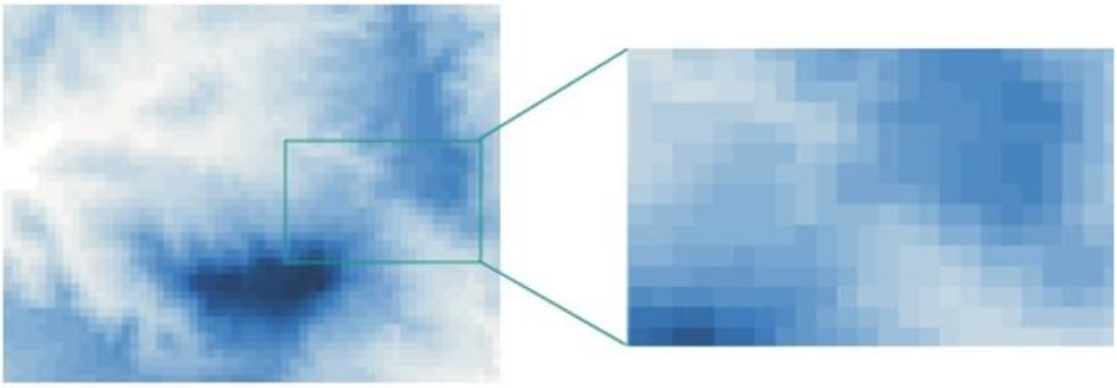
Rasterski podaci mogu biti fotografije, ali mogu takođe prikazivati nefotografske informacije. Slika 2.1.2 prikazuje primer nefotografskog prikaza realnog sveta gde je Zemljina površina klasifikovana u kategorije zemljišnog prekrivača. Svaka ćelija ima jedinstvenu vrednost koja predstavlja šta je na zemlji u datoj koordinatnoj mreži i ne postoji prazne ćelije na mreži (Aber & Aber, 2017). Ova reprezentacija će biti značajna u nastavku ove disertacije.



Slika 2.1.2: Rasterski prikaz sveta. Pravougaone linije ne oslikavaju realnost, već su posledica strukture rasterskog formata (Aber & Aber, 2017).

Na slici 2.1.3 je ilustrovano kako raster predstavlja kontinualno polje poput elevacije zemljišta. Različite nijanse plave boje ukazuju na različitu elevaciju tla (tamnija plava ukazuje na veću visinu). Izbor boja je samo estetske prirode – umesto

toga mogli bismo prikazati stvarnu vrednost u svakoj od ćelija, ali bi to bili prilino nepregledno.



Slika 2.1.3: Rasterski prikaz elevacije tla. Stvarne visine su predstavljene nijansama plave boje. Na desnoj strani je prikazan uveličan segment slike na levoj strani (Huisman & de By, 2009).

Dakle, podaci daljinskog uzorkovanja se u konačnici mogu zabeležiti u vidu matrice (rastera). Kolone se obično nazivaju uzorcima (eng. *samples*) a redovi linijama (Levin, 1999). Postoji više načina za beleženje pojedinačnih vrednosti svakog od piksela, kao što su RGB (odnosno crvena, zelena i plava boja, eng. *Red*, *Green*, *Blue*), sa rasponom vrednosti između 0 i 255, i HSL (odnosno nijansa, odnosno osenčenost boje, zatim zasićenje, koja opisuje čistoću boje, i osvetljenost, eng. *Hue*, *Saturation*, *Lightness*, respektivno) (Canada Centre for Remote Sensing, 2007; Levin, 1999). Kada je slika prikazana na ekranu, crvena, zelena i plava (RGB) informacija je kombinovana kako bi slika bila prikazana na način da je ljudsko oko adekvatno interpretira. Ipak, ova RGB informacija je zapamćena u posebnom opsegu boja (Sutton, Dassau, & Sutton, 2009). S obzirom na to da je imati slike koje sadrže višestruke opsege svetlosti veoma korisno u GIS-u, rasterski podaci često postoje kao slike multi-opsega. Svaki opseg u slici je kao poseban sloj a GIS će kombinovati ova tri opsega i prikazati ih kako bi ih ljudsko oko videlo. Broj opsega u rasterskoj slici se naziva *spektralna rezolucija* (Sutton, Dassau, & Sutton, 2009).

2.2 Prikupljanje prostornih podataka

Prema osnovnoj definiciji GIS uvek sadrži module za unos podataka, pakovanje i kreiranje baza podataka, analizu i prikazivnje prostornih podataka. Ulazne informacije GIS pribavlja iz geografskih karata, satelitskih i avionskih snimaka, direktnim prikupljanjem na terenu ili su to opisni podaci vezani za posmatranu lokaciju (Jovanović, Đurđev, Srdić, & Stankov, 2012).

Geografski podaci se dobijaju iz različitih izvora. Mogu biti prikupljeni „od nule“, pomoću tehnike prikupljanja prostornih podataka, kao što je sakupljanje podataka *direktno* sa terena (npr. uz pomoć GPS-a) i daljinskom detekcijom, ili *indirektno*, korišćenjem postojećih prostornih podataka skupljenih od drugih, kao što su digitalizacija i skeniranje karata i unošenje postojećih (digitalnih) baza podataka i pisanih zapisa (Huisman & de By, 2009; Jovanović, Đurđev, Srdić, & Stankov, 2012). Sistem za globalno pozicioniranje (eng. *Global Positioning System*, GPS) je originalno postavljen kao kontrolni sistem Ministarstva odbrane SAD, ali danas, sem vojne, ima veliki broj civilnih primena. GPS obezbeđuje specijalno kodirane satelitske signale koji se obrađuju i šalju u prijemnik na kome se očitavaju podaci o preciznoj poziciji, brzini i tačnom vremenu (Jovanović, Đurđev, Srdić, & Stankov, 2012). Podaci prikupljeni direktno iz okruženja se nazivaju *primarnim podacima*, a podaci sakupljeni indirektno (digitalizacijom postojećih mapa, kupovinom postojećih podataka komercijalnih kompanija, i sl.) se nazivaju *sekundarnim podacima* (Huisman & de By, 2009; Campbell & Shin, 2011).

Svi podaci, bilo iz primarnih ili sekundarnih izvora, imaju tri dimenzije: vremensku, tematsku i prostornu. Tokom rada sa geografskim podacima potrebno je identifikovati ove tri dimenzije za svaki podatak. Vremenska dimenzija nam daje određenost – kada su podaci prikupljani, a tematska opisuje pojavu stvarnog sveta na koju se podatak odnosi. U GIS-u tematski podaci se često nazivaju i neprostorni ili atributski podaci. Prostorna dimenzija podataka može biti predstavljena kao vrednost, niz karaktera ili simbola koji šalju korisniku informaciju o lokaciji pojave ili objekta koji se posmatra. Svim prostornim podacima upotrebljenim u GIS-u moraju se dati matematičke reference a jedan od najčešćih primera matematičke prostorne reference je koordinatni sistem (koordinate *x*, *y*). (Jovanović, Đurđev, Srdić, & Stankov, 2012).

U predmetnom kontekstu slika se odnosi na sirove date koje proizvodi elektronski senzor, koji nisu ilustracija, već niz digitalnih brojeva vezanih za neko svojstvo prikazanog objekta, kao što je količina reflektovane svetlosti. Za sliku, još uvek nije izvršena nikakva interpretacija vrednosti refleksije kao tematskih ili geografskih karakteristika. Kada se vrednosti refleksije „prevedu“ u varijable koje imaju tematsko značenje, nazivamo ih rasterom (Huisman & de By, 2009). Rasterski podaci se mogu dobiti na različite načine, od kojih su najčešći pomoću fotografija snimljenih iz vazduha i satelitskih snimaka. Fotografije snimljene iz vazduha se dobijaju tako što avion, ili, što je sve popularnije, bespilotna letelica, nadleće površinu sa pričvršćenom kamerom (fotoaparatom). Fotografije se potom uvoze u računar i georeferenciraju. Satelitski snimci se dobijaju pomoću satelita koji kruže Zemljinom orbitom i usmeravaju digitalne kamere prema površinama od interesa . Slika se potom šalje nazad na zemlju pomoću radio signala posebnim stanicama za prijem. Proces prikupljanja rasterskih podataka iz letelice ili satelita se naziva *daljinska detekcija* ili *daljinsko uzorkovanje* (eng. *Remote Sensing*) (Sutton, Dassau, & Sutton, 2009).

Daljinsko detekcija (*uzorkovanje*) je metod prikupljanja i interpretacije informacija o udaljenim objektima bez fizičkog dodira s objektom (Lapaine & Frančula, 2000/2001; Manson, Bonsal, Kernik, & Lambin, 2015). Slike prikupljenje daljinskim uzorkovanjem su veoma važan izvor podataka za GIS (Huisman & de By, 2009). Avioni, sateliti i svemirske sonde su uobičajene platforme za opažanja u daljinskim istraživanjima. Daljinska detekcija koristi metode koje upotrebljavaju elektromagnetnu energiju kao sredstvo za otkrivanje i merenje značaja objekata. Ona obuhvata upotrebu različitih vrsta snimaka: fotografskih, termalnih, radarskih itd. Dva značajna uža područja daljinske detekcije su: teledetekcija i fotogrametrija. *Teledetekcijom* se naziva daljinsko istraživanje u užem smislu, tj. prikupljanje informacija o Zemljinoj površini, uz pomoć uređaja smeštenim u satelitima i interpretaciju tako dobijenih informacija (Lapaine & Frančula, 2000/2001). *Fotogrametrija* je nauka i tehnika merenja pomoću koje se iz fotografskih snimaka izvodi oblik, veličina i položaj snimljenog predmeta i konvertuju takve mere u količine od značaja na terenu (Lapaine & Frančula, 2000/2001; Tempfli, Kerle, Huurneman, & Janssen, 2009).

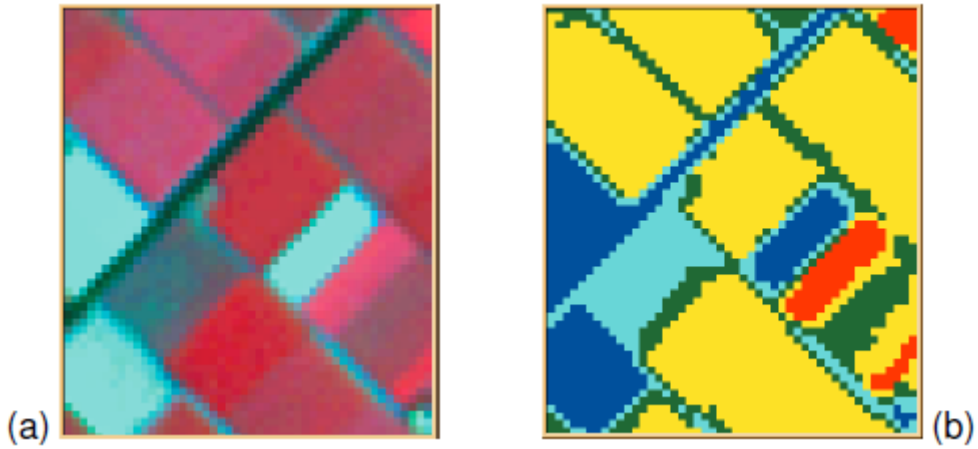
Daljinsko uzorkovanje i GIS analize su komplementarni pristupi koje naučnici koriste da prikupe, obrade i analiziraju prostorne podatke. Ove dve tehnologije se široko koriste u različitim društvenim i prirodnim naukama za procenu prirodnih resursa, praćenje promena sredine i razumevanje različitih sociooloških fenomena koji se odnose na životnu sredinu i ekološka pitanja (Manson, Bonsal, Kernik, & Lambin, 2015). GIS olakšava interpretaciju podataka daljinskog uzorkovanja vezujući biofizičke informacije izmerene udaljenim senzorima za mape atributa prirodnih i kulturoloških pejzaža (npr. vrsta zemljišta, topografija, itd.), za terenska merenja (npr. procena prinosa useva ili kvalitet vode) i socioekonomske informacije izvedene kroz istraživanja domaćinstava i druge izvore (npr. pokazatelji uslova kvaliteta života, prirast populacije, korišćenje poljoprivrednih sirovina, itd.). GIS to postiže pomoću upraivanja podataka iz slika sa ostalim vrstama prostornih podataka, pomoću njihove lokacije. Daljinsko uzorkovanje i GIS su važni pristupi za istraživanja usmerena ka razumevanju dinamike upotrebe zemljišta, njenih pokretačkih sila, i uticaja na društvo (Manson, Bonsal, Kernik, & Lambin, 2015).

Satelitski snimci postaju sve popularniji kako se sateliti opremljeni tehnološki naprednim senzorima sve više šalju u svemir, od strane vladinih agencija i privatnih kompanija širom sveta. Sateliti se koriste u vojne i u civilne svrhe posmatranja zemljine površine, komunikacije, navigacije, vremenske prognoze, istraživanja itd. Trenutno postoji preko 3,000 satelita u svemiru, od kojih su većinu lansirali SAD i Rusija. Sateliti i njihovi senzori se nazivaju *aktivnim* ukoliko detektuju refleksione odgovore objekata ozračenih od veštački generisanih izvora energije, i *pasivnim* ukoliko detektuju reflektovano ili emitovano elektromagnetno zračenje iz prirodnih izvora (Campbell & Shin, 2011).

Fotografije iz vazduha, poput satelitskih snimaka, predstavljaju ogroman izvor informacija za upotrebu u GIS-u. Oprema koja je potrebna za snimanje fotografija iz

vazduha uključuje avione, helikoptere, balone, rakete (Campbell & Shin, 2011) a u poslednje vreme sve su više popularne bespilotne letelice ili popularni „dronovi“ (eng. *Unmanned Aerial Vehicle*, UAV). Iako fotografije iz vazduha uglavnom uključuju slike prikupljene u vidljivom spektru, mogu uključivati i senzore koji prikupljaju podatke iz opsega nevidiljivog svetlosnog spektra (npr. ultraljubičaste, infracrvene, blizu-infracrvene). Slično, vazdušne fotografije mogu biti aktivne ili pasivne i mogu biti uzete pod vertikalnim ili kosim uglom (Campbell & Shin, 2011; Tempfli, Kerle, Huurneman, & Janssen, 2009). Iako deluje jednostavno, slike prikupljene iz vazduha nisu toliko jednostavne kao prosto slikanje iz vazduha. Zemljina površina nije ravna i sva sočiva unose distorziju u prikupljene slike. *Ortofoto* je vertikalna fotogrfija iz vazduha koja je geometrijski „ispravljena“ da se uklone ove distorzije, tj. da se ukloni zakrivljenost i greške nastale kao posledica nepravilnog oblika terena (Campbell & Shin, 2011; Aber & Aber, 2017).

Neprocesirane slike sadrže mnogo piksela, pri čemu svaki piksel nosi vrednost refleksije oblasti na koju se odnosi. Postoje razne tehnike obrade digitalnih slika u klasifikovane slike koje se mogu sačuvati u GIS-u kao rasterski podaci. Klasifikacija slika služi da se svaki piksel pridruži jednoj iz skupa konačnih klasa, čime se dobija interpretacija sadržaja slike (Huisman & de By, 2009). Prepoznate klase mogu npr. biti vrste useva, kao što je dato u primeru na slici 2.2.1. Slika prikazuje neobrađenu sliku (a), kao i klasifikovanu verziju slike (b).



Slika 2.2.1: Neprocesirana digitalna slika (a) i klasifikovani rasterski prikaz (b) poljoprivrednog područja (Huisman & de By, 2009).

2.3 Analize prostornih podataka

Geografska informaciona analiza je novi koncept rada generisan kroz različte kontekste a njeno nastajanje je vezano za znatno stariji termin prostorna analiza. Prema (Jovanović, Đurđev, Srdić, & Stankov, 2012) četiri prepoznatljive, bliske oblasti koje se susreću u literaturi imenovane su kao:

- *Rukovanje prostornim podacima*: upućuje na GIS kontekst koji označava upotrebu tehničko-tehnološkog sklopa za rad sa prostornim podacima:
- *Analiza prostornih podataka* je deskriptivna i istraživačka naučna oblast. To je prvi korak u svim prostornim analizama koje uključuju velike i kompleksne skupove podataka.
- *Prostorna statistička analiza* koristi statističke metode za ispitivanje prostornih podataka radi upotrebe ili odbacivanja u kreiranju statističkog modela. Već nekoliko decenija kao posebna naučna disciplina razvija se geostatistika, a njena primena je spregnuta sa pojavom softverskih paketa namenjenih obradi podataka. Približavanje geostatistike i GIS-a je u polju prostornih analiza. Savremene prostorne analize koriste geostatistiku za proučavanje prostorne distribucije, putem variograma, kao i za predviđanje prostornih atributa na različitim lokacijama na osnovu uzorkovanih vrednosti primenom Kriging interpolacionih metoda.
- *Prostorno modeliranje* obuhvata konstrukciju modela za predviđanje prostornih ishoda. U društvenoj geografiji model se koristi za predviđanje kretanja ljudi i dobara između različitih mesta ili optimizaciju kapaciteta s obzirom na činjenicu da model treba da simulira dinamiku prirodnih procesa, odnosno stanja u životnoj sredini.

Jedna od primarnih uloga GIS-a je da pruži podršku u odlučivanju analizom prostornih podataka (Huisman & de By, 2009; Longley, Goodchild, Maguire, & Rhind, 2005). Prostornom (geografskom) analizom je omogućena transformacija, manipulacija i metodski postupci koji mogu biti primenjeni nad geografskim podacima dodajući im novu vrednost koja doprinosi lakšem donošenju odluka i stvaranju uvida u relevantne prostorne odnose i anomalije koje nisu mogle biti uočene na prvi pogled. Drugim rečima, prostorna analiza je proces kojim se sirovi podaci pretvaraju u korisne informacije, u cilju istraživanja ili u postupku donošenja odluka (Longley, Goodchild, Maguire, & Rhind, 2005). Mnogi modeli koji su razvijeni korišćenjem GIS-a su u suštini statistički i predstavljaju prostorni ekvivalent deskriptivne statistike (Longley, Goodchild, Maguire, & Rhind, 2005).

Postoji više načina klasifikacije analitičkih funkcija GIS-a. Jedan način klasifikacije dat u (Huisman & de By, 2009) je:

1. *Funkcije klasifikacije, pronalaženja i merenja*: koje se vrše nad jednim slojem (eng. *layer*) podataka (vektorskom ili rasterskom), često pomoću pridruženih atributa.
2. *Funkcije preklapanja*: omogućuju kombinovanje dva (ili više) sloja prostornih podataka poredeći ih poziciju po poziciju, i tretirajući oblasti preklapanja i nepreklapanja na različit način.
3. *Funkcije susednosti*: dok preklapanje kombinuje osobine iste lokacije, funkcije susednosti evaluiraju karakteristike oblasti koja okružuje lokaciju određenog svojstva ili objekta.
4. *Funkcije povezivanja*: rade na principu mreža, uključujući mreže puteva, vodnih tokova, komunikacionih linija mobilne telefonije i sl. Ove mreže predstavljaju prostornu povezanost između objekata.

U (Jovanović, Đurđev, Srdić, & Stankov, 2012) takođe se navodi klasifikacija analitičkih funkcija GIS-a u dve velike grupe: analitičke funkcije na geografskim podacima (u koje spadaju tačka u poligonu, rastojanje dve tačke, izdvajanje, klasifikacija, operacije susedstva, tampon-buffer zone, preklapanje) i analitičke funkcija na atributskim podacima. Najčešći izlazni rezultat GIS-a je karta. U najvećem broju slučajeva to će biti tematska karta koja će ilustrovati prostorne varijacije ili šemu određene promenljive. U ostalim slučajevima, GIS može biti upotrebljen za proizvodnju topografskih karata (Jovanović, Đurđev, Srdić, & Stankov, 2012).

Klasifikacija rasterskih podataka će biti od posebnog značaja za nas u nastavku ove disertacije, budući da se on suštinski bavi upravo metodologijom klasifikacije podataka prikupljenih daljinskim uzorkovanjem.

2.4 Primena GIS-a i analize prostornih podataka

Primena GIS-a kontinuirano raste. Njegovo korišćenje omogućilo je inženjerima da projektuju saobraćajne sisteme; naučnicima da istražuju promene u životnoj sredini; vladama da planiraju korišćenje zemljišta; vatrogasci i policajci mogu da planiraju rute svojih hitnih intervencija; kompanije pomoću GIS-a pronalaze i analiziraju tržišta, optimalno planiraju svoje servisne usluge, itd. (Jovanović, Đurđev, Srdić, & Stankov, 2012). Pet osnovnih i najčešćih vidova upotrebe GIS-a su: kartiranje, merenje, nadgledanje, modelovanje i menadžment (Longley, Goodchild, Maguire, & Rhind, 2005). Tabela 2.4.1 daje jedan pregled najvažnijih oblasti i načina primene GIS-a (Jovanović, Đurđev, Srdić, & Stankov, 2012).

Tabela 2.4.1: Oblasti i primena GIS-a (Jovanović, Đurđev, Srdić, & Stankov, 2012).

Agronomija	Nadgledanje i upravljanje od nivoa farmi do nacionalnog nivoa
Akvakultura	Pogodnost, zoniranje, planiranje razvoja, popis i nadzor (Kapetsky & Aguilar-Manjarrez, 2007)
Arheologija	Opis nalazišta i procena arheoloških scenarija
Društvene nauke	Analize demografskih kretanja i razvoja
Epidemiologija i zdravstvo	Lokacija zaraznih bolesti u odnosu na faktore sredine
Hitne usluge	Optimizacija vatrogasnih, policijskih i ambulantnih koridora; bolje sagledavanje zločina i njihovih lokacija
Komunalne službe	Lokacije, upravljanje i planiranje vodovoda, kanalizacije, gasovoda, električnih i kablovskih servisa
Marketing	Položaji i ciljne grupe; optimizacija dostavljanja robe
Navigacija	Vazdušna, morska i kopnena
Nepokretnosti	Zakonski aspekti katastra, vrednosti imovine u odnosu na lokaciju, osiguranje
Obrazovanje	Primena kao osnovnog ili pomoćnog sredstva u izvođenju nastave svih nivoa obrazovanja
Predmer radova i troškova	Useci i nasipi, računanje količine materijala
Putevi i železnice	Planiranje i menadžment
Regionalno i lokalno planiranje	Izrada planova, troškovi, održavanje, menadžment
Šumarstvo	Menadžment, planiranje i optimizacija seče i ponovnog sađenja
Telekomunikacije	Određivanje pokrivenosti signala, lokacije predajnika
Trgovina i ekonomija	Analiza stanja na berzi, osiguranje, direktna prodaja, ciljana prodaja, lokacija maloprodaja
Turizam	Lokacije i upravljanje kapacitetima i turističkim atrakcijama
Vojska i odbrana	Pronalaženje cilja, pomoć u taktičkom planiranju, modeliranje mobilnih naredbi, integracija obaveštajnih podataka
Životna sredina	Nadgledanje, modelovanje i menadžment degradacije zemljišta; procena zemljišta i planiranje poljoprivrede; klizišta: kvalitet i količina voda; nesreće; kvalitet vazduha; modelovanje vremenske prognoze.

GIS nalazi veliku primenu u *poljoprivrednoj proizvodnji*. Za velike parcele moguće je kreirati osnovne karte koje uključuju puteve, kuće, ambare i granice poseda i preklapati ih sa kartama drenažnih sistema, pedološkim kartama, kartama upotrebe zemljišta, kartama upotrebe pesticida i širenja biljnih bolesti, kao i topografskim kartama. Mnoge kompanije i instituti razvijaju precizne poljoprivedne sisteme uz upotrebu GIS-a da bi planirali fertilizaciju, upotrebu pesticida i herbicida i navodnjavanje na optimalan način. Poljoprivrednicima je na raspolaganju veliki broj GIS podataka koji se odnose na tip tla, satelitske i avionske snimke, topografsko, geomorfološko stanje i stanje vlažnosti tla (Jovanović, Đurđev, Srdić, & Stankov, 2012). Rasterski podaci se često koriste u poljoprivredi i šumarstvu za upravljanje proizvodnjom useva. Na primer, pomoću satelitskih snimaka poljoprivrednog zemljišta,

moguće je identifikovati zone gde biljke rastu slabije i onda upotrebiti te informacije za primenu više đubriva samo na identifikovanim oblastima. Šumari koriste rasterske podatke da procene koliko se drvne građe može dobiti iz određenih oblasti (Sutton, Dassau, & Sutton, 2009). Kako poljoprivredna proizvodnja u velikoj meri zavisi od niza prirodnih faktora, a istovremeno je zavisna i od uticaja određenih socio-ekonomskih komponenti, to je njeno uspešno planiranje i upravljanje direktno zavisno od stepena upoznatosti sa prostorom u kome se ova delatnost odvija (Manić, Gajović, & Popović, 2016). GIS je tehnologija kojom se lako, jednostavno, efikasano i kvalitetno mogu prikupiti podaci vezani za prostor u kome se odvija ta proizvodnja. Primena GIS-a u poljoprivredi može se posmatrati na makro i mikro nivou, a posmatrano po vrsti poljoprivrednih delatnosti, najčešća je u oblasti zemljoradnje (Manić, Gajović, & Popović, 2016):

- Na makro nivou: GIS se najčešće koristi za upravljanje i planiranje poljoprivredom jednog regiona ili države. Pri tome se prate, ne samo fizičkogeografski, već i društvenogeografski faktori poljoprivredne proizvodnje. Pružaju se informacije donosiocima odluka u javnom sektoru kako bi minimizirali potencijalne greške u upravljanju poljoprivrednim sektorom. Ovakva primena GIS-a moguća je i na nivo jednog preduzeća koje se bavi poljoprivrednom proizvodnjom;
- Monitoring: proces praćenja različitih pokazatelja i indikatora: fizičkih i hemijskih svojstava zemljišta, klimatskih elemenata, hidroloških uslova, stanja useva i autohtone vegetacije (obavlja se primenom različitih tehnologija za prikupljanje podataka koji se potom pohranjuju u baze podataka i odmah bivaju raspoloživi da se putem GIS-a analiziraju i vizuelizuju);
- Primena agrotehničkih mera: na osnovu monitoringa stanja setvenih površina i samih useva moguće je, u gotovo realnom vremenu, određivati na kojim lokacijama je potrebno vršiti primenu određenih agrotehničkih mera, u kojoj količini i vrsti.

Razvojem različitih GIS aplikacija u segmentu poljoprivredne proizvodnje, došlo se i do specifičnog koncepta agrarne proizvodnje u kome ključnu ulogu, upravo, imaju GIS bazirane tehnologije. Koncept se naziva *precizna poljoprivreda* (eng. *Precision Agriculture*), a radi se o pažljivo kreiranom menadžmentu u poljoprivrednoj proizvodnji u kome poljoprivrednik na bazi prikupljenih podataka i raspoloživih informacija, odlučuje kada, gde, čime i koliko delovati kako bi poboljšao krajnje rezultate svog rada (povećanje prinosa) (Manić, Gajović, & Popović, 2016).

Postoje brojne aktivnosti, kako u SAD tako i u EU, koje se odnose na upotrebu daljinski prikupljenih podataka kao ulaz za oficijelnu statistiku poljoprivrednog sektora, za predviđanje stanja i prinosa useva, procenu i mapiranje. Među aktivnostima u SAD su istraživanja o upotrebi administrativnih GIS podataka, istraživanja o upotrebi slika

visoke rezolucije za brojanje stabala limuna, i programi mapiranja useva. U EU su napredci programa MARS (eng. *Monitoring Agriculture with Remote Sensing*) u veći broj zemalja, kontinuirano unapređenje diseminacije podataka i korišćenje senzora poput za praćenje stanja useva (Hanuschak & Delincé, 2004).

Primena daljinskog uzorkovanja i analiza takvih podataka je od posebnog značaja za ovo istraživanje, jer ćemo se kroz studije primera baviti klasifikacijom prikupljenih podatka (slika) za primene u poljoprivrednoj proizvodnji, te je zato u ovom poglavlju primeni GIS-a i analizama takvih prostonih podataka posvećena značajna pažnja.

3 DALJINSKO UZORKOVANJE

Posmatranje Zemlje je prikupljanje informacija o fizičkim, hemijskim, biološkim i geometrijskim svojstvima naše planete; pomaže nam da procenimo status i pratimo promene prirodnog i kulturološkog okruženja. Prema tome, mapiranje, praćenje i predviđanje su oblici upotrebe posmatranja Zemlje. Posmatranje Zemlje nam daje geoprostorne podatke. Prikupljanje geoprostornih podataka se može smatrati polaznom tačkom u razvoju ciklusa osmatranja-analize-planiranja-razvoja-osmatranja, itd. (Tempfli, Kerle, Huurneman, & Janssen, 2009).

Daljinska detekcija ili *daljinsko uzorkovanje* (eng. *Remote Sensing*) je naučna oblast i tehnika za prikupljanje informacija o objektu (najčešće Zemljinoj površini) bez dolaženja u kontakt sa njim (dakle na daljinu, npr. iz letelice, satelita, itd.) a sprovodi se uzorkovanjem (očitavanjem) putem beleženja reflektovane ili emitovane energije (elektromagnetno zračenje, akustičnost, itd.) objekta, procesiranjem, analiziranjem i primenom informacija (Khorram, Wiele, Koch, Nelson, & Potts, 2016; Tempfli, Kerle, Huurneman, & Janssen, 2009; Aber & Aber, 2017; Levin, 1999). Ovakva definicija je prilično široka, ali u opštem slučaju se odnosi na upotrebu vazdušnih platformi, poput aviona, dronova, zmajeva, balona, i satelita za prikupljanje rasterskih fotografija. Rasterski podaci određuju prostor kao kontinuirana serija redova i kolona ćelija ili piksela, svaki od kojih poseduje svoju vrednost atributa (Aber & Aber, 2017). „Daljinsko“ je jer se opažanje radi na daljinu bez fizičkog kontakta sa objektom posmatranja. Možemo koristiti uzorkovanje (detekciju) i uređaje za prikaz u realnom vremenu, ili uređaje za snjimanje energije, koju emituje ili reflektuje objekat ili prizor. Energija može biti svetlost ili drugi oblik elektromagnetnog zračenja, sila uticaja ili akustične energije (Tempfli, Kerle, Huurneman, & Janssen, 2009; Levin, 1999). U poslednjih 50 godina razvijene su brojne platforme daljinskog uzorkovanja na velikim visinama (letelice i sateliti) i senzori za slikanje koji skupljaju podatke iz različitih oblasti elektromagnetnog spektra. Simultano sa napredkom u prikupljanju slika su se razvijale tehnike prikazivanja i obrade slika i povezani algoritmi. Danas, daljinsko uzorkovanje je prepoznata interdisciplinarna oblast širom sveta. Često je uparena sa disciplinama obrade slika, geografskog informacionog sistema (GIS) i GPS-om za široku oblast geospatijalne nauke i tehnologije (Khorram, Wiele, Koch, Nelson, & Potts, 2016). Iako je daljinsko uzorkovanje zasebna oblast, često je komplementarna sa GIS analizama, dodajući jedinstvene informacije i tehnike analize paleti GIS alata (Aber & Aber, 2017). Slične definicije sa gore navedenim se takođe mogu naći u (Krapivin, Varotsos, & Soldatov, 2015; Deekshatulu, et al., 1995).

Osnovne ideje o daljinskoj detekciji, odnosno osmatranju Zemlje i kosmosa, nastaju u doba renesanse otkrićem i upotrebom prvih instrumenata. Kasnije, sa razvojem tehnike započinje novo razdoblje snimanja površi Zemlje i nastaju prve fotografije. Gaspard Felix Tournachon poznatiji kao Nadar, francuski fotograf, načinio je prvu fotografiju sa određene visine iz balona iznad Pariza. Godine 1908. Francuski

pilot Wilburg Wright iz letilice je fotografisao deo teritorije Francuske i Italije. Taj događaj se u naučnim krugovima označava kao početak razvoja daljinske detekcije u današnjem smislu (Jovanović, Đurđev, Srdić, & Stankov, 2012). Savremeno doba daljinske detekcije počinje sa vazdušnim fotografijama u dvadesetom veku i sazreva lansiranjem satelitskih senzora koji skupljaju multispektralne podatke širom planete, počev od 1970-e. Čitav niz prirodnih i ljudski kreiranih svojstava se može identifikovati na slici putem različitih metoda (Manson, Bonsal, Kernik, & Lambin, 2015). U američkim istraživačkim centrima, intenzivna istraživanja i unapređenje tehničkih rešenja za daljinsku detekciju podstaknuta su pripremama za let na Mesec. Kosmički brod Apolo 9 je prvi put bio opremljen multispektralnom kamerom za snimanja, što je kasnije razvijano i postavljeno na satelitima serije Landsat. U julu 1972. godine NASA je lansirala prvi satelit ERTS-1 (Earth Resources Technology Satellite). Prvi snimci urađeni multispektralnim uređajima korišćeni su za istraživanja i dobijanje informacija o površima pod poljoprivrednim kulturama, zemljištu, mineralima, rastu urbanih sistema i mnogim drugim pojavama. Ovaj satelit je preimenovan u Landsat 1 i svi sateliti kasnije su nosili isto ime Landsat 2, Lansat 3, Landsat 4, itd. (Jovanović, Đurđev, Srdić, & Stankov, 2012). Daljinsko uzorkovanje pomoću satelita se transformisalo u manje od 30 godina od sporadičnog istraživačkog alata do robe dostupne širokom broju korisnika. Osnovna prednost satelitskih snimaka je što variraju u spektralnoj, prostornoj i vremenskoj rezoluciji, te se mogu koristiti u velikom broju primena i pružiti komletniji pogled na posmatrani objekat (Srivastava, Mukherjee, Gupta, & Islam, 2014).

Neke od osobina daljinskog uzorkovanja je da pruža detaljne podatke o velikim površinama, pruža podatke o stvarnom mestu posmatranja, opaža površinska svojstva, omogućava visoku frekventnost pregleda, da je „multi-kanalna“ i da je ekonomično (Tempfli, Kerle, Huurneman, & Janssen, 2009). Dok je mapa uvek subjektivna, jer odlučujemo šta da prikažemo na njoj i kako da je prestavimo, daljinsko uzorkovanje podrazumeva objektivan prikaz elektromagnetnog signala koji dolazi do senzora. Pored toga, za razliku od mape koja je prikaz zemlje „na papiru“, slike daljinskog uzorkovanja prikazuju i reljefne razmeštaje i geometrijsku distorziju (Levin, 1999).

Postoje dve vrste daljinskog uzorkovanja, aktivno i pasivno, i generalno se koriste za različite primene. Aktivno daljinsko uzorkovanje uključuje slanje signala i čekanje na njegov povratak do senzora. Primeri su RADAR i LIDAR. Pasivno daljinsko uzorkovanje ne odašilje signal koji bi se vratio, već snima informacije koristeći energiju koja je prisutna u okruženju. To znači da se pasivne slike generalno prikupljaju tokom dana, kada sunce pruža veliku količinu zračenja koje se reflektuje od Zemljinu površinu. Jedan od najbitnijih elemenata daljinskog uzorkovanja je što nam dozvoljava da opažamo informacije proizvedene izvan vidljivog svetlosnog spektra (Aber & Aber, 2017).

Kako je ranije opisano, GIS softver generalno upravlja i vektoriskim i rasterskim podacima. Podaci daljinskog uzorkovanja pripadaju grupi rasterskih podatka i u opštem

slučaju zahtevaju manipulacije prostornim podacima koje GIS ne nudi. Međutim, nakon što su analize daljinski uzorkovanih podataka završene, rezultati se obično kombinuju sa GIS-om ili prostornim bazama podataka za dalje analize. U poslednje vreme sve više vektorskih mogućnosti se dodaje u softver za daljinsko uzorkovanje a takođe neke funkcije daljinskog uzorkovanja se dodaju u GIS (Levin, 1999).

Elementi neophodni za izvođenje daljinskog uzrokovanja su (Canada Centre for Remote Sensing, 2007):

1. Izvor energije ili osveljenost: koji obezbeđuju elektromagnetnu energiju objektu od interesa.
2. Zračenje i uticaj atmosfere: uticaj na energiju koja putuje ka objektu ili od objekta ka senzoru.
3. Interakcija energije (svetlosti) sa objektom: apsorpcija, propuštanje i reflektovanje. Merenjem energije koja se reflektuje (ili je emitovana) od objekta možemo identifikovati spektralni odgovor tog objekta. Poređenjem različitih obrazaca odgovora koje daju različiti objekti i njihove osobine, možemo ih razlikovati. Npr. voda i vegetacija mogu slično reflektovati svetlost u vidljivom spektru, ali drugačije u infracrvenom.
4. Snimanje energije putem senzora: prikupljanje elektromagnetnog zračenja koje reflektuje ili emituje objekat. Platforme sa senzorima se mogu nalaziti na zemlji, na letelici u zemljinoj atmosferi ili letelici ili satelitu koji se nalazi izvan zemljine atmosfere.
5. Emitovanje, prijem i procesiranje energije: energija koju registruje senzor se mora zabeležiti, obično u elektronskoj formi, i preneti do stanice za procesiranje, gde se podaci beleže u vidu slike.
6. Interpretacija i analiza: slika se interpretira i procesira, vizuelno ili računarskim putem, za dobijanje informacija o posmatranom objektu.
7. Primena: poslednji korak je primena dobijenih informacija o objektu, za njegovo bolje razumevanje, otkrivanje novih osobina, ili za pomoć pri rešavanju određenih problema.

U nastavku ove disertacije za nas je od posebnog značaja praktično samo tačka 6, koja se odnosi na interpretaciju i analizu podataka zabeleženih u formi digitalnih slika, i to njihova interpretacija računarskim putem, odnosno statističkim metodama.

Dakle, za nas ključni deo procesa daljinskog uzorkovanja je ekstrakcija smislenih informacija iz slike putem njene interpretacije i analize. Ona uključuje identifikaciju i/ili merenje različitih ciljnih objekata na slici kako bi se dobile korisne informacije o njima. Ciljni objekti mogu biti tačke, linije ili površine i moraju imati svojstvo da se mogu razlikovati od ostalih oblika na slici. Dakle, automatsko procesiranje i analiza digitalnih slika se sprovode za automatsku identifikaciju ciljnih objekata i dobijanje informacija,

načelno, bez ručnih intervencija analitičara. Ipak, procesiranje i analiza se u praksi retko sprovode u potpunosti bez ljduskih intervencija i nadzora, već kao pomoć analizi koju sprovodi analitičar (Canada Centre for Remote Sensing, 2007). Međutim, kada govorimo o simultanoj analizi većeg broja spektara, koji podrazumevaju i veću količinu podataka, nesporan je značaj uloge automatskog procesiranja na ukupnu efikasnost procesa. U tome se takođe ogleda i jedan od doprinosa ove disertacije – povećanje nivoa automatizacije procesa automatskog uzorkovanja. U tom smislu prepoznavanje ciljnih objekata ili segmenata slike je ključ interpretacije podataka i dobijanja informacija, što podrazumeva posmatranje razlika između ciljnog objekta i njegove okoline poređenjem nekih, ili svih, njegovih vizuelnih elemenata, poput boje, oblika, veličine, obrasca, teksture, senke i asocijativnosti.

Većina standardnih funkcija obrade i analize slika se mogu kategorizovati u sledeće četiri kategorije: predprocesiranje, poboljšanja, transformacije, i klasifikacija i analiza (Canada Centre for Remote Sensing, 2007). Koraci koji se odnose na predprocesiranje, poboljšanja i transformacije slika su od značaja za ovo istraživanje u disertaciji jedino u smislu da se bave pripremom ulaznih podataka, te ćemo se samo na tom nivou njima i baviti, dok ih nećemo dublje razrađivati. Detaljnije ćemo se baviti tehnikama klasifikacije i analize podataka. Standardne procedure klasifikacije se obično grupišu u dve grupe na osnovu korišćenih metoda (Canada Centre for Remote Sensing, 2007):

- *Klasifikacija s nadzorom* (eng. *Supervised Classification*): gde analitičar identifikuje delove slike kao reprezentativne uzorke tipova površine (klase informacija), koji se nazivaju oblastima učenja, a na osnovu čijih se spektralnih svojstva piksela algoritam „trenira“ da prepoznaje spektralno slične oblasti za svaku od klasa. Jedna od oblasti multivarijacione analize koja se bavi ovakvom klasifikacijom je Diskriminaciona analiza.
- *Klasifikacija bez nadzora* (eng. *Unsupervised Classification*): u ovom slučaju su spektralne klase prvo grupisane, jedino na osnovu svojih numeričkih podatka, a zatim se analitičar bavi njihovim validiranjem i mapiranjem u klase informacija (ukoliko je moguće). Jedna od oblasti multuvarijacione analize koja se bavi ovom vrstom klasifikacije je Analiza grupisanja (Klaster analiza).

3.1 Primena daljinskog uzorkovanja

Primenu daljinskog uzorkvoanja (Khorram, Wiele, Koch, Nelson, & Potts, 2016) grupišu u sledeće grupe: zemaljske primene daljinskog uzorkovanja, atmosferske primene, promatranje priobalnog i okeanskog ekosistema, planterne i ekstrasolarna promatranja, međunarodni zakoni, mapiranja i politike. Oblasti primene (Levin, 1999) navodi kao:

- Poljoprivreda: klasifikacija useva, procena stanja useva, procena prinosa, mapiranje karakteristika i prakse tretiranja zemljišta, zakonsko nadgledanje, itd.
- Šumarstvo: mapiranje vrsta i inventar šuma, nadzor krčenja i obnova šuma, nadzor požara, procena biomase, nadzor stanja i vitalnosti, itd.
- Geologija: mapiranje sastava i strukture površina, geologija životne sredine, geobotanika, nadzor i mapiranje sedimentacija, nadzor geo-hazarda, mapiranje planete, itd.
- Hidrologija: mapiranje i nadzor močvara, snežne mase, leda, glečera, poplava, ušća reka, i sl., procene vlažnosti zemljišta, planiranje irigacije, itd.
- Morski led: koncentracija leda, detekcija santi, topografija površina, bezbedno rutiranje brodova, stanje leda, praćenje zagađenosti i staništa divljih životinja, itd.
- Zemljišni pokrivač i upotreba zemljišta: upravljanje prirodnim resursima, zaštita prirodnih staništa, mapiranje za GIS, širenje urbanih zona, rutiranje i planiranje logistike resurasa, parcelizacija, itd.
- Mapiranje: planimetrija, modeli elevacije, mapiranje osnova, topografsko mapiranje
- Nadzor okeana i obala: predviđanje oluja, procena ribljeg fonda i morskih sisara, nadzor naftnih mrlja, nadzor brodskog transporta, mapiranje zona plima i oseka, itd.

Kao posebnu oblast primene daljinskog uzorkovanja (detekcije) u ovoj disertaciji izdvojićemo Preciznu poljoprivredu, kao jedna od oblasti koja očekuje veliki procvat u budućnosti ali još uvek nije dovoljno zastupljena danas, te stoga su više nego smisleni svi napori unapređenja te grane.

Precizna poljoprivreda (eng. *Precision Agriculture*, PA) je primena geoprostornih tehnika i senzora (npr. GIS-a, daljinske detekcije, GPS-a) za identifikaciju varijacija na polju i njhovog tretiranja različitim (alternativnim) strategijama (Zhang & Kovacs, 2012). Napretkom u elektronici i informacionim tehnologijama, različiti sistemi detekcije su razvijeni za proizvodnju useva i precizni podaci o prostornom varijabilitetu na poljima su veoma značajni za efikasnu proizvodnju (Lee, Alchanatis, Yang, Hirafuji, & Moshou, 2010). Konkretnije, satelitski snimci (slike) visoke rezolucije se danas često koriste za proučavanje ovih varijacija u stanju useva i zemljišta. Međutim, dostupnost i često ograničavajuća cena takvih snimaka iziskuje pronalaženje alternativnih koncepata za konkretnu primenu u preciznoj poljoprivredi. Slike prikupljenje sa platformi daljinskog uzorkovanja na niskim visinama ili mali bespilotni sistemi (eng. *Unmanned Aerial System*, UAS) su se pokazale kao potencijalno važne alternative s obzirom na njihovu nisku cenu, velike prostorne i vremenske rezolucije, i veliku fleksibilnost kod

programiranja prikupljanja slika (Zhang & Kovacs, 2012). Time i nije iznenađenje što se u poslednje vreme javlja veliki broj studija u vezi sa primenom bespilotnih letelica za preciznu poljoprivredu i što ćemo mi upravu na te primene obratiti najveću pažnju u nastavku ove disertacije. Opšte faze u praksi precizne poljoprivrede su: prikupljanje podataka, mapiranje varijabiliteta polja, odlučivanje, i na kraju sprovođenje upravljanja i treitranja polja. Daljinsko uzorkovanje može biti uključeno u prve tri od navedenih faza. Konkretnije ključno je prikupiti ažurne slike-mape tokom procesa donošenja odluka, te bi varijabilitet polja mogao biti mapiran pomoću daljinski uzorkovanih slika (Zhang & Kovacs, 2012). Generalno primene daljinskog uzorkovanja u poljoprivredi uključuju nadzor i mapiranje svojstava zemljišta, klasifikaciju vrsta useva, kontrolu štetočina kod useva, detekciju stresa biljaka uzrokovanog nedostatkom ili viškom vode, analizu hemijskih svojstava lišća, i kontrolu i nadzor korova (Zhang & Kovacs, 2012).

U pogledu podizanja nivoa automatizacije procesa, smisleno je govoriti o automatskoj proceduri identifikacije, koja uključuje primenu bespilotnih sistema, a koja bi identifikovala anomalije u usevima na polju (npr. korov, bolesti, stres usled suše, itd.). To bi moglo uključivati ekstrakciju različitih bioloških pokazatelja i varijabli (npr. indeks oblasti lista – eng. *Leaf Area Index*, koncentraciju hlorofila, prinos, itd.) pomoću bespilotnih sistema (Zhang & Kovacs, 2012). Detalji prethodnih definicija se takođe mogu pronaći i u (Ballesteros, Ortega, Hernández, & Moreno, 2014). Autori navode kako povećanje prostorne i vremenske rezolucije geomatskih proizvoda dobijenim pomoću bespilotnih letelica moraju biti propraćeni novim algoritmima i tehnikama apstrakcije informacija iz tih proizvoda. Jasan primer je upotreba vegativnih indeksa poput NDVI (eng. *Normalized Difference Vegetation Index*), koja se može zameniti tehnikama računarske vizije (eng. *Computer Vision*) ili drugim indeksima koji baziraju na informacijama iz RGB spektra, koje se mogu dobiti pomoću jeftinih senzora (Ballesteros, Ortega, Hernández, & Moreno, 2014). Dalje, prema (Pérez-Ortiz, i drugi, 2015) kombinacija mašinskog učenja (eng. *Machine Learning*) i dronova za preciznu poljoprivredu, iako pokazuju dobru sinergiju, je i dalje istraživačka oblast u nastajanju, uglavnom nerazvijena. Međutim, danas kada je UAV (dron) tehnologija dovoljno napredovala i kada su mnogi problemi prevaziđeni, uključujući i cenu, mašinsko učenje i segmentacija slika su pogodne tehnologije za navedene zadatke, te i studije koje se bave njihovim kombinovanjem su sve češće u oblasti daljinskog uzorkovanja (Pérez-Ortiz, i drugi, 2015). Pored toga, senzorski sistemi, poput multispektralnih i hiperspektralnih, mogu pribaviti podatke visoke rezolucije o poljoprivrednim usevima, pa i napredak u tehnologiji senzora, zajedno sa napredkom u informatici i geografskim informacionim sistemima, pružaj nove mogućnosti za preciznu poljoprivredu i čini osnovu za ranu detekciju i identifikaciju problema sa usevima poput korova i bolesti (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

Budućnost daljinskog uzorkovanja leži u stvaranju brojnih vrsta preciznih, ažurnih i podataka visoke rezolucije dobijenih daljinskom detekcijom i izvedenih geoprostornih informacija koje su spremne za korišćenje u svim oblastima od interesa. Naučna

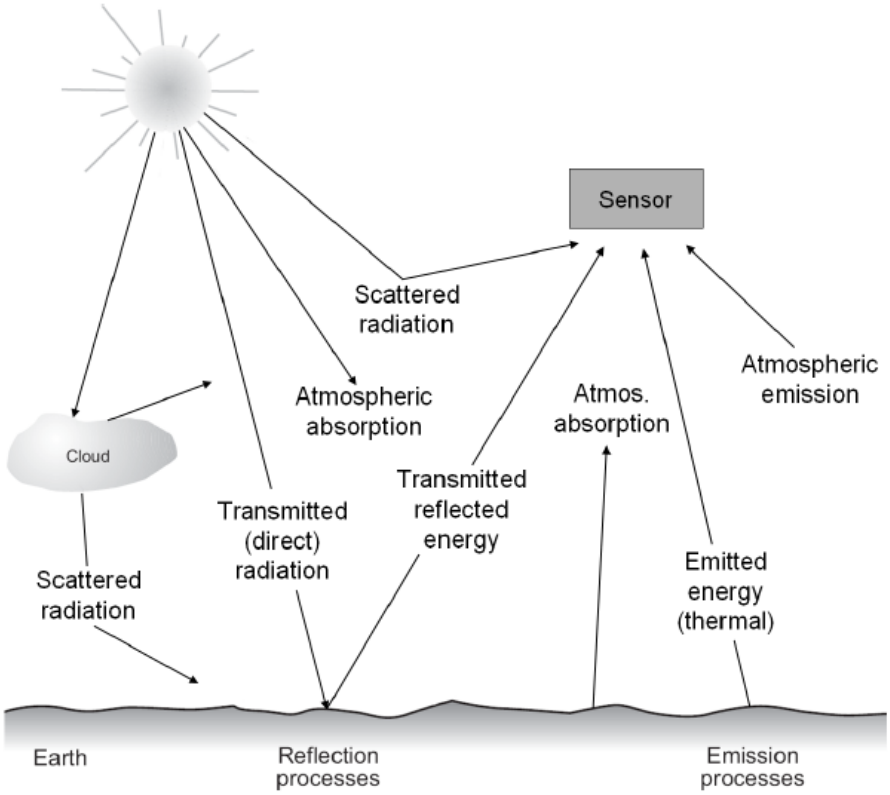
zajednica iščekuje dan kada će skupovi podataka biti dostupni za analize u nespecijalizovanim softverskim paketima (Srivastava, Mukherjee, Gupta, & Islam, 2014). Buduće trendove u daljinskom uzorkovanju (Khorram, Wiele, Koch, Nelson, & Potts, 2016) navode u više segmenata. Proboji u nauci o podacima (eng. *Data Science*) će inicirati i zahtevati napredak svih aspekata daljinskog uzorkovanja i geoprostornih tehnologija. Predviđa se napredak i u hardveru i softveru, kao i u primeni tehnologije daljinskog uzorkovanja na nove i različite načine (Khorram, Wiele, Koch, Nelson, & Potts, 2016). Trendovi koji će uticati i produbljivati primenu daljinskog uzorkovanja u budućnosti su: dalja minijaturizacija i integracija elektronike, razvoj prikupljanja podataka pomoću bespilotnih letelica (UAV), rast računarske moći pomoću paralelizacije, *cloud computing*-a i kvantnih i bioloških računara, razvoj novih i moćnijih senzora i antena, rast transmisijske snage predajnika aktivnih sistema, minijaturizacija optike, napredak tehnologije skladištenja podataka, razvoj malih satelita, napredak tehnologije ekrana i mobilnog računarstva, i napredak tehnika za obradu velikih podataka (*Big Data*) (Khorram, Wiele, Koch, Nelson, & Potts, 2016). Pored ubrzanog napretka tehnoloških apsketa prikupljanja i obrade podataka daljinskog uzorkovanja, zajednica istraživača će direktno profitirati i kroz zajedničke napore developera i zajednice korisnika poput biologa, inženjera, socijalnih naučnika, javnog zdravlja i pravnog sistema, medija, prostornog planiranja i ekološkog menadžmenta. Prilike za korišćenje daljinskog uzorkovanja su neograničene (Khorram, Wiele, Koch, Nelson, & Potts, 2016).

3.2 Elektromagnetno zračenje

Elektromagnetna energija se odnosi na svaku energiju koja se kreće brzinom svetlosti u harmoničnom obrascu (harmoničnost implicira da se komponentni talasi ravnomerno i sa ponavljanjem rasprostiru u vremenu) (Levin, 1999; Khorram, Wiele, Koch, Nelson, & Potts, 2016). Koncept talasa objašnjava širenje elektromagnetne energije, ali ta energija se može opaziti jedino u smislu njene interakcije sa materijom. Elektromagnetni talasi se mogu opisati svojom brzinom, talasnom dužinom (udaljenošću između istih pozicija u dva susedna ciklusa, merenom standardnim metričkim sistemom, najčešće u mirkometrima ili nanometrima) i frekvencijom (Levin, 1999).

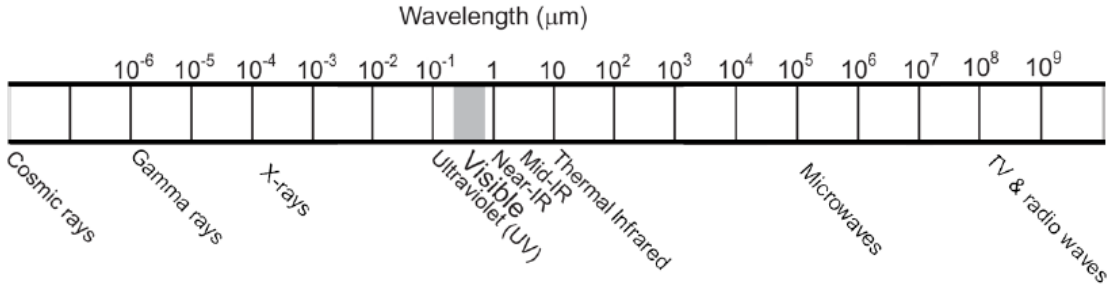
Kada se elektromagnetna energija susretne sa materijom, bilo čvrstom, tečnom ili gasovitom, postoji nekoliko vrsta interakcija. Nauka daljinskog uzorkovanja detektuje i snima promene koje time nastaju. Rezultujuće slike i podaci se interpretiraju kako bi se daljinski identifikovale osobine materije koja je izazvala promene zabeležene elektromagnetnog zračenja. Tim interakcijama zračenje može biti: propušteno, apsorbovano, emitovano, raspršeno i reflektovano (Levin, 1999). Grafički prikaz interakcija energije sa atmosferom i zemljinom površinom dat je na slici 3.2.1. Sistemi pasivne daljinske detekcije beleže energiju koju prirodno zrače ili reflektuju objekti, dok

sistemi aktivne daljinske detekcije obezbeđuju sopstveni izvor energije, koji se usmerava prema objektu kako bi se beležila vraćena energija (Levin, 1999).



Slika 3.2.1: Interakcija energije sa atmosferom i Zemljinom površinom (Tempfli, Kerle, Huurneman, & Janssen, 2009).

Vidljiva svetlost je samo jedna kategorija elektromagnetnog zračenja, a kategorije inače uključuju: gama zrake, X-zrake, ultraljubičaste, vidljivo zračenje (svetlost), infracrvene, mikrotalase i radio talase (slika 3.2.2). Zajedno, one čine elektromagnetni spektar (Khorram, Wiele, Koch, Nelson, & Potts, 2016; Tempfli, Kerle, Huurneman, & Janssen, 2009). Svaka od ovih imenovanih sekcija predstavlja raspon talasnih dužina, a ne jednu specifičnu talasnu dužinu. Elektromagnetni spektar je kontinualan i nema jasno definisane granice klasa (Tempfli, Kerle, Huurneman, & Janssen, 2009).



Slika 3.2.2: Elektromagnetni spektar

Pregled osobina elektromagnetnog zračenja pokazuje da različiti oblici energije zračenja mogu pružiti različite informacije o svojstvima površine terena i da različite primene osmatranja zemlje profitiraju od korišćenja različitih opsega elektromagnetnog zračenja (Tempfli, Kerle, Huurneman, & Janssen, 2009). Inženjer geoinformatike koji želi razlikovati objekte za topografsko mapiranje preferira upotrebu optičkih senzora koji rade u vidljivom spektru. Stručnjaci koji se bave zaštitom životne sredine koji prate toplotne gubitke nuklearnih elektrana će koristiti senzore koji detektuju toplotna zračenja. Geolog koga zanima struktura stena će se osloniti na mikrotalase, itd. (Tempfli, Kerle, Huurneman, & Janssen, 2009). Tabela 3.2.1 prikazuje primere upotrebe različitih opsega u različite svrhe.

Tabela 3.2.1: Tematski opsezi NASA LANDSAT satelita (Gonzalez & Woods, 2002).

Opseg	Naziv	Tal. dužina (μm)	Karakteristike i upotreba
1	Vidljiva plava	0.45-0.52	Maksimalno prodiranje vode
2	Vidljiva zelena	0.52-0.60	Dobra za merenje zdravlja biljaka
3	Vidljiva crvena	0.63-0.69	Razlikovanje vegetacije
4	Blizu infracrveni	0.76-0.90	Mapiranje obala i biomase
5	Srednji infracrveni	1.55-1.75	Sadržaj vlage zemljišta i vegetacije
6	Termalni infracrveni	10.4-12.5	Vlažnost zemljišta, termalno mapiranje
7	Srednji infracrveni	2.08-2.35	Mapiranje minerala

Ranije spomenuti vegetativni indeksi (eng. *Vegetation Indices*, VI) su algebarske kombinacije nekoliko spektralnih opsega osmišljenih da naglase stanje i svojstva vegetacije (biomasa, apsorbovano zračenje, nivo hlorofila, itd.) (Candiago, Remondino, Giglio, Dubbini, & Gattelli, 2015). Za studije vegetacije, istraživači koriste činjenicu da je refleksija niska u plavim i crvenim opsezima spektra, dok je visoka u zelenom. U blizu-infracrvenom spektru (eng. *Near-Infrared*, NIR), refleksija vegetacije je značajno jača nego u vidljivom ospegu. Među različitim VI, najrasprostranjenija i izvodljiva iz multispektralnog senzora sa tri opsega, je upotreba: NDVI, GNDVI i SAVI (Candiago, Remondino, Giglio, Dubbini, & Gattelli, 2015). Npr. NDVI je direktno povezan sa

fotosintetičkim kapacitetom i apsorbcijom energije biljaka. U opštem slučaju ukoliko postoji mnogo više reflektovanog zračenja u blizu-infracrvenim talasnim dužinama u odnosu na vidljive talasne dužine, onda je verovatnije da je vegetacija u datom pikselu daljinski uzorkovane slike gušća i verovatno sadrži neku vrstu šume, useva i sl. (Sellers, 1985; Mynemi, Hall, Sellers, & Marshak, 1995).

Ovde je još korisno napomenuti da se u odsustvu merenja sa tla pomoću spektro-radiometra, koje je neophodno da bi se izlaz iz senzora konvertovao u konkretnu refleksiju, izračunavanje pokazatelja može bazirati i na *digitalnim brojevima* (eng. *Digital Number*, DN) (Candiago, Remondino, Giglio, Dubbini, & Gattelli, 2015). Ovo je za nas naročito interesantno jer će se upravo takav pristup najviše koristiti u nastavku ove disertacije.

3.3 Senzori i platforme za daljinsko uzorkovanje

Daljinski senzor je uređaj koji detektuje elektromagnetnu energiju, kvantifikuje je, i obično beleži je, u analognom ili digitalnom obliku (Tempfli, Kerle, Huurneman, & Janssen, 2009). Senzori se nazivaju *aktivnim* ukoliko detektuju refleksione odgovore objekata ozračenih od veštački generisanih izvora energije, i *pasivnim* ukoliko detektuju reflektovano ili emitovano elektromagnetno zračenje iz prirodnih izvora (Campbell & Shin, 2011). Neki primeri senzora su: altimetar, radiometar, senzor gama zračenja, filmske kamere, digitalne kamere, video kamere, multispektralne kamere, hiperspektralne kamere, termalni skeneri, mikrotalasni radiometri, laserski skeneri, radarska kamera, radarski radiometar, sonar (Tempfli, Kerle, Huurneman, & Janssen, 2009; Levin, 1999). Platforme i senzori za daljinsko uzorkovanje variraju u svojim *spektralnim*, *prostornim*, *temporalnim* i *radiometrijskim* karakteristikama ili domenima, odnosno rezolucijama (Manson, Bonsal, Kernik, & Lambin, 2015; Levin, 1999; Campbell & Shin, 2011). (Levin, 1999) dodatno uvodi i pojam digitalne rezolucije (kod digitalnih senzora).

Spektralni raspon se odnosi na deo elektromagnetnog spektra u kome senzor deluje (Manson, Bonsal, Kernik, & Lambin, 2015). *Spektralna rezolucija* je sposobnost senzora da razlikuje različite intervale talasnih dužina elektromagnetnog spektra (Campbell & Shin, 2011). Zračenje je senzorom opaženo za spektralni opseg, ne za pojedinačnu talasnu dužinu. Spektralni opseg, ili opseg talasnih dužina, je interval elektromagnetnog spektra za koji se beleži prosečno zračenje. Senzori poput panhromatskih kamera, radara, ili laserskih skenera mere samo u jednom specifičnom opsegu tok multispektralni skeneri ili digitalne kamere mere u nekoliko spektralnih opsega istovremeno (Tempfli, Kerle, Huurneman, & Janssen, 2009). Pojam *multispektralni* označava senzor koji može zabeležiti više različitih delova opsega, dok se *panhromatski* odnosi na senzore koji skupljaju jedan uski elektromagnetni opseg.

Pojedini senzori se nazivaju *hiperspektralni* jer mogu prikupiti stotine diskretnih delova spektra (Manson, Bonsal, Kernik, & Lambin, 2015). Različite klase svojstava i detalji slike se često mogu razlikovati poređenjem njihovog odziva kroz spektar talasnih dužina svetlosti. Zato upravo i postoje sistemi daljinskog uzorkovanja koji beleže energiju (svetlost) kroz nekoliko različitih opsega talasnih dužina u različitim spektralnim rezolucijama (multispektralni), ili njihove napredne verzije, koje detektuju stotine veoma uskih spektralnih opsega odjednom (hiperspektralni), npr. vidljive, blizu-infracrvenih, srednje-infracrvene i sl. delove elektromagnetnog spektra. Takvi senzori olakšavaju razlikovanje različitih objekata na osnovu njihovih različitih odziva u različitim uskim spektralnim opsezima (Canada Centre for Remote Sensing, 2007; Levin, 1999).

Prostorni domen ili *rezolucija* se odnosi na rezoluciju senzora (najmanja diskretna oblast koju može identifikovati na tlu), njegov domet ili obuhvat (veličina površine koju može uhvatiti odjednom), i prostorne obrasce koji se mogu uočiti sa slike (Manson, Bonsal, Kernik, & Lambin, 2015). Nivo detalja koje je moguće opaziti zavisi od prostorne rezolucije senzora, koja se odnosi na veličinu najmanje moguće svojstvo koje je moguće opaziti (Canada Centre for Remote Sensing, 2007; Levin, 1999; Campbell & Shin, 2011). Nekoliko faktora određuju prostornu rezoluciju slike. Za podatke daljinskog uzorkovanja, prostorna rezolucija je obično određena mogućnostima senzora koji pravi slike. Npr. SPOT5 satelit može prikupiti slike gde je svaki piksel 10x10m. Drugi sateliti, npr. SPOT5 prikupljaju rezoluciju preciznosti svega 500x500m po pikselu. Kod fotografija iz vazduha, česte su rezolucije 50x50m po pikselu (Sutton, Dassau, & Sutton, 2009). Fotografije uslikane pomoću bespilotnih letelica (dronova), s obzirom na njihovu malu visinu letenja i malu brzinu kretanja, ča i sa jeftinijom opremom se lako postižu rezolucije od po 1cm do 5cm po pikselu.

Temporalni (vremenski) domen senzora se odnosi količinu vremena koju jedan snimak pokriva i koliko često se određena oblast uzorkoje putem senzora (Manson, Bonsal, Kernik, & Lambin, 2015). *Temporalna rezolucija* je vreme koje protekne između dva snimanja istog terena ili objekta (Campbell & Shin, 2011). Npr. interval ponovnog posećivanja površine od strane satelitskoz senzora je nekoliko dana. Prema tome, temporalna rezolucija ovog senzora je jednaka tom periodu. Spektralne karakteristike svojstava se mogu promeniti vremenom, i ove izmene se mogu detektovati skupljanjem i poređenjem multi-temporalnih slika (Levin, 1999).

Radiometrijska rezolucija sistema za prikupljanje slika definiše mogućnost razlikovanja veoma blage razlike u elektromagnetnoj energiji, odnosno osetljivost senzora na varijacije intenziteta (osvetljenosti) (Levin, 1999; Campbell & Shin, 2011). Što je finija radiometrijska rezolucija senzora, on je osetljiviji u detektovanju malih razlika reflektovane ili emitovane energije (Levin, 1999).

Digitalna rezolucija je broj bitova koji čine svaki digitalni uzorak. Podaci slike su predstavljeni (u rasterskom obliku) pomoću pozitivnih digitalnih brojeva koji variraju od 0 do (jedan manje od) eksponenta broja 2 (Levin, 1999).

Kako bi senzor prikupio i zapamtio reflektovanu ili emitovanu energiju sa cilja ili površine, mora se nalaziti na stabilnoj platformi udaljenoj od cilja ili površine koja se posmatra (Levin, 1999). U daljinskom uzorkovanju, nosač senzora (npr. satelit koji orbitira Zemljom) se naziva platforma, dok je sam senzor njegovo korisno opterećenje (teret). (Khorram, Wiele, Koch, Nelson, & Potts, 2016) navode kako se platforme za daljinsko uzorkovanje se grubo mogu podeliti u dve kategorije: platforme u vazduhu i platforme u svemiru, i kako se tokom dve poslednje decenije, značajno se povećao broj i jednih i drugih (Khorram, Wiele, Koch, Nelson, & Potts, 2016). (Levin, 1999) navodi kako se platforme mogu nalaziti na zemlji, vazduhoplovu ili balonu (odnosno generalno platformi Zemljinoj atmosferi), ili na svemirskom brodu ili satelitu izvan Zemljine atmosfere. (Tempfli, Kerle, Huurneman, & Janssen, 2009) navode kako senzori koji se koriste mogu funkcionisati sa svega nekoliko centimetara od tla pa do daleko izvan atmosfere. Senzori se često montiraju na vozila u pokretu (platforme), kao što su avioni ili sateliti, ali ponekad se i statične platforme koriste, npr. stub (Tempfli, Kerle, Huurneman, & Janssen, 2009). Senzori koji se nalaze na zemlji (npr. merdevinama, kranu, itd.) prikupljaju podatke koji se često upoređuju sa onima prikupljenih iz vazduha. Vazdušne platforme su primarno avioni i helikopteri (Levin, 1999). Zbog svojih orbita, sateliti omogućuju kontinuiranu i repetitivnu pokrivenost Zemljine površine. U izboru platforme koju koristiti, cena je obično među najznačajnijim faktorima (Levin, 1999). Satelitski snimci dolaze iz raznih izvora, neki su javni, neki privatni. Neki prikupljeni podaci su dostupni besplatno, poput onih koje generiše Landsat program, dok se kod drugih izvora naplaćuje pristup slikama (Aber & Aber, 2017).

Nekoliko poslednjih godina su godine velikog proboja komercijalne upotrebe bespilotnih letelica, popularnih dronova (UAV ili UAS). Razvoj jeftinih bespilotnih letelica i laganih senzora za slikanje u poslednjoj deceniji je rezultiralo velikim interesom za njihovu upotrebu. Taj interes proizilazi iz brojnih prednosti koje male bespilotne letelice imaju u odnosu na ostale platforme daljinskog uzorkovanja (Hung, Xu, & Sukkarieh, 2014). Prvo, male bespilotne letelice zahtevaju minimalnu infrastrukturu i mogu se koristiti iz udaljenih ili nepristupačnih lokacija, one su fleksibilne i lake za korišćenje, zatim male bespilotne letelice mogu biti veoma jeftine, što omogućava da budu lako zamenjive platforme prikupljanja podataka u visoko rizičnim uslovima (Hung, Xu, & Sukkarieh, 2014; Jannoura, Brinkmann, Uteau, Bruns, & Joergensen, 2015). Dalje, ono što je možda i najbitnije, u odnosu na prikupljanje slika satelitom, slike prikupljene bespilotnim letelicama obično imaju veću temporalnu (npr. dnevno prikupljanje, kad god dozvole vremenski uslovi) i veću prostornu rezoluciju (npr. u centimetrima), budući da lete na nižim visinama (Zhang & Kovacs, 2012; Hung, Xu, & Sukkarieh, 2014; Jannoura, Brinkmann, Uteau, Bruns, & Joergensen, 2015).

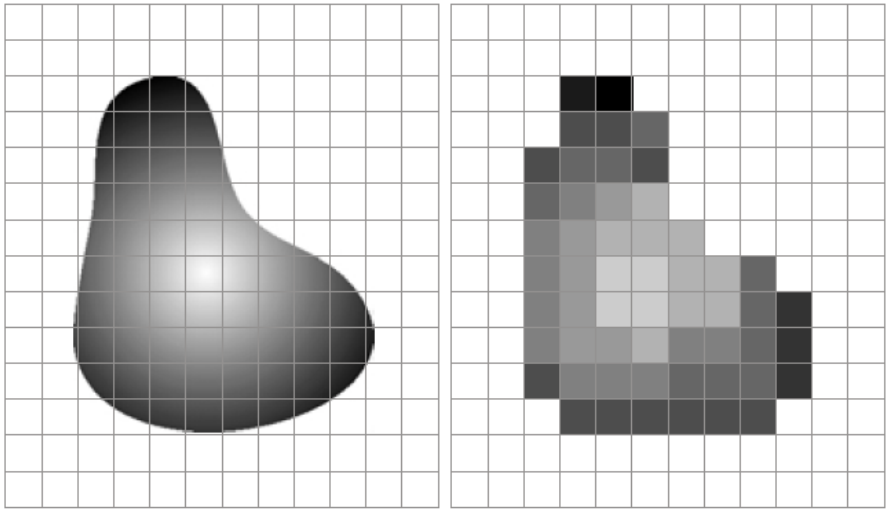
U delu koji je najinteresantniji za ovo istraživanje, to otvara potpuno novi skup mogućnosti primene u preciznoj poljoprivredi (PA). Npr. neke određene odluke u vezi sa upravljanjem farmama, poput detekcija i suzbijanje korova zahtevaju često slike visoke prostorne rezolucije, do nivoa centimetara, dok su podaci daljinskog uzorkovanja koji su najčešće na raspolaganju su slike srednje rezolucije (npr. Landsat, ASTER, SPOT5) i korisni su samo za studije većih razmera. Čak i najnoviji sateliti za prikupljanje slika (npr. WorldView-2, GeoEye-1) ne mogu obezbediti visoko frekventne podatke za hitne situacije (npr. praćenje stresa koji je posledica hraniva ili bolesti) sa ograničenim periodima ponovnog snimanja tla (1-2 dana). Vremenski uslovi su takođe česta prepreka za satelitske snimke, naročito kada je sezona useva u kišnoj sezoni. Veruje se da su troškovi, raspoloživost, fleksibilnost i obrada daljinskih podataka iz satelitskih snimaka ti koji čine njihovu primenu (u PA) nepodobnim i nepraktičnim (Zhang & Kovacs, 2012). S druge strane, bespilotne letelice bi mogle biti jeftina i praktična zamena za satelite i generalno avionske snimke za daljinski uzorkovane podatke visoke rezolucije. Pored različitih platformi koje su na raspolaganju u te svrhe, postoji veliki broj senzora za daljinsko uzorkovanje koji se mogu koristiti, kao što su čak i jeftine digitalne kamere, koje mogu biti i modifikovane ili specijalne kamere za upotrebu pomoću bespilotnih letelica, poput multispektralnih kamera i sl. (Zhang & Kovacs, 2012). Primene daljinskog uzorkovanja pomoću bespilotnih letelica u preciznoj poljoprivredi uključuje, ali se ne ograničava na, procenu prinosa, merenje hemijskog sastava, praćenje stanja useva, praćenje useva pod stresom, itd. (Zhang & Kovacs, 2012). Postoji više studija koje se detaljno bave analizom efikasnosti primene bespilotnih letelica, naročito u preciznoj poljoprivredi, koje ukazuju na i kvantifikuju različite efekte i ograničenja, poput (Peña, Torres-Sánchez, Serrano-Pérez, Castro, & López-Granados, 2015).

Za nas će u nastavku disertacije od najvećeg značaja biti digitalna kamera kao senzor, koji je elektro-optički daljinski senzor (Tempfli, Kerle, Huurneman, & Janssen, 2009), i bespilotna letelica kao platforma, odnosno bavićemo se detaljnije analizom podataka prikupljenih pomoću bespilotnih letelica i digitalnih kamera. S druge strane, više informacija o primeni radiometrije i algoritmima obrade tako prikupljenih podataka se mogu naći u (Krapivin, Varotsos, & Soldatov, 2015).

3.4 Obrada i analiza prikupljenih slika

Iz prethodnih poglavlja smo videli da ima više načina da se prikupe slike. Međutim, naš cilj je da iz uzorkovanih podataka generišemo digitalne slike. Konverzija kontinuiranih formi elektromagnetnih talasa, čija amplituda i prostorno ponašanje zavisi od fizičkog fenomena koji se posmatra, u digitalnu sliku uključuje procese *uzorkovanja* (eng. *sampling*) i *kvantizacije* (eng. *quantization*) (Gonzalez & Woods, 2002). Osnovna ideja iza uzorkovanja i kvantizacije je data na Slici 3.4.1. Slika levo prikazuje kontinualnu sliku, $f(x,y)$, koju želimo konvertovati u digitalni oblik. Slika može biti

kontinualna u odnosu na *x* i *y* koordinate, ali takođe u amplitudi. Da bi se konvertovala u digitalni oblik, moramo sam uzorkovati funkciju i u koordinatama i u amplitudi. Digitalizacija vrednosti koordinata se naziva *uzorkovanje* a digitalizacija amplitude se naziva *kvantizacija* (Gonzalez & Woods, 2002).



Slika 3.4.1: (a) kontinualna slika projektovana na senzorskom nizu; (b) rezultat uzorkovanja i kvantizacije slike (Gonzalez & Woods, 2002).

Tako dobijena digitalna slika je polazna osnova za dalje korake procesiranja i analize. Većina standardnih funkcija obrade i analize slika se mogu kategorizovati u sledeće četiri kategorije (Levin, 1999; Canada Centre for Remote Sensing, 2007; Khorram, Wiele, Koch, Nelson, & Potts, 2016):

- *Predprocesiranje*: uključuje operacije koje su zahtevane pre glavne analize podataka i ekstrakcije informacija, što može uključivati spajanje i normalizaciju slika sakupljenih pomoću različitih senzora, mozaiking više slika sakupljenih pomoću istog senzora, geometrijsko mapiranje slika na površinu (geotagovanje), geometrijske korekcije slika (resemplovanje kao što su *Nearest neighbour*, *Bilinear interpolation* i *Cubic convolution*), itd.
- *Poboljšanja*: uključuje prikaz slike, od značaja samo prilikom vizuelne interpretacije, i analiza slika, npr. povećanje kontrasta (histogramske operacije), i povećanje oštrine ivica ili detalja na slici, suzbijanje neželjenih detalja slike, uklanjanje šuma, ekstrakcija svojstava i objekata, prepoznavanje obrazaca (primene operacija filtera). Histogramske operacije posmatraju piksele bez obzira gde se nalaze na slici i dodeljuju novu vrednost pikselu pomoću *lookup* tabela, koje se dobijaju iz statistike slike, dok je filtriranje „lokalna operacija“, kojom se računaju nove vrednosti za piksel na osnovu vrednosti piksela u okolini (Tempfli, Kerle,

Huurneman, & Janssen, 2009). Primeri su *Linear contrast stretch*, *Histogram-equalized stretch*, *Spatial filtering*, *Low-pass filter*, *High-pass filters* i *Directional*, ili *Edge detection filter*, itd. Detaljno objašnjenje ovih i drugih filtera i alrgoritama poboljšanja slika može se naći u (Bovik, 2009; Gonzalez & Woods, 2002).

- *Transformacije*: za razliku od poboljšanja slike, transformacija obično uključuje kombinovano procesiranje podataka iz više spektralnih opsega. Aritmetičke operacije se sprovode kako bi se kombinovali ili transformisali originalni opsezi u „nove" slike koje bolje prikazuju ili naglašavaju pojedine osobine na slici, pri čemu mogu i smanjiti broj različitih opsega koji su ulaz u procedure klasifikacije, bez značajnih gubitaka u podacima. Tehnike koje se koriste su npr. *skeptralna racionalizacija* ili *racionalizacija opsega* (eng. *Spectral* ili *Band Ratioing*), *vegetativni indeksi* (eng. *Vegetation Index*) ili Analiza glavnih komponenti (PCA). Najčešći spektralni odnosi koji se koriste uključuju odnos infracrvenog spektra sa crvenim spektrom za detekciju vegetacije, zelenog sa crvenim za mapiranje površinskih voda i močvarnih površina, crvenog sa infracrvenim za mutne vode, itd. Najčešći indeks koji se koristi za mapiranje vegetacije je *vegetativni indeks normalizovane razlike* (eng. *Normalized Difference Vegetation Index*, NDVI) (Khorram, Wiele, Koch, Nelson, & Potts, 2016).

- *Klasifikacija i analiza*: koristi se da identifikuje i klasifikuje delove ili piksele slike, odnosno podatke. Generalno, način analize i ekstrakcije informacija iz daljinskih slika se može grupisati u dve grupe: ekstrakcija informacija na osnovu vizuelne interpretacije slika (npr. od strane ljudskog analitičara) i ekstrakcija informacija na osnovu polu-automatskog prosesiranja od strane računara (npr. klasifikacija digitalnih slika) (Tempfli, Kerle, Huurneman, & Janssen, 2009). Polu-automatsko i automatsko procesiranje su od daleko većeg značaja za nas u nastavku ove disertacije. Klasifikacija slika se obično izvodi na višestrukim skupovima podataka a cilj je dodeljivanje svakog piksela slike određenoj klasi (npr. voda, vrsta šume, kukuruz, pšenica), na osnovu statističkih karakteristika intenziteta i obojenosti piksela. Klasifikacija pokušava da svrsta u određenu klasu svaki piksel na osnovu njegovih spektralnih informacija, što se naziva i Spektralno prepoznavanje obrazaca (eng. *Spectral Pattern Recognition*). Ovde treba napraviti razliku između *spektralnih klasa* (koje su grupe piksela koji su uniformni, ili slični, u odnosu na svoju osvetljenost u datom spektru) i *klasa informacija* (koje su kategorije od interesa za analitičara, koje on pokušava da identifikuje na slici, kao što su različite vrste useva, vrste šuma, sastav stena, itd.). Osnovni cilj je mapiranje spektralnih klasa podataka na klase informacija od značaja, pri

čemu je ovo mapiranje veoma retko realizovano kao 1:1, već često jedna klasa informacija podrazumeva više spektralnih klasa, koje su obično proizvod odstupanja u okviru iste klase podataka, poput senke, gustine ili starosti stabala šume, itd. U ovome se ogleda još jedan doprinos ovog istraživanja jer će pokušati definisati metodologije grupisanja kako bi se što efikasnije realizovalo ovo mapiranje.

Neke od specifičnih tehnika kalibracije i korekcija su (Levin, 1999; Khorram, Wiele, Koch, Nelson, & Potts, 2016; Tempfli, Kerle, Huurneman, & Janssen, 2009):

- *Radiometrijska kalibracija*: normalizacija *digitalnih brojeva* (DN) koji predstavljaju zračenje površine kako bi bili predstavljeni određenim rasponom vrednostima i označavali iste ili približno iste nivoe zračenja (iako npr. prikupljeni različitim senzorima). Tehnike korekcije snimljenih digitalnih brojeva služe da bi podaci bili pogodniji za ekstrakciju informacija, a služe nekoj od sledećih svrha: ispravljanje podataka zbog imperfekcije senzora, korekcija podataka zbog osobenosti pejzaža i atmosferskih smetnji, i poboljšanja slika kako bi one bile pogodnije za vizuelnu interpretaciju.
- *Atmosferske korekcije*: od zračenja do refleksije ili temperature i emisivnosti, ove korekcije su neophodne kako bi se mogle porediti slike iz različitih senzora ili vremenskih trenutaka.
- *Geometrijske korekcije*: ispravljanje geometrijske iskrivljenosti koja nastaje zbog raznih faktora, uključujući perspektivu optike senzora, kretanje sistema skeniranja, kretanje platforme, visine, brzine, reljefa terena, zarkivljenosti i rotacije Zemlje. Ove korekcije kompenzuju ove iskrivljenosti kako bi geometrijski prikaz slike bila što približnija realnom svetu. Dovođenje u vezu pozicje piksela na slici sa odgovarajućom pozicijom na zemlji je svrha *georeferenciranja* slike (Tempfli, Kerle, Huurneman, & Janssen, 2009).

Pojedinačnim detaljima i različitim tehnikama predprocesiranja, poboljšanja i transformacije slika u ovom istraživanju se nećemo detaljnije baviti, već ćemo se fokusirati na samu klasifikaciju slika u svrhe analize. Više detalja o tehnikama predprocesiranja, poboljšanja i transformacije slika se može pronaći u (Bovik, 2009).

Interpretacija vazdušnih fotografija koristi karakteristike poput tonaliteta, teksture, obrazaca, oblika, veličina, i lokacije, za identifikaciju objekata i površina na fotografijama. Nasuprot tome, *fotogrametrija* (eng. *photogrammetry*) koristi vazdušne fotografije za dobijanje pouzdanih prostornih merenja objekata (Khorram, Wiele, Koch, Nelson, & Potts, 2016). *Digitalna obrada slika* (eng. *Digital Image Processing*) je

koncept koji prati veliki broj varijeteta računarskih algoritama i pristupa za vizuelizaciju, poboljšanja i interpretacija daljinski uzorkovanih slika. *Prepoznavanje oblika*, odnosno *modela* ili *obrazaca* (eng. *Pattern Recognition*) se koristi u algoritmima digitalne obrade slika sa ciljem pružanja razumnih odgovora za sve moguće ulaze i za sprovođenje „najverovatnijeg“ uparivanja ulaza, uzimajući u obzir statističke osobine podatka slike (Khorram, Wiele, Koch, Nelson, & Potts, 2016). Ključni proizvodi digitalne obrade slika uključuju tematske mape i bojama obeležene klasifikovane slike koje daju prostorne obrasce pojedinih karakteristika objekata i svojstava, kao što su granice između površina vode i kopna (Khorram, Wiele, Koch, Nelson, & Potts, 2016).

Kada govorimo u slikama prikupljenim pomoću bespilotnih letelica, one uglavnom poseduju veću radiometrijsku homogenost u odnosu na slike prikupljene iz aviona ili satelita, zbog niske visine sa koje se slike prikupljaju (Zhang & Kovacs, 2012). Ipak, i slike prikupljene pomoću bespilotnih letelica sadrže i jedinstvene probleme sa kvalitetom slike. Npr. mala težina letelica često znači manje stabilnu poziciju kamere, što rezultira različitim rezolucijama ili uglovima između slika tokom istog leta. Niska visina može dovesti to većih geometrijskih iskrivljenosti, ali što je još bitnije rezultira većim brojem slika za isto polje, a može doći i do zamućenosti slika. Da bi se to kompenzovalo, često se koristi prekomerno prikupljanje slika, što dovodi do većeg obima podataka (Zhang & Kovacs, 2012). Time mozaiking (sklapanje) slika postaje neophodan i važan korak predprocesiranja, ali takođe su i geometrijske korekcije i ortorektifikacija zahtevane pre nego se slike mogu spojiti zbog malih oblasti preleta i nestabilnosti platforme. Neke od metoda koje su razvijene da adresiraju ove probleme uključuju: ručno georeferenciranje pomoću kontrolnih tačaka na zemlji (eng. *Ground Control Points*, GCP), uparivanje slika i automatsko georeferenciranje pomoću navigacionih podataka (Zhang & Kovacs, 2012). Kada su ortofotografije prikupljene, dalje je potrebno izvući korisne informacije iz njih, npr. za primene u preciznoj poljoprivredi lisni pokrivač, procene različitih indeksa i pokazatelja, kao što je *zeleni lisni pokrivač* (eng. *Green Canopy Cover*, GCC) NDVI, i sl. Međutim, korišćenje slika dobijenih pomoću bespilotnih letelica omogućuju dobijanje ortofoto slika veoma visoke rezolucije, što omogućuje dobijanje GCC i drugih indeksa pomoću drugih tehnika, kao što je računarska vizija i raznih pokazatelja izračunatih iz vidljivog spektra (Ballesteros, Ortega, Hernández, & Moreno, 2014). Ukratko, problemi za primenu slika prikupljenih pomoću bespilotnih letelica su generalno identični onima za primenu slika prikupljenih iz aviona ili satelita. Međutim, za većinu primena, procedure obrade slika moraju biti automatizovane kako bi se slika kao finalni proizvod mogla isporučiti pravovremeno. To je od velikog značaja za daljinsko uzorkovanje pomoću bespilotnih letelica, uzimajući u obzir količine prikupljenih podataka i zahtevana vremena obrade (Zhang & Kovacs, 2012).

3.5 Statističke metode u daljinskom uzorkovanju

Proučavanje geografskih fenomena često zahteva primenu statističkih metoda kako bi se dobili novi uvidi u podatke. (Rogerson, 2001) navodi i kategorizuje brojne aspekte i karakteristike povezane sa problemom prostornih analiza. Iako su sve kategorije relevantne za prostorne statističke analize, među najvažnijim su: problem promenjivih jedinica oblasti, problem ograničavanja, procedure prostornog uzorkovanja i prostorna autokorelacija (Rogerson, 2001).

(Stein, i drugi, 1998) adresiraju sledeće elemente od značaja za uspešnu primenu prostorne statistike za daljinski uzorkovane podatke: (a) obim: skala na kojoj su podaci predstavljeni je data prostornom rezolucijom slika i definiše način na koji su ti podaci integrisani sa drugim izvorima informacija, (b) klasifikacija: daljinski uzorkovani podaci se koriste za klasifikaciju. Savremeni razvoj nauke uključuje podršku klasifikaciji pomoću geostatisike, fuzzy klasifikaciju i Bajesove procedure klasifikacije, (c) uzrokovanje: kada su homogene jedinice identifikovane, prostorno uzorkovanje (na polju) je često neophodno da okarakteriše ove jedinice i da se primeti koje prostorne varijable se menjaju i u kom obimu, i (d) podrška odlučivanju: daljinski uzorkovani podaci se sve češće koriste za podršku u donošenju odluka na različitim nivoima (Stein, i drugi, 1998).

Digitalni podaci koji se odnose na spektralne odzive dobijene daljinskim uzorkovanjem su obimni po svojoj prirodi. To čini da je primena statistike neophodna za analizu daljinski uzorkovanih podataka. Statističke metode se intenzivno koriste u izučavanju performansi senzora, obraci slika, analizi i generisanju tematskih mapa (Deekshatulu, i drugi, 1995). U obradi slika statistički parametri igraju domnantnu ulogu za poboljšanja otklanjanjem šuma i modifikacije histograma, dok prepoznavanje obrazaca (eng. *Pattern Recognition*) igra ključnu ulogu u analiziranju daljinski uzorkovanih podataka. Obrazac se definiše kao primer, šablon ili model i predstavlja bilo koji prepoznatljiv međuodnos podataka, bilo analognih ili digitalnih. Prepoznavanje obrazaca je dvojak zadatak, odnosno čine ga razvoj neke vrste pravila odlučivanja koje bazira na prethodnom znanju (komponenta učenja) i njegova upotreba za donošenja odluka u vezi sa nepoznatim obrascima (komponenta klasifikacije) (Deekshatulu, i drugi, 1995). Generalno su u upotrebi tehnike klasifikacije sa i bez nadzora. Statistički pristupi se široko koriste u analizi tekstura daljinski uzorkovanih podataka. Tekstura je prostorno svojstvo koje je korisno kod identifikacije objekata ili oblasti od značaja. Statistike prvog reda nisu prikladne za opisivanje tekstura i u mnogim metodama analize tekstura, statistike drugog reda se intenzivno koriste (Deekshatulu, i drugi, 1995). Veštačka inteligencija (eng. *Artificial Intelligence*, AI) stiče sve veću popularnost, naročito razvoj ekspertskih sistema u različitim primenama poput istraživanja zemljišta, itd. (Deekshatulu, i drugi, 1995).

Multivarijaciona statistička analiza i matematičko modelovanje su važne tehnike za geološke i ostale studije koje generalno mogu uključivati daljinsko uzorkovanje (Deekshatulu, i drugi, 1995). (Deekshatulu, i drugi, 1995) navode sledeće tehnike:

- *Tehnike klasifikacije*: dodeljivanje tačke u prostoru svojstava (npr. daljinski uzorkovan piksel okarakterisan refleksijom u različitim spektralnim ospezima) određenoj klasi obrasca (eng. *pattern class*).
- *Analiza teksture*: tekstura je važno prostorno svojstvo korisno za identifikaciju objekata ili oblasti na slici. Tekstura je definisana kao struktura sastavljena od većeg broja sličnih pod-elemenata ili obrazaca ili tonalnih primitiva. Analiza tekstura se koristi da opiše i diskriminiše kompleksne regione na slici, koje je lakše okarakterisati statistički nego u detalje.
- *Tehnike uzorkovanja*: su veoma korisne u proceni preciznosti mapa izvedenih iz daljinski uzorkovanih podataka (npr. studijama korišćenja zemljišta, itd.).
- *Multivarijaciona statistička analiza*: tokom istraživanja prirodnih resursa, obično su brojne varijable analizirane. Tehnike statističkog i matematičkog modelovanja se koriste za analizu geonaučnih podataka o prirodnim resursima (npr. minerali, nafta, voda, biljke).
- *Analiza vremenskih serija*: na terenima sa čvrstim stenama, sloj oštećenja vremenskim prilikama je od značaja za istraživanja površinskih voda.
- *Veštačka inteligencija*: je oblast izučavanja koja obuhvata računarske tehnike za izvođenje zadataka koji naizgled zahtevaju inteligenciju kada ih izvode ljudi. Istraživači rade sa statističarima tražeći rešenja za probleme koji uključuju neizvesnost, koja može nastati kada su dostupne informacije nepotpune, nisu pouzdane ili način prikaza je neprecizan.

Automatska ekstrakcija informacija pomoću računara iz daljinski uzorkovanih slika se primenjuje već duže vreme. Podaci koji se koriste u obradi su najčešće multispektralni podaci, i metode statističkog prepoznavanja obrazaca (multivarijaciona klasifikacija) su već široko poznate (Benediktsson, Swain, & Ersoy, 1990). Podaci daljinskog uzorkovanja su često raštrkani u odnosu na prostorno-vremsnki domen geofizičkih procesa, tako da se podaci iz različitih senzora koriste u sprezi kako bi se iskoristila njihova komplementarna pokrivenost. Dalje, podaci su često obimni, sa nekomptibilnom podrškom i često polarizovani. Ovim problemima pristupamo sa statističkog stanovišta, i ciljamo da procenimo tačne ali ne direktno opažene procese, kao i da procenimo neizvesnosti ovih procena (Nguyen, 2009). *Fuzija podataka* je proces kombinovanja informacija iz heterogenih izvora u jednu kompozitnu sliku relevantnog procesa, takvu da je kompozitna slika preciznija i kompletnija u odnosu na

podatke dobijene samo iz pojedinačnih izvora (Nguyen, 2009). Fuzija podataka u kontekstu daljinskog uzorkovanja se najčešće koristi na proces integrisanja podataka prikupljenih na različitim prostornim, spektralnim i/ili temporalnim rezolucijama. Primarni cilj je olakšavanje interpretacije ili klasifikacije na nivu preciznosti koji se ne bi mogao postići korišćenjem samo jednog izvora. Ovo se takođe može smatrati kao umanjenje neizvesnosti u vezi sa inicijalnim izvorom podataka (Khorram, Wiele, Koch, Nelson, & Potts, 2016).

Kod daljinskog uzorkovanja, pored tehnika klasifikacije, se takođe koristi i broj drugih statističkih metoda, kao što su regresija, kanonočka korelaciona analiza, Bajesove mreže uslovnih verovatnoća, itd. (Evans, 1998; Deekshatulu, i drugi, 1995) ali one neće biti u fokusu ovog istraživanja, već date pregledno. Na primer, (Radojičić, Vukmirović, & Glišin, 1997) se bave primenom statstičke analize u obradi slika zvezda, prikupljenih pomoću astronomskih kamera sa CCD čipom. Dva glavna problema koja su se javila su uklanjanje šuma sa sirovih slika, što je rešeno analizom intenziteta raspodele piksela, i otkrivanje nepoznatih sijanja zvezda na prethodno filtriranoj slici, za šta je korišćena regresiona analiza. Dalje, dobar pregled različitih statističkih tehnika i metoda se može pronaći u (Lee, Alchanatis, Yang, Hirafuji, & Moshou, 2010). Dodatni primeri primene dobro poznatih statističkih metoda kod daljinskog uzorkovanja, naročito onih koji uključuju primenu u poljoprivredi i upotrebu različitih vegetativnih indeksa (VI) i koji uključuju različite tehnike obrade prikupljenih slika, se mogu naći u (Peña-Barragan, López-Granados, Jurado-Exposito, & Garcia-Torres, 2010; Gonzalez-Dugo, i drugi, 2013; Hočevar, Širok, Godeša, & Stopar, 2014; Gallego, Craig, Michaelsen, Bossyns, & Fritz, 2009), itd. Na kraju, sveobuhvatan uvod u tehnike daljinskog uzorkovanja, sa svim elementima, može se naći u (Khorram, Wiele, Koch, Nelson, & Potts, 2016; Canada Centre for Remote Sensing, 2007; Levin, 1999)

4 MULTIVARIJACIONA STATISTIČKA ANALIZA

Česte su pojave koje se zasnivaju na većem broju različitih promenljivih, i koje, da bi mogle biti izmerene, zahtevaju posmatranje većeg broja različitih pokazatelja (Bulajić, 2002; Radojičić, 2007). Da bi se definisao i okarakterisao višedimenzionalni koncept analize podataka, koristi se pojam *multivarijacione statističke analize* (Dobrota M. P., 2015). Multivarijaciona analiza je veoma pogodna, čak i neophodna za upotrebu kada se isti entiteti opisuju većim brojem atributa (pokazatelja, indikatora) koji mogu biti međusobno povezani (Bulajić, 2002; Radojičić, 2007; Jeremić, 2012).

Multivarijaciona analiza omogućuje analizu složenih nizova podataka, kada postoji mnogo nezavisnih i zavisnih promenljivih koje su u korelaciji, kako bi se obezbedile što sveobuhvatnije statističke analize (Dobrota M. P., 2015; Radojičić, 2007; Jeremić, 2012). Za multivarijacionu statističku analizu, odgovarajući skupovi podataka se moraju formirati od vrednosti koje odgovaraju broju promenljivih u odnosu na broj entiteta. Oni mogu biti organizovani kao matrice podataka, korelacione matrice, matrice varijansi-kovarijansi, matrica sume kvadrata, kao niz reziduala (Dobrota M. P., 2015).

Kada je cilj objašnjenje prirode neke pojave, često je teško da se kompleksna priroda entiteta sagleda kroz jednu karakteristiku, pa je moguće obuhvatiti različite karakteristike jedne višedimenzionalne pojave, koje se nazivaju promenljive ili indikatori (Dobrota M. P., 2015; Đoković, 2013; Jeremić, 2012). Multivarijaciona analiza ispituje prirodu entiteta istovremenim merenjem više promenljivih za svaki entitet (Vuković, 1977; Vuković, 1987). Prema (Kovačić, 1994), multivarijaciona analiza je skup statističkih metoda koje simultano analiziraju višekriterijumska merenja dobijena za svaki entitet iz skupa koji se posmatra.

Neka su u procesu istraživanja sakupljeni podaci za *i*-ti entitet, *i=1,2,...,n,* o njihovom *j*-tom atributu, *j=1,2,..,p*. Takvi podaci predstavljaju osnovu multivarijacione analize i prikazuju se u vidu matrice podataka (Tabela 4.1). Po redovima su dati entiteti a po kolonama izmereni indikatori (promenljive). Ovakva matrica podataka nema svojstva matrice, već predstavlja uređeni skup podataka koji se analizira. Pretpostavimo da postoji *n* redova (entiteta) i *p* kolona (indikatora). Matrica podataka data je Tabelom 4.1, gde X_{ij} predstavlja vrednost *j*-tog indikatora izmerenog na *i*-tom entitetu.

Tabela 4.1: Matrica podataka

	Ind. 1	Ind. 2	...	Ind. j	...	Ind. p
Entitet 1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}
Entitet 2	X_{21}	X_{22}	...	X_{2j}	...	X_{2p}
...	...	...		...		...
Entitet i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ip}
...	...	...		...		...
Entitet n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}

Multivarijaciona metoda za analizu matrice podataka se bira u zavisnosti od sledećih faktora: vrsta problema, tipovi podataka, karakteristike metode ili ciljevi istraživanja. S obzirom na to da su takve matrice podataka uglavnom velikih dimenzija, veoma je teško zaključivati o međuzavisnosti promenjivih bez odgovarajućih dubljih analiza. Kako bi se takvi problemi prevazišli moguće je koristiti metode multivarijacione analize. Ovim metodama se postiže pojednostavljenje složene strukture posmatrane pojave, a u cilju lakše interpretacije podataka. Metode multivarijacione analize se takođe koriste u procesu zaključivanja, tako što se, na primer, ocenjuje stepen međuzavisnosti promenjivih i/ili testira njihova statistička značajnost (Bulajić, 2002; Dobrota M. P., 2015). Neke od metoda multivarijacione analize se koriste, na primer, ne za testiranje a priori definisanih hipoteza, već za njihovo generisanje.

Metode multivarijacione analize klasifikovane su prema različitim klasifikacionim kriterijumima (Jeremić, 2012; Radojičić, 2007). Moguće ih je klasifikovati prema tome da li su metode orjentisane ka ispitivanju međuzavisnosti indikatora ili međuzavisnosti entiteta. Osnovu prve grupe metoda (indikatori) predstavlja kovarijaciona ili korelaciona matrica. Kod druge grupe metoda (entiteti), porede se entiteti, odnosno definišu se različite mere bliskosti između dva entiteta. Osnovu ovih metoda multivarijacione analize predstavlja matrica odstojanja (Bulajić, 2002; Đoković, 2013; Vuković, 1977; Kovačić, 1994).

Prema drugoj klasifikaciji, metode se dele u dve grupe: metode zavisnosti i metode međuzavisnosti (Dobrota M. P., 2015). Ukoliko se ispituje zavisnost između dva skupa indikatora, gde jedan skup predstavlja zavisne, a drugi nezavisne promenjive, tada se govori o metodama zavisnosti. Ukoliko nema teorijskog osnova za podelu svih promenjivih na dva ovakva podskupa (zavisnih i nezavisnih), tada se koriste metode međuzavisnosti. Treba imati u vidu da metode zavisnosti teže da ocene ili predvide jednu ili više zavisnih promenjivih na osnovu skupa nezavisnih indikatora. Metode međuzavisnosti se ne koriste za predviđanje već se, pomoću njih, prodire u kompleksnu unutrašnju strukturu podataka (Kovačić, 1994).

Metode zavisnosti (Bulajić, 2002; Radojičić, 2001; Kovačić, 1994):

1. *Multivarijaciona regresija* je najpoznatija metoda multivarijacione analize. Razlikuju se dva slučaja: (1) analiza zavisnosti jedne promenjive (zavisna promenjiva) od skupa drugih promenjivih (nezavisne promenjive), metoda analize poznata pod nazivom metod višestruke regresije; (2) skup zavisnih promenjivih sadrži više od jednog člana, što predstavlja opštiji model multivarijacione regresije. Kod oba modela zadatak je ocenjivanje ili predviđanje srednje vrednosti (srednjih vrednosti) zavisne promenljive na bazi poznatih vrednosti nezavisnih promenjivih.
2. *Kanonička korelaciona analiza* se može smatrati uopštenom višestrukom regresionom analizom. Ona analizira linearnu zavisnost između skupa

nezavisnih i skupa zavisnih promenljivih, a prilikom računanja kanoničke korelacije, formiraju se dve linearne kombinacije, jedna za nezavisne, a druga za zavisne promenljive. Koeficijenti ovih linearnih kombinacija određuju se tako da koeficijent korelacije između njih bude maksimalan.

3. *Diskriminaciona analiza* se bavi problemom razdvajanja grupa i raspoređivenjem entiteta u predefinisane grupe. Primena diskriminacione analize omogućava, na osnovu vrednosti skupa nezavisnih promenjivih, identifikaciju indikatora koji je najviše doprineo razdvajanju grupa kao i predviđanje verovatnoće da će entitet pripasti jednoj od grupa.
4. *Multivarijaciona analiza varijanse (MANOVA)* je pogodna metoda kada je cilj ispitivanje uticaja jedne ili više "eksperimentalnih" promenjivih na dve ili više zavisnih promenjivih. Ona predstavlja uopštenje jednodimenzionalne analize varijanse (ANOVA). Osnovni cilj je analiza razlike u grupama, što se ispituje testiranjem hipoteze koja se tiče varijanse efekata grupa dve ili više zavisnih promenjivih.
5. *Logit analiza* se koristi kada je u regresionom modelu zavisna promenjiva dihotomnog tipa (na primer, promenljiva *Pol* sa vrednostima: muško-žensko). Takav model se naziva regresioni model sa kvalitativnom zavisnom promenjivom, gde je zavisna promenjiva, tzv. Logit funkcija (logaritam količnika verovatnoća da će dihotomna zavisna promenjiva uzeti jednu ili drugu vrednost).

Metode međusobne zavisnosti (Radojičić, 2001; Kovačić, 1994):

1. *Analiza glavnih komponenti* je metoda za redukciju većeg broja indikatora na manji broj novih promenjivih (glavne komponente). Najčešće se manjim brojem glavnih komponenata objašnjava veći deo varijanse originalnih promenjivih. Osnovni zadatak jeste konstruisanje linearne kombinacije originalnih indikatora (glavnih komponenata) uz uslov da obuhvate što je moguće više varijanse originalnog skupa indikatora.
2. *Faktorska analiza* je slična metodi glavnih komponenti jer se koristi za redukciju promenljivih na manji broj faktora, međutim originalna promenljiva se iskazuje kao linearna kombinacija faktora uz dodatak greške modela. Na taj način se celokupna kovarijansa ili korelacija objašnjava zajedničkim faktorima, a neobjašnjeni deo se pridružuje grešci (specifičan faktor). Za razliku od glavnih komponenti, gde se objašnjava varijansa, interes faktorske analize je usmeren ka objašnjenju kovarijanse, odnosno onog dela ukupne varijanse koji promenljiva deli sa ostalim iz posmatranog skupa.
3. *Analiza grupisanja* je metoda za redukciju podataka, koje se bazira na entitetima. Ovom analizom se entiteti kombinuju u grupe po sličnosti, koje su

relativno homogene. Zadatak je podela entiteta u manji broj grupa, tako da su oni koji pripadaju jednoj grupi sličniji jedan drugom, u odnosu na one koji pripadaju drugim grupama.

4. *Višedimenziono proporcionalno prikazivanje*. Pripada klasi metoda koji su orjentisani ka objektima, a koristi meru sličnosti, odnosno razlike između njih u cilju njihovog prostornog prikazivanja. Izvedena prostorna reprezentacija sadrži geometriski raspored tačaka na mapi, gde se svaka tačka odnosi na jedan od objekata. Ukoliko se za ovo proporcionalno prikazivanje koristi mera bliskosti dobijena na osnovu merljivih (kvantitativnih) promenjivih nazivu metode dodajemo pridev kvantitativno, a ako smo za računanje mera sličnosti koristili kvalitativne promenjive, tada nazivu metode dodajemo pridev kvalitativno.
5. *Loglinearni modeli* ispituju međusobnu zavisnost kvalitativnih promenjivih koje formiraju višedimenzionalnu tabelu kontigencije. Ukoliko se jedna od promenjivih u tabeli kontigencije može smatrati zavisnom, tada se na osnovu ocenjenih loglinearnih modela mogu izvesti logit modeli.

Za nas je od posebnog značaja prepoznavanje obrazaca (eng. *Pattern Recognition*), gde je analiza podataka usmerena ka prediktivnom modelovanju: na osnovu podataka za učenje (trening), želimo predvideti ponašanje još uvek nepoznatih podataka. Ovaj zadatak se naziva i *učenje* (Jain, 2010). Često se pravi jasna definicija između problema učenja koji su (1) *nadgledani* (*s nadzorom*, eng. *supervised*) (klasifikacija) i (2) *nenadgledani* (*bez nadzora*, eng. *unsupervised*) (grupisanje). Prvi uključuju označene (eng. *labeled*) podatke (uzore učenja sa poznatim oznakama kategorija) a drugi uključuju samo neoznačene podatke. Postoji i rastući interes u pravcu hibridnih metoda, koje se još nazivaju *polunadgledano* učenje (*sa delimičnim nadzorom*), gde oznake postoje samo za mali deo skupa podataka za učenje. Neoznačeni podaci, umesto da se odbacuju, se takođe koriste u procesu učenja (Jain, 2010).

U nastavku ove disertacije ćemo se detaljnije baviti metodama multivarijacione analize koje se generalno koriste u analizi, odnosno klasifikaciji, podataka prikupljenih daljinskim uzorkovanjem. Drugim rečima, klasifikacija podataka je od nas od posebnog značaja jer prestavlja jedan od ciljeva analize podataka daljinskog uzorkovanja. Pregled istorije razvoja rešavanja problema automatske klasifikacije se može naći u (Radojičić, 2001). Jedan način da se opiše značenje grupisanja elemenata (klasifikacije) je pomoću skupova. Pretpostavimo da se pomoću jednog kriterijuma mogu vršiti grupisanja elemenata skupa *S*. Svaka tako obrazovana grupa *A* predstavlja jedan podskup od *S*. Neka je *D* dobijeni skup podskupova od *S*. Ako je unija svih delova jednaka skupu *S*, za *D* kažemo da predstavlja jedno pokriće skupa *S*. Ako su svi podskupovi od *D* neprazni, međusobom disjunktni, a unija im je jednaka skupu *S*, za *D* kažemo da predstavlja jednu podelu skupa *S*. Za delove jedne podele kažemo da predstavljaju klase skupa *S*

(Radojičić, 2001). U nastavku ovog istraživanja je od značaja određivanje klasa podataka daljinskog uzorkovanja i grupisanje podataka (oblasti predstavljene pikselima) u jednu od klasa.

4.1 Analiza grupisanja

Analiza grupisanja ili *klaster analiza* (eng. *Cluster Analysis*) je metoda multivarijacione analize koja se koristi za grupisanje entiteta, tako da su entiteti unutar jedne grupe međusobno slični, a znatno različiti u odnosu na entitete iz drugih grupa (Radojičić, 2001; Dobrota M. P., 2015; Webb & Copsey, 2011). Grupisanje entiteta vrši se na osnovu mera sličnosti koje se definišu na osnovu njihovih atributa. Klaster analiza je formalno proučavanje metoda i algoritama za grupisanje (klasterovanje) objekata prema njihovim izmerenim ili percipiranim suštinskim karakteristikama ili sličnosti (Jain, 2010). Klasterovanje (eng. *Clustering*) je jedna od najčešće korišćenih tehnika eksploratorne analize podataka (Luxburg, 2007). Svrha klasterovanja je grupisanje individuelnih entiteta u populacije i time pronalaženje strukture u podacima, čime je klasterovanje eksploratorno po svojoj prirodi i predstavlja skup metoda istraživanja podataka (Jain, 2010; Webb & Copsey, 2011). Ima dugačku i bogatu istoriju u raznim naučnim oblastima, sa primenama od statistike, informatike, biologije do sociooloških nauka ili psihologije. U praktično svim naučnim oblastima koje se bave empirijskim podacima, istraživači pokušavaju da dobiju prve impresije o svojim podacima pokušavajući da identifikuju grupe „sličnog ponašanja“ u svojim podacima (Luxburg, 2007). Jedan od najpopularnijih i jednostavnih algoritama klasterovanja je K-means, koji je prvi put objavljen 1955. Iako je uveden još pre više od 60 godina i na hiljade algoritama klasterovanja je od tada objavljeno, K-means je i dalje široko u primeni. Ova činjenica govori o poteškoćama u nalaženju algoritama klasterovanja opšte namene, iako savremena povećanja obima i raznovrsnosti podataka zahtevaju napredke u metodologiji automatizovanog razumevanja, obrade i sinteze podataka (Jain, 2010). Eksplozija količine informacija ne kreira samo velike količine podataka već i različitost podataka, i struktuiranih i nestruktuiranih. Većina pristupa klasterovanju ignoriše strukture u objektima koji se klasteruju i koriste reprezentacije pomoću vektora svojstava i za struktuirane i nestruktuirane podatke (Jain, 2010).

Ciljevi analize grupisanja, odnosno klasterovanja, su: *istraživanje fundamentalne strukture*, ako struktura skupa entiteta nije poznata, analizom grupisanja otkrivamo nepoznatu strukturu, istaknuta svojstva, anomalije, generišemo hipoteze (npr. oko broja klastera, itd.), *redukcija i kompresija podataka*, kao metod za orgnizovanje podataka i njihovo sumiranje kroz prototipove klastera (npr. ako su klasteri kompaktni, originalni veliki skup podataka se može svesti na manji broj grupa, bez značajnog gubitka informacija, čime se grupa individua predstavlja jedinstvenim obrascem), *prirodna klasifikacija*, za identifikaciju stepena sličnosti među formama ili organizmima, i *predviđanje* (Anderberg, 1973; Jain, 2010; Webb & Copsey, 2011). Klaster analiza je

vrlo slična diskriminacionoj analizi, kada se diskriminaciona analiza koristi kao sredstvo za klasifikaciju entiteta. Razlika je što su kod diskriminacione analize grupe predefinisane (klase), dok to kod klaster analize nisu (klasteri). Klaster analiza ne koristi oznake kategorija koje označavaju objekte priornim identifikatorima, npr. naziv klase. Odsustvo informacije o kategoriji diferencira klasterovanje (učenje bez nadzora) od klasifikacije ili diskriminacione analize (učenje s nadzorom) (Jain, 2010). Procedure koje koriste imenovane uzorke se nazivaju s nadzorom, dok procedure koje koriste neimenovane uzorke se nazivaju bez nadzora (Duda, Hart, & Stork, 2000). Postoji više razloga za interesovanje za procedure bez nadzora, kao što je to što prikupljanje i obeležavanje velikih skupova obrazaca može biti skupo, pa može biti neophodno da se, nakon što se klasifikator istrenira nad malim skupom podataka, dalje pusti da radi bez nadzora na velikim neobeleženim skupovima, ili se može pristupiti u suprotnom smeru – prvo grupisati velike količine neobeleženih podataka a onda koristiti tehnike s nadzorom da se obeležavaju tako pronađene grupe. Dodatno, u mnogim primenama karakteristike obrazaca se mogu postepeno menjati u vremenu, što su promene koje bi se mogle pratiti tehnikama bez nadzora, pre nego se primene tehnike s nadzorom, itd. (Duda, Hart, & Stork, 2000). Dakle metode klasterovanja se koriste u istraživanju podataka i za obezbeđivanje prototipova za klasifikatore s nadzorom (Webb & Copsey, 2011).

Metode grupisanja entiteta (Bogosavljević, 1985) mogu se podeliti u dve grupe:

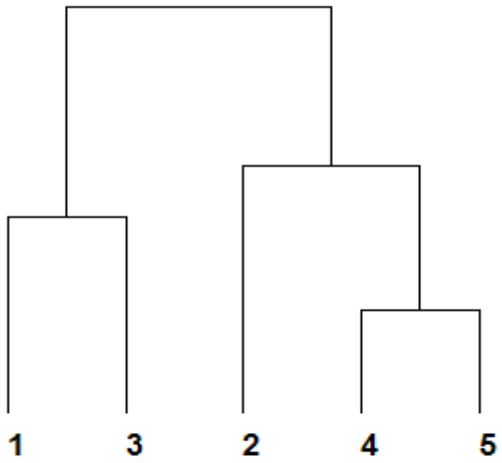
- *Hijerarhijske metode*: Aglomerativne, Dividivne, Preklapajuće, Fazi;
- *Nehijerarhijske metode*: K-means algoritam, Frogy algoritam, itd.

Kod *hijerarhijskih metoda*, postupak je iterativni proces u kome se entiteti spajaju u grupe, pa se u narednoj iteraciji spajaju entiteti i prethodno formirane grupe, tako da se jednom formirane grupe zapravo samo proširuju novim entitetima, bez mogućnosti prelaska entiteta iz jedne grupe u drugu (Bogosavljević, 1985; Radojičić, 2007). S druge strane, *nehijerarhijske metode* dozvoljavaju mogućnost prelaska (Radojičić, 2001).

Hijerarhijske metode grupisanja se mogu svrstati u dve kategorije prema tome da li su zasnovane na iterativnom spajanju (aglomerativne metode) ili deljenju grupa i entiteta (dividivne metode). Prva kategorija polazi od pojedinačnih entiteta koje udružuje u grupe, a zatim u sledećim iteracijama spaja prethodno formirane grupe i pojedinačne entitete, s tim da jednom formirane grupe ostaju zajedno, tj. nema mogućnosti prelaska entiteta iz jedne u drugu grupu (Vukmirović, 1992). Metode koje spadaju u ovu grupu se zajednički nazivaju *hijerarhijske metode udruživanja*. Na početku postupka hijerarhijskog udruživanja postoji n grupa sa po jednim entitetom, a nadalje se postupak odvija na sledeći način: (1) na osnovu matrice odstojanja biraju se dve najbliže grupe i udružuju u novu grupu (neka su r-ta i s-ta grupa udružene u novu grupu t), (2) određuje se odstojanje ostalih grupa i novoformirane grupe, i ponovo računa matrica odstojanja, i (3) prethodna dva koraka se ponavljaju (n-1) put sve dok se ne formira jedna grupa.

Druga kategorija metoda radi u suprotnom smeru. One polaze od jedne grupe u kojoj se nalaze svi entiteti, i iz nje izdvajaju po jedan entitet ili grupu sve dok se ne formira onoliko grupa koliko ima pojedinačnih entiteta. Ove metode se nazivaju se *hijerarhijske metode deobe*. Najpopularnije metode grupisanja pripadaju hijerarhijskim metodama udruživanja (Vuković, 1987).

Grupisanje entiteta zasnovano je na atributima koje merimo. Uzmimo, na primer, dva atributa; tada za grafički prikaz podataka u cilju određivanja grupa možemo uzeti dijagram rasturanja. Na osnovu dijagrama rasturanja možemo definisati grupe kao oblasti u dvodimenzionalnom prostoru sa velikom gustinom tačaka, koje su razdvojene od drugih oblasti delovima sa malom gustinom tačaka. Ako definišemo takve grupe na osnovu kriterijuma bliskosti, možemo smatrati da entiteti unutar grupe treba da budu bliži jedni drugima, nego entitetima u drugim grupama (Dobrota M. P., 2015; Đoković, 2013; Kovačić, 1994). Osim grafičkih metoda, gde se subjektivnom procenom formiraju grupe, postoje i analitičke metode, gde se prema nizu pravila vrši grupisanje entiteta. U osnovi svih ovih metoda je matrica podataka (Tabela 4.1), tj. matrica sa n redova (entiteta) i p kolona (promenljivih). Na osnovu (n x p) matrice podataka formiramo (n x n) matricu bliskosti (sličnosti) P, čiji elementi mere nivo sličnosti ili razlike između svih parova profila iz matrice podataka. Na primer, element p_{rs} (r,s = 1,2,...,n) je mera bliskosti između r-tog i s-tog entiteta.



Slika 4.1.1: Primer izgleda dendrograma.

Nakon što se formira matrica bliskosti, bira se metoda grupisanja. Metoda grupisanja je skup pravila pridruživanja entiteta u grupe na osnovu mere bliskosti, i između njih treba izabrati onu koja najviše odgovara posmatranom problemu (Vukmirović, 1992). Kao što je spomenuto, najčešće se koriste hijerarhijske metode grupisanja. Na kraju se dobija hijerarhijska struktura skupa entiteta koja se zove

hijerarhijsko drvo ili *dendrogram* (Bulajić, 2002), čiji je primer dat na Slici 4.1.1. Generalno, sve metode grupisanja mogu rezultirati podelom skupa podataka u međusobno isključive (nepreklapajuće) grupe. Ipak, bitno je naglasiti da različite metode će često rezultirati različitim grupisanjem, s obzirom da implicitno nameću srukturu nad podacima, kao i da će tehnike rezultirati grupisanjem čak i onda kada ne postoji „prirodno“ grupisanje podataka (Webb & Copsey, 2011).

4.1.1 Mere sličnosti i razlike između entiteta

U cilju grupisanja entiteta, mera bliskosti iskazuje međusobne razlike i sličnosti između dva entiteta. Mera bliskosti p_{rs} predstavlja *meru razlike entiteta r i s* ako su ispunjeni sledeći uslovi:

- Uslov ne-negativnosti: $p_{rs} > 0$ ako se entiteti *r* i *s* razlikuju, a $p_{rs} = 0$ ako i samo ako su entiteti *r* i *s* identični;
- Uslov simetričnosti: $p_{rs} = p_{sr}$;
- Uslov triangularnosti: $p_{rs} \leq p_{rq} + p_{qs}$, za sve *r*, *s* i *q*.

Mera bliskosti p_{rs} predstavlja *meru sličnosti entiteta r i s* ako su ispunjeni sledeći uslovi:

- Uslov normiranosti: $0 \leq p_{rs} \leq 1$, za sve *r* i *s*;
- $p_{rs} = 1$, samo ako su entiteti identični;
- Uslov simetričnosti: $p_{rs} = p_{sr}$ (Bulajić, 2002; Đoković, 2013; Radojičić, 2007).

Najpoznatija mera razlike je *Euklidsko odstojanje*, mera odstojanja na bazi kvantitativnih promenljivih. Na primer, ako su x_r i x_s *r*-ti i *s*-ti red matrice podataka tada je kvadrat Euklidskog odstojanja:

$$d_{rs}^2 = \sum_{j=1}^{p} \left(x_{rj} - x_{sj} \right)^2$$

Euklidsko odstojanje je specijalan slučaj *Minkowskog odstojanja* koje glasi:

$$M = \left[\sum_{j=1}^{p} \left| x_{rj} - x_{sj} \right|^{\lambda} \right]^{1/\lambda}$$

Odstojanje Minkowskog se svodi na Euklidsko odstojanje kada je $\lambda = 2$ (Kovačić, 1994). Na osnovu odstojanja Minkowskog se takođe može definisati i "odstojanje gradskog bloka" tj. *Menhetn odstojanje* koje se dobija za $\lambda = 1$. Što je λ veće, to je mera odstojanja manje osetljiva na prisustvo nestandardnih opservacija.

Mahalanobisovo odstojanje je odstojanje koje vodi računa i o kovarijacionoj strukturi podataka. Ono eliminiše efekat korelisanosti promenljivih, tako da ga ne treba koristiti kada je u analizi taj efekat bitan za razlikovanje entiteta (Radojičić, 2007).

Merenje bliskosti entiteta može se bazirati i na merama sličnosti. Ako se posmatraju dva entiteta *r* i *s* u *p*-dimenzionalnom prostoru, može se uzeti ugao između dva vektora x_r i x_s, kako bi se izmerio stepen sličnosti između tih entiteta. Što je taj ugao manji, entiteti *r* i *s* su sličniji među sobom, tako da se kao mera sličnosti može koristiti kosinus ugla:

$$c_{rs} = \frac{\sum_{j=1}^{p} x_{rj} x_{sj}}{\sqrt{\sum_{j=1}^{p} x_{rj}^2 \sum_{j=1}^{p} x_{sj}^2}}$$

Pošto je u gornjem izrazu kvadrat dužine vektora, c_{rs} ne zavisi od dužine dva vektora. Mera sličnosti c_{rs} se zove *konusni koeficijent* ili *koeficijent podudarnosti* (Jeremić, 2012).

Ako je dat skup entiteta x_1, ... x_n, i notacija sličnosti s_{ij}, između svih parova entiteta, intuitivni cilj klasterovanja je deljenje entiteta u nekoliko grupa, pri čemu su oni koji pripadaju jednoj grupi slični, ali su različiti sa onima iz drugih grupa. Ako je jedini podatak kojim raspolažemo sličnost između entiteta, podaci se takođe mogu predstaviti u formi grafa sličnosti $G = (V, E)$, gde svaki čvor v_i predstavlja entitet x_i, dok su ivice date težinama s_{ij}. Time se problem klasterovanja može predstaviti kao problem adekvatnog particionisanja takvog grafa (Luxburg, 2007).

4.1.2 Mere sličnosti i razlike među grupama

Način merenja sličnosti ili razlike između grupa je upravo osobina po kojoj se metode analize grupisanja razlikuju. Najpoznatije mere sličnosti i razlike su:

- Jednostruko povezivanje,
- Potpuno povezivanje,
- Prosečno povezivanje,
- Metod centroida,
- Wardov metod (metod minimalne sume kvadrata).

Prema jednostrukom povezivanju, odstojanje između dve grupe je najmanje odstojanje parova entiteta iz posmatranih grupa; prema potpunom povezivanju, odstojanje između dve grupe je najveće odstojanje između parova entiteta; prema prosečnom povezivanju, odstojanje između dve grupe se određuje na osnovu prosečnog odstojanja svih parova entiteta (Kovačić, 1994).

Uzmimo dve grupe entiteta (*r* i *s*) koje sadrže n_r i n_s entiteta; označimo opservacije *p* promenljivih za *n* entiteta u *r*-toj grupi sa x_{rjm} ($j=1,2,...,p$; $m=1,2,...n_r$), i za n_s entiteta u *s*-toj grupi sa x_{sjm}; označimo centroide *r*-te grupe sa $x'_r = [x_{r1*},x_{r2*},...x_{rp*}]$ i centroide *s*-te grupe sa $x'_s = [x_{s1*},x_{s2*},...x_{sp*}]$; tada prvu meru odstojanja između ove dve grupe možemo definisati kao:

$$d_{rs}^2 = \sum_{j=1}^{p} (x_{rj*} - x_{sj*})^2$$

Pošto postoji ukupno ($n_r n_s$) odstojanja između dve grupe, druga mera odstojanja definiše meru ukupnog odstojanja između dve grupe kao $n_r n_s d_{rs}^2$, a prosečno rastojanje je $n_r n_s d_{rs}^2 / (n_r + n_s)$. Može se pokazati da je ova mera odstojanja između grupa ekvivalentna promeni u sumi kvadrata unutar grupa do koje je došlo zbog udruživanja *r*-te i *s*-te grupe (Bulajić, 2002; Jeremić, 2012).

Suma kvadrata odstupanja opservacija od svoje sredine tj. suma kvadrata unutar grupe, se za *r*-tu grupu definiše kao

$$SKW_r = \sum_{m=1}^{n_r} \sum_{j=1}^{p} \left(x_{rjm} - \bar{x}_{rj*} \right)^2$$

dok je za *s*-tu grupu

$$SKW_s = \sum_{m=1}^{n_s} \sum_{j=1}^{p} \left(x_{sjm} - c_{sj*} \right)^2$$

Kada se ove dve grupe udruže, dobija se kombinovana grupa (na primer *t*). Ako se posmatraju odstupanja opservacija grupe *t* od novog centroida $x'_t = [x_{t1*},x_{t2*},...x_{tp*}]$ dobija se nova suma kvadrata unutar *t*-te grupe

$$SKW_t = \sum_{m=1}^{n_r+n_s} \sum_{j=1}^{p} \left(x_{tjm} - \bar{x}_{tj*} \right)^2$$

Zbog udruživanja *r*-te i *s*-te grupe dolazi do povećanja ukupne sume kvadrata unutar grupe koje je dato izrazom

$$SKW_t - \left(SKW_r + SKW_s \right)$$

i ekvivalentno je prosečnom odstojanju između grupa ($n_r n_s d_{rs}^2/(n_r+n_s)$). Do ove relacije se dolazi ako se uspostavi veza između analize varijanse i određivanja odstojanja između grupa (Radojičić, 2007). Ukupna suma kvadrata unutar kombinovane grupe *t* (SKW_t) je kao ukupna suma kvadrata u analizi varijanse. Ona se sastoji iz dva dela: suma kvadrata unutar grupa (SKW_r+SKW_s) i suma kvadrata između grupa (SKB_t) do koje se može doći ili izrazom $SKW_t - \left(SKW_r + SKW_s\right)$, ili direktno:

$$SKB_t = \sum_{j=1}^{p}\left[n_r \left(\overline{x}_{rj*} - \overline{x}_{tj*}\right)^2 + n_s \left(\overline{x}_{sj*} - \overline{x}_{tj*}\right)^2 \right]$$

$$SKB_t = \frac{n_r n_s}{(n_r + n_s)} \sum_{j=1}^{p} \left(\overline{x}_{rj*} - \overline{x}_{sj*}\right)^2$$

$$SKB_t = \frac{n_r n_s}{(n_r + n_s)} d_{rs}^2$$

Druga mera odstojanja između grupa je ekvivalentna sumi kvadrata između grupa, tj. priraštaju u sumi kvadrata unutar grupa do koga je došlo udruživanjem *r*-te i *s*-te grupe. Ovakva mera odstojanja predstavlja osnovu Wardove metode hijerarhijskog udruživanja (Radojičić, 2007).

Nakon formiranja nove grupe potrebno je izračunati i odstojanja te grupe i ostalih grupa:

$$d_{tu}^2 = \alpha_r d_{ru}^2 + \alpha_s d_{su}^2 + \beta d_{rs}^2 + \gamma \left| d_{ru}^2 - d_{su}^2 \right|$$

gde je *t* novoformirana grupa, *u* jedna od ostalih grupa (različita od *r* i *s*), a α_r, α_s, β i γ su koeficijenti koji zavise od toga koji se metod udruživanja koristi. U datom izrazu korišćen je kvadrat Euklidskog odstojanja, što je neophodno samo ako se koristi metod centroida ili Wardov metod (Radojičić, 2007). Za ostale metode može se koristiti i neka druga mera odstojanja.

Vrednosti parametara α_r, α_s, β i γ se menjaju u zavisnosti od korišćene mere odstojanja između grupa (Bulajić, 2002):

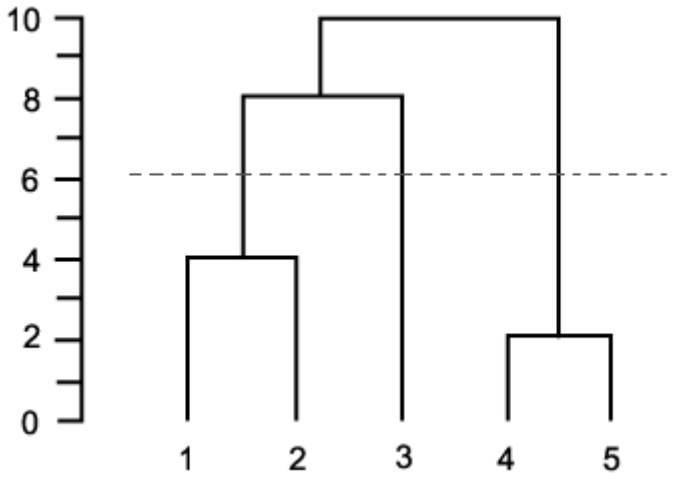
- Jednostruko povezivanje: $\alpha_r = \alpha_s = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$
- Potpuno povezivanje: $\alpha_r = \alpha_s = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$
- Prosečno povezivanje: $\alpha_r = \frac{n_r}{n_r + n_s}, \alpha_s = \frac{n_s}{n_r + n_s}, \beta = \gamma = 0$
- Metod centroida: $\alpha_r = \frac{n_r}{n_r + n_s}, \alpha_s = \frac{n_s}{n_r + n_s}, \beta = -\frac{n_r n_s}{(n_r + n_s)^2}, \gamma = 0$

- Wardov metod: $\alpha_r = \frac{n_r + n_u}{n_t + n_u}, \alpha_s = \frac{n_s + n_u}{n_t + n_u}, \beta = -\frac{n_u}{n_t + n_u}, \gamma = 0$

4.1.3 Određivanje broja grupa (klastera)

Automatsko određivanje broja klastera je jedan od najtežih problema u klasterovanju podataka. Većina metoda za automatsko određivanje broja klastera prebacuju to u problem odabira modela (Jain, 2010). Naime, u većini algoritama automatskog klasterovanja, broj klastera mora biti određen unapred (Mok, Huang, Kwok, & Au, 2012). Uprkos brojnim studijama klasterovanja, i dalje dva ključna pitanja ostaju nerešena: (1) kako automatski odrediti broj klastera, i (2) kako efektivno obaviti klasterovanje uzimajući u obzir raštrkane i šumovite podatke. Istraživači su uveliko okupirani prvim problemom, određivanje broja klastera automatski (Xiang & Gong, 2008). Za odabir broja klastera *k*, različite, manje ili više uspešne metode su osmišljene. Kada je za osnovni model malo ili nimalo pretpostavki dostupno, veliki broj različitih indikatora može biti korišćen za izbor broja klastera. Primeri koji se mogu naći idu od ad hoc mera, poput odnosa između sličnosti unutar klastera i između klastera, preko informaciono-teoretskih kriterijuma, gap statistika, do stabilnosnih pristupa (Dobrota, Delibašić, & Delias, 2016).

Jedan od metoda koji se može primeniti definiše da je moguće, na osnovu dendrograma, formirati *izvedenu matricu odstojanja*, tako što se svim parovima entiteta iz dve različite grupe koje se udružuju u jednu pripisuje ista vrednost odstojanja, ona pri kojoj su udruženi u dve grupe. Međusobnim poređenjem odgovarajućih elemenata originalne i izvedene matrice odstojanja može se utvrditi u kom stepenu formirane grupe predstavljaju dobro rešenje problema grupisanja (Đoković, 2013; Jeremić, 2012; Kovačić, 1994). Da bi se odredio broj grupa, grafički prikaz hijerarhijskog grupisanja, odnosno dendrogram, može se „preseći“ na određenoj visini izborom željenog broja grupa (Slika 4.1.2). Time se dolazi do jednog od mogućih rešenja. Problem izbora broja grupa se može rešiti praćenjem vrednosti mere odstojanja pri kojoj se dve grupe udružuju u jednu. Krećući se od prvog ka *n-1* koraku, vrednost mere odstojanja će rasti, ali u početku sporije, a kasnije brže tj. eksponencijalno. Ako se u okolini očekivanog broja grupa u određenom koraku zabeleži velika promena vrednosti mere odstojanja između grupa (Slika 4.1.2), tada se broj grupa koji je prethodio definisanom koraku proglašava optimalnim.



Slika 4.1.2: Presek dendrograma i podela na odgovarajući broj grupa.

4.2 Bajesova teorija odlučivanja

Bajesova teorija odlučivanja je fundamentalan statistički pristup problemu klasifikacije. Pristup bazira na kvantifikaciji kompromisa između različitih odluka klasifikacije pomoću verovatnoće i cene ili napora koji se javljaju tokom odlučivanja. Ona predpostavlja da je problem odlučivanja postavljen u probabilističkom kontekstu, i da su sve relevantne verovatnoće poznate (Duda, Hart, & Stork, 2000).

Ako sa ω označimo stanje prirode, sa $\omega = \omega_1$ možemo označiti jednu a $\omega = \omega_2$ drugu odluku (npr. kod automatske klasifikacije predmeta kao predmet A ili predmet B, pomoću automatskih merenja). Možemo pretpostaviti da postoje neke *priorne verovatnoće* (ili jednostavnije *priori*) $P(\omega_1)$ da je sledeći predmet A ili priorna verovatnoća $P(\omega_2)$ da je B. Smatraćemo x da je kontinualna slučajna promenjiva, koja je npr. rezultat automatskog očitavanja svojstva predmeta, kao što je osvetljenost ili težina, čija distribucija zavisi od stanja prirode, izražena kao $p(x|\omega_1)$. Ovo je funkcija gustine verovatnoće za x kada je dato da je stanje prirode ω_1. Onda razlika između $p(x|\omega_1)$ i $p(x|\omega_2)$ opisuje različitost tog svojstva između dve populacije, predmeta A i B.

Predpostavimo da su nam je poznato i jedno i drugo, i priorne verovatnoće $P(\omega_j)$ i uslovne verovatnoće $p(x|\omega_j)$. Predpostavimo dalje da merimo svojstvo predmeta (npr. osvetljenost ili težinu) i ustanovimo da je njegova vrednost x. Kako ovo merenje utiče na naš stav u pogledu stvarnog stanja prirode – tj. kategorije predmeta? Primetimo prvo da se (zbirna) gustina verovatnoće nalaženja predmeta iz kategorije ω_j koji ima svojstvo x može zapisati na dva načina: $p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$. Sređivanje ovog izraza nas dovodi do odgovora na naše pitanje, što se naziva *Bajesovom formulom*:

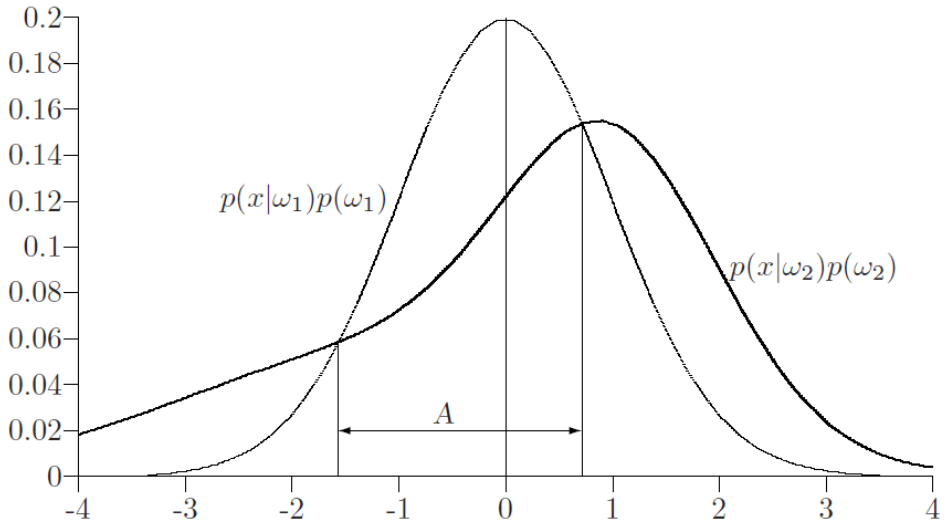
$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},$$

gde u slučaju dve kategorije

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j).$$

Bajesova formula pokazuje da posmatrajući vrednost x možemo konvertovati priornu verovatnoću $P(\omega_j)$ u *posteriornu verovatnoću* (ili *posterior*) $P(\omega_j|x)$ – verovatnoću da je stanje prirode ω_j ako je izmerena vrednost svojstva x. Verovatnoću $p(x|\omega_j)$ nazivamo *izglednost* (eng. *likelihood*) ω_j u odnosu na x (termin izabran da ukaže da je kategorija ω_j za koju je $p(x|\omega_j)$ veliko „izglednija" da bude tačna kategorija). Može se primetiti da je proizvod izglednosti i priorne verovatnoće taj koji je najvažniji za određivanje posteriorne verovatnoće, dok se $p(x)$ može smatrati samo faktorom za skaliranje koji obezbeđuje da je suma posteriornih verovatnoća jednaka 1. U mnogim praktičnim primenama komponente x su binarne ili uopšte celobrojne vrednosti, tako da x može uzeti samo jednu od m diskretnih vrednosti $v_1,...,v_m$. U tom slučaju funkcija gustine verovatnoće $p(x|\omega_j)$ postaje singularna i integrali se moraju zameniti odgovarajućim sumama. Bajesova formula onda uključuje verovatnoće umesto gustina verovatnoće.

Bajesovo pravilo odlučivanja sada nalaže da se minimizira verovatnoća greške na sledeći način: odlučiti ω_1 ako je $P(\omega_1|x) > P(\omega_2|x)$, u suprotnom odlučiti ω_2. Primer je skiciran na Slici 4.2.1.



Slika 4.2.1: $p(x|\omega_i)p(\omega_i)$ za klase (kategorije) ω_1 i ω_2: u oblasti A x je dodeljen klasi ω_1.

Predpostavimo da smo osmotrili određeno x i da razmišljamo o preuzimanju akcije α_i. Ako je stvarno stanje prirode ω_j, možemo reći da ćemo izazvati „trošak“ ili „gubitak“ $\lambda(\alpha_i|\omega_j)$. Pošto je $P(\omega_j|x)$ verovatnoća da je stvarno stanje prirode ω_j, očekivani gubitak vezan za preuzimanje akcije α_i je:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x).$$

U terminologiji teorije odlučivanja, očekivani gubitak se naziva rizikom a $R(\alpha_i|\mathrm{x})$ se naziva uslovnim rizikom. Kada god naiđemo na određenu opservaciju x, možemo minimizirati očekivani gubitak odlučivanja birajući onu akciju koja minimizira uslovni rizik. Ova *Bajesova procedura odlučivanja* zapravo daje optimalne performanse za ukupni rizik. Formalno govoreći, naš problem je nalaženje pravila odlučivanja u odnosu na $P(\omega_j)$ koje minimizira ukupan rizik, odnosno potrebno je izabrati akciju α_i za koju je $R(\alpha_i|\mathrm{x})$ minimalno, odnosno:

$$\alpha^* = \arg\min_i R(\alpha_i|\mathrm{x}).$$

U Bajesovom slučaju može se uočiti da je tipičan efekat opserviranja dodatnih uzoraka izoštravanje posteriorne funkcije verovatnoće, sa maksimum u blizini tačnih vrednosti parametara. Ovaj fenomen je poznat kao *Bajesovo učenje*. U svakom slučaju koriste se posteriorne gustine kao pravilo klasifikacije. Nama je od značaja i uočiti razliku između učenja sa i bez nadzora. U oba slučaja, uzorci x se smatraju prikupljenim birajući stanje prirode ω_i sa verovatnoćom $P(\omega_i)$ a zatim nezavisno birajući x prema zakonu verovatnoće $p(\mathrm{x}|\omega_i)$. Razlika je u tome da kod nadgledanog učenja su nam poznata stanja prirode (imenovane klase) za svaki uzorak, dok kod nenadgledanog učenja nisu. (Duda, Hart, & Stork, 2000; Radojičić, 2001; Webb & Copsey, 2011)

4.2.1 Parametarske tehnike procene gustine

U prethodnom delu smo opisali osnovnu teoriju klasifikacije entiteta. Sve informacije u vezi sa funkcijom gustine $p(x|\omega_i)$ su smatrane poznatima. U praksi te informacije su često nedostupne ili samo delimično poznate. Prema tome, sledeće pitanje koje se postavlja je procena samih funkcija gustine. Ako pretpostavimo neku parametarsku formu za raspodelu, možda dobijena teoretskim razmatranjima, ili procenom problemskog domena, onda se se problem može svesti na procenu konačnog skupa takvih parametara. Često se parametarska forma bira zbog pogodnosti (Webb & Copsey, 2011).

Procena gustine $p(\mathrm{x}|\omega_j)$ bazira na uzorku opservacija $D_j = \{x_1^j, \dots, x_{n_j}^j\}$ $\left(x_i^j \in R^d\right)$ iz klase ω_j. U ovom poglavlju razmatraćemo parametarske pristupe proceni gustine. U takvom parametarskom pristupu, smatraćemo da su uslovne verovatnoće koje zavise od klase ω_j poznate forme ali imaju nepoznati parametar, ili skup parametara, θ_j, i to ćemo

zapisati kao $p(\mathrm{x}|\theta_j)$. Alternativni, neparametarski pristup proceni gustine, koji ćemo kasnije razmatrati, ne pretpostavlja jednostavnu funkcionalnu formu za gustinu. Ovde ćemo razmatrati dva pristupa proceni parametara kod parametarskih klasno-zavisnih gustina, pristup procene (eng. *estimative approach*) i pristup predviđanja ili Bajesov pristup (eng. *Bayesian approach*) (Webb & Copsey, 2011).

4.2.1.1 Procena maksimalne izglednosti

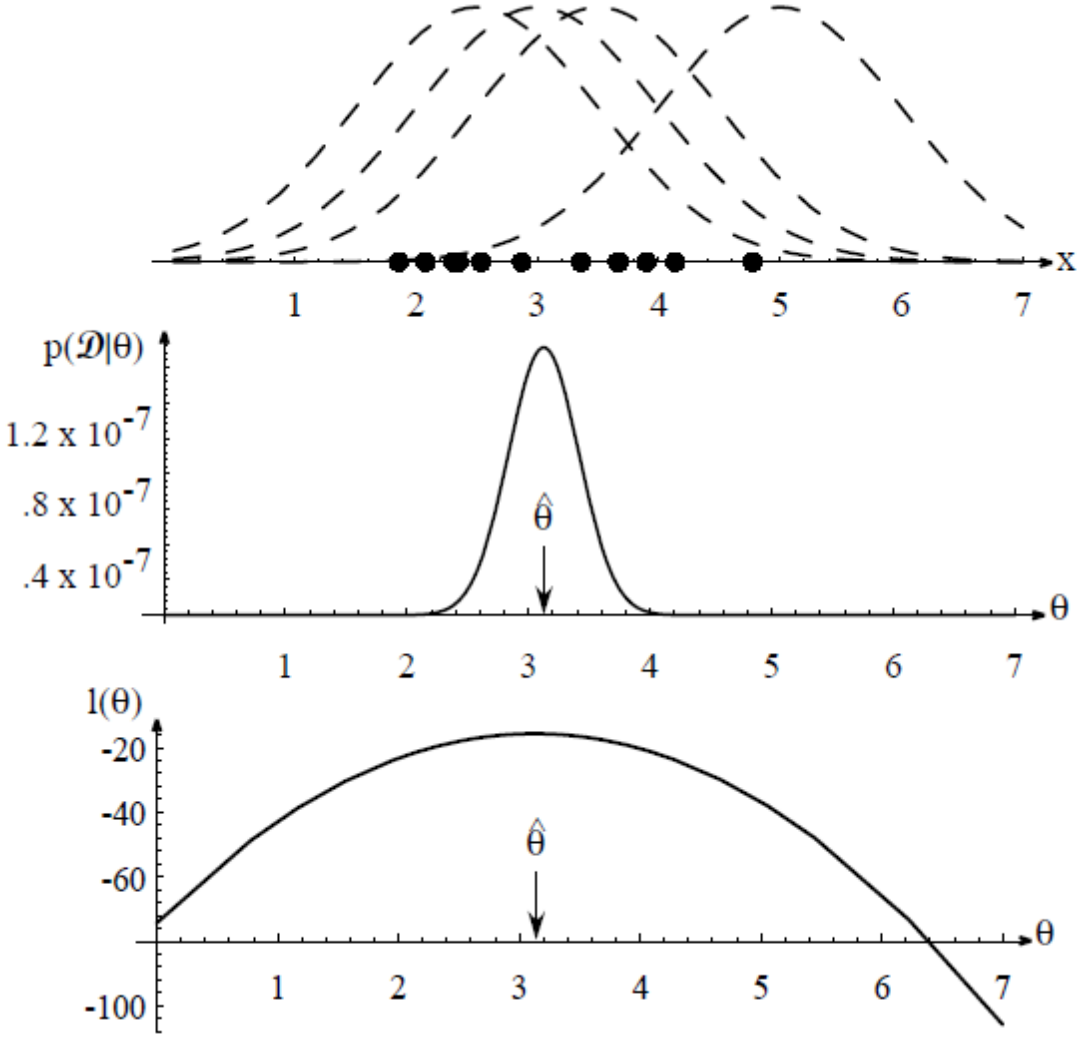
Procena maksimalne izglednosti (eng. *Maximum Likelihood Estimation*) metode imaju nekoliko pozitivnih karakteristika. Prvo, gotovo uvek imaju dobra svojstva konvergencije kako veličina uzorka za trening (učenje) raste. Zatim, procena maksimalne izglednosti je često jednostavnija nego alternativne metode, poput Bajesovih tehnika (Duda, Hart, & Stork, 2000).

Pretpostavimo da $p(\mathrm{x}|\omega_j)$ ima poznatu parametarsku formu, i da je tako određena jedino vrednošću parametarskog vektora θ_j. Na primer, možemo imati $p(\mathrm{x}|\omega_j) \sim N(\mu_j, \Sigma_j)$, gde se θ_j sastoji od komponenti μ_j and Σ_j. Da bismo eksplicitno iskazali zavisnost $p(\mathrm{x}|\omega_j)$ od θ_j, možemo zapisati $p(\mathrm{x}|\omega_j)$ kao $p(\mathrm{x}|\omega_j,\theta_j)$. Naš problem je sada da koristimo informacije koji pruža uzorak za trening da dobijemo dobru procenu za nepoznate parametarske vektore θ_1, ..., θ_c vezanih za svaku od kategorija.

Pretpostavimo da D sadrži n uzoraka, x_1, ..., x_n. Onda, pošto su uzorci uzeti nezavisno, imamo

$$p(D|\theta) = \prod_{k=1}^{n} p(\mathrm{x}_k|\theta).$$

Posmatrano kao funkcija od θ, $p(D|\theta)$ se naziva *izglednošću* (eng. *likelihood*) od θ u odnosu na skup uzoraka. *Procena maksimalne izglednosti* od θ je, po definiciji, vrednost $\hat{\theta}$ koja maksimizira $p(D|\theta)$. Intuitivno, ova vrednost odgovara onom θ koje u nekom smislu se najbolje „slaže“ ili „podržava“ posmatrani skup uzoraka za trening (Slika 4.2.2).



Slika 4.2.2: Gornji dijagram prikazuje nekoliko trening tačaka u jednoj dimenziji, za koje se zna ili se pretpostavlja da su Gausove raspodele određene varijanse, ali nepoznate srednje vrednosti. Četiri od konačnog skupa kandidata polazne raspodele su prikazani isprekidanim linijama. Dijagram u sredini prikazuje izglednost $p(D|\theta)$ kao funkciju očekivane vrednosti. Iako smo imali veoma veliki broj trening tačaka, ova izglednost će biti veoma uska. Vrednost koja maksimizira izglednost je označena sa $\hat{\theta}$; ona takođe maksimizira logaritam izglednosti, odnosno *log-izglednost* $l(\theta)$, prikazano na dnu.

Za analitičke svrhe, je često lakše baratati sa logaritmom izglednosti nego sa samom izglednosti. Pošto je logaritam monotono rastuć, $\hat{\theta}$ koji maksimizira log-izglednost takođe maksimizira izglednost. Ako se $p(D|\theta)$ ponaša dobro, diferencijabilna funkcija od θ, $\hat{\theta}$ se može naći metodama diferencijelnog računa. Ako definišemo $l(\theta)$ kao funkciju log-izglednosti

$$l(\theta) = \ln p(D|\theta).$$

Možemo formalno zapisati rešenje kao argument θ koji maksimizira log-izglednost,

$$\hat{\theta} = \arg\max_{\theta} l(\theta),$$

gde je zavisnost od skupa podataka implicitna. Tako, iz prethodne jednačine imamo

$$l(\theta) = \sum_{k=1}^{n} \ln p(\mathrm{x}_k|\theta)$$

i

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln p(\mathrm{x}_k|\theta).$$

Pa se skup nephodnih uslova za *procenu maksimalne izglednost* za θ može dobiti iz skupa p jednačina

$$\nabla_\theta l = 0.$$

Povezana klasa estimatora – *maksimalni a posteriori* estimatori (MAP) nalaze vrednost θ koja maksimizira $l(\theta)p(\theta)$. Tako je estimator maksimalne izglednosti MAP estimator za uniformne ili „ravne“ priore (Duda, Hart, & Stork, 2000).

4.2.1.2 Bajesova procena

Sada ćemo razmotriti *Bajesovu procenu* ili *pristup Bajesovog učenja* problemu klasifikacije. Iako rezultati koje ovaj pristup postiže će biti približno isti onima dobijenim pomoću maksimalne izglednosti, postoji konceptualna razlika: dok kod metoda maksimalne izglednosti smatramo pravi vektor parametara koji tražimo, θ, fiksnim, kod Bajesovog učenja smatramo θ slučajnom promenljivom, i podaci za učenje nam dozvoljavaju da konvertujemo raspodelu ove promenljive u posteriorne gustine verovatnoće (Duda, Hart, & Stork, 2000).

Računanje posteriornih verovatnoća $P(\omega_i|\mathrm{x})$ je u osnovi Bajesove klasifikacije. Bajesova formula nam dozvoljava da računamo ove verovatnoće iz priornih verovatnoća $P(\omega_i)$ i klasno-uslovnih gustina $p(\mathrm{x}|\omega_i)$, ali kako možemo nastaviti dalje ako su ove vrednosti nepoznate? Opšti odgovor na ovo pitanje je da je najbolje što možemo da izračunamo $P(\omega_i|\mathrm{x})$ korišćenjem svih raspoloživih informacija. Deo ovih informacija može biti priorno znanje, kao što je poznavanje funkcionalnih formi za nepoznate gustine i raspone vrednosti nepoznatih parametara. Deo informacija leži u skupu trening uzoraka. Ako ponovo označimo skup uzoraka sa D, onda možemo naglasiti ulogu uzoraka tako što ćemo kao cilj postaviti da računamo posteriorne verovatnoće $P(\omega_i|\mathrm{x},D)$. Iz ovih verovatnoća možemo doći do *Bajesovog klasifikatora*.

Za dati uzorak *D*, Bajesova formula postaje

$$P(\omega_i|\mathrm{x}, D) = \frac{p(\mathrm{x}|\omega_i, D)P(\omega_i|D)}{\sum_{j=1}^{c} p(\mathrm{x}|\omega_j, D)P(\omega_j|D)}.$$

Kao što sugeriše ova jednačina, možemo koristiti informacije koje pružaju trening uzorci kako bismo odredili i klasno-uslovne gustine i priorne verovatnoće (Duda, Hart, & Stork, 2000).

I skoro svim slučajevima, rešenja maksimalne izglednosti i Bajesov su identični u graničnoj vrednosti beskonačnih trening podataka. Ipak, kako su praktični problemi klasifikacije zasnovanina ograničenom skupu podataka za učenje, logično je razmatrati kada se ova dva pristupa mogu razlikovati, i koji bi trebalo koristiti. Postoji više kriterijuma koji utiču na izbor. Jedan je kompleksnost računanja, i tu su metode maksimalne izglednosti obično preferirane s obzirom da praktično zahtevaju samo diferencijalni račun, umesto potencijalno kompleksnih multidimenzionalnih integracija koje zahteva Bajesova procena.To vodi dalje do pitanja lakoće interpretacije. U mnogim slučajevima rešenja maksimalne izglednosti će biti lakše interpretirati i razumeti pošto vraćaju jedinstveni najbolji model iz skupa koje je dao analitičar/projektant.

Prolikom projektovanja klasifikatora nekom od ove dve metode, određujemo posteriorne gustine za svaku kategoriju, i kasifikujemo testne tačke prema maksimalnim posteriorima. Ukoliko postoji pridrućen trošak (gubitak), to se takođe može dodati u model. Postoje tri izvora greške klasifikacije:

- Bajesova ili greška nemogućnosti razlikovanja: greška usled preklapajućih gustina $p(\mathrm{x}|\omega_i)$ za različite vrednosti *i*. Ova greška je nasledna odlika problema i ne može se eliminisati.
- Greška modela: greška usled definisanja pogrešnog modela, koja se može eliminisati ako projektant precizira tačan model koji uključuje onaj koji generiše podatke.
- Greška procene: greška koja dolazi iz činjenice da su parametri procenjeni iz konačnog skupa uzoraka. Ova greška se najbolje umanjuje povećavajući skup trening podataka (Duda, Hart, & Stork, 2000).

4.2.2 Neparametarske tehnike procene gustine

Do sada smo razmatrali učenje sa nazdorom (nadlgedano učenje) pod pretpostavkom da su oblici ili forme funkcija gustine u osnovi poznate. Kada znamo konkretne funkcije i njihove parametre možemo primeniti Bajesovo pravilo ili test odnosa izglednosti i odlučiti kako da klasifikujemo entitet. Kada možemo dati

pretpostavke o formi funkcija gustine, možemo, kako je gor opisano, na osnovu poznatih uzoraka baviti se procenama njihovih parametara.

Međutim, u većini primena prepoznavanja obrazaca i klasifikacije ova pretpostavka je upitna; uobičajene parametarske forme retko odgovaraju gustinama koje se često javljaju u praksi. Konkretno, sve klasične parametarske gustine su unimodalne (imaju jedinstven lokalni maksimum), dok mnogi praktični problemi uključuju multimodalne gustine. Dalje, naša očekivanja da se višedimenzione gustine mogu jednostavno predstaviti kao proizvod jednodimenzionih funkcija su retko ostvarena (Duda, Hart, & Stork, 2000). Drugim rečima, moguće je da ne postoje formalne strukture kojima se mogu opisati i definisati gustine (Webb & Copsey, 2011). Ovde ćemo opisati neke neparametarske procedure koje se mogu koristiti sa proizvoljnim raspodelama i bez pretpostavki da se forme gustina u osnovi poznate.

Postoji nekoliko vrsta neparametarskih metoda od značaja za prepoznavanje obrazaca i klasifikaciju entiteta (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011):

- Metode koje se sastoje od *procedura za procenu funkcija gustine* $p(\mathrm{x}|\omega_j)$ iz trening uzoraka. Ako su ove procene zadovoljavajuće, mogu se koristiti kao stvarne gustine kada se projektuje klasifikator.
- Metode koje se sastoje od *procedura za direktnu procenu posteriornih verovatnoća* $P(\omega_j|\mathrm{x})$. Ovo je usko vezano sa neparametarskim procedurama poput pravila najbližeg suseda, koje zaobilazi procenu verovatnoće i ide direktno na funkciju odlučivanja.
- Metode koje su neparametarske *procedure za transformaciju prostora* svojstva u očekivanju da je moguće primeniti parametarske metode nad tako transformisanim prostorom. Ove metode diskriminacione analize uključuju Fišerovu diskriminacionu (linearnu) funkciju, koja daje važnu vezu između parametarskih i adaptivnih tehnika. Ovim metodama ćemo se baviti posebno u odvojenom poglavlju ove glave.

4.2.2.1 Procena gustine metodom *k*-najbližih suseda

Možemo reći da je verovatnoća P_k da se k entiteta nalaze u prostoru u kome je prosečna verovatnoća prostora P funkcija od k/n, gde je n ukupan broj entiteta. Ako pretpostavimo da je $p(\mathrm{x})$ kontinualna i da je region R dovoljno mali da p ne varira značajno unutar njega, možemo zapisati

$$\int_R p(x')dx' \simeq p(x)V,$$

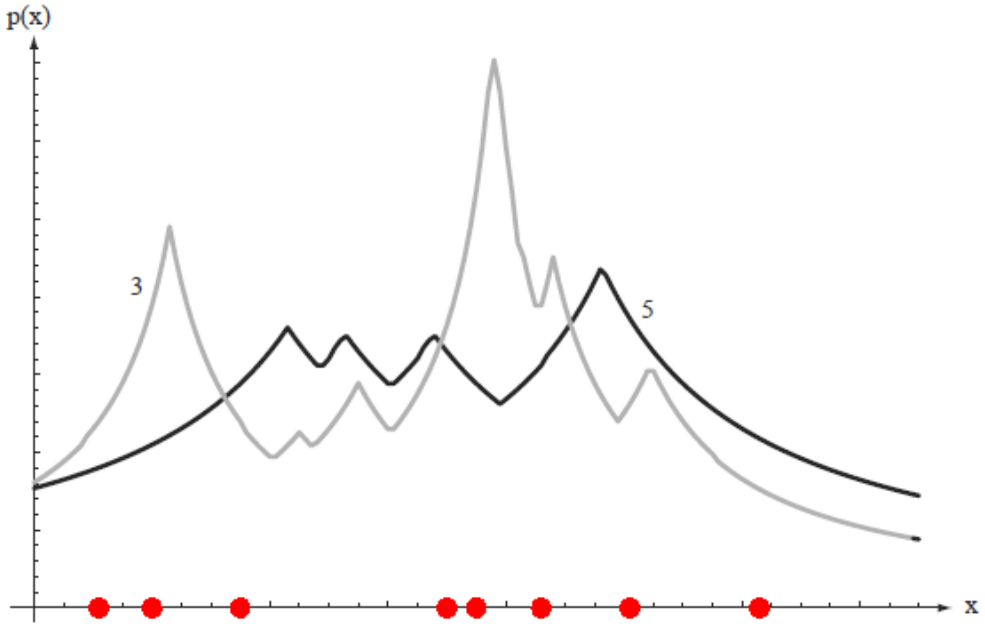
gde je x tačka unutar R i V je prostor (zapremina) obuhvaćen sa R. Dolazimo time do sledeće očigledne procene za p(x) (Duda, Hart, & Stork, 2000):

$$p(x) \simeq \frac{k/n}{V}.$$

Moguće rešenje za problem nalaženja „najbolje“ funkcije je da zapremina ćelije bude u funkciji trening podataka, pre nego neka proizvoljna funkcija ukupnog broja uzoraka. Na primer, da procenimo p(x) iz n trening uzoraka ili prototipova možemo centrirati ćeliju (zapremine V) oko x i proširivati je dok ne obuhvati k_n uzoraka, gde je k_n neka određena funkcija od n. Ovi uzorci su k_n najbližih suseda tački x. Ako je gustina velika oko x, ćelija će biti relativno mala, što daje dobru rezoluciju. Ako je gustina mala, ćelija će porasti velika, ali će prestati rasti čim uđe u oblast veće gustine. U oba slučaja ako uzmemo

$$p_n(x) = \frac{k_n/n}{V_n}$$

hoćemo k_n da ide ka beskonačno kako n ide ka beskonačno, s obzirom da to obezbeđuje da k_n/n bude dobra procena verovatnoće da će tačka upasti u ćeliju zapremine V_n. Međutim, takođe hoćemo da k_n raste dovoljno sporo kako bi veličina ćelije potrebna da obuhvati k_n trening bila smanjena ka nuli. Primer je dat na Slici 4.2.3 (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011).



Slika 4.2.3: osam tačaka u jednoj dimenziji i k-najbližih suseda (eng. *k-nearest-neighbor*) procena gustine, za k=3 i k=5. Možemo primetiti da se prekidi u nagibima procena u opštem slučaju javljaju van pozicija samih tačaka.

Opisana tehnika se može koristiti za procenu posteriornih verovatnoća $P(\omega_i|\mathrm{x})$ iz skupa od n klasifikovanih uzoraka korišćenjem uzoraka za procenu predmetnih gustina. Pretpostavimo da postavimo ćeliju zapremine V oko x i obuhvatimo k uzoraka, od kojih je k_i klasifikovano kao ω_i. Onda je očigledna procena za ukupnu verovatnoću $p(\mathrm{x},\omega_i)$

$$p_n(\mathrm{x}, \omega_i) = \frac{k_i/n}{V},$$

i prema tome razumna procena za $P(\omega_i|\mathrm{x})$ je

$$P_n(\omega_i|\mathrm{x}) = \frac{p_n(\mathrm{x}, \omega_i)}{\sum_{j=1}^{c} p_n(\mathrm{x}, \omega_j)} = \frac{k_i}{k}.$$

Drugim rečima, procene posteriornih verovatnoća da je ω_i stanje prirode je jednostavno udeo uzoraka unutar ćelije koji su klasifikovani kao ω_i. Posledično, za najmanju stopu greške biramo onu kategoriju koja je najviše zastupljena unutar ćelije. Ako postoji dovoljno uzoraka i ako je ćelija dovoljno mala, može se pokazati da će ovo dati rezultate koji se približavaju najbolje mogućim (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011).

Kada se radi o izboru veličine ćelije, možemo koristiti pristup *Parzen-prozora* ili ili pristup k_n*-najbližih suseda*. U prvom slučaju V_n bi bio neka određena funkcija od n, kao što je $V_n = 1/\sqrt{n}$. U drugom slučaju V_n bi se širio dok ne obuhvati neki određeni broj uzoraka, poput $k = \sqrt{n}$. U oba slučaja, kako n ide ka beskonačnosti beskonačan broj uzoraka će upasti u beskonačno malu ćeliju. Činjenica da veličina ćelije može postati proizvoljno mala i ipak sadržati proizvoljno veliki broj uzoraka bi nam omogućila da saznamo nepoznate verovatnoće sa približnom sigurnošću i time dobijemo optimalne performanse. Šta više, može se pokazati da možemo postići uporedive performanse čak i ako baziramo odluke samo na klasama jednog jedinog najbližeg suseda od x (Duda, Hart, & Stork, 2000).

4.2.2.2 Procena gustine metodom histograma

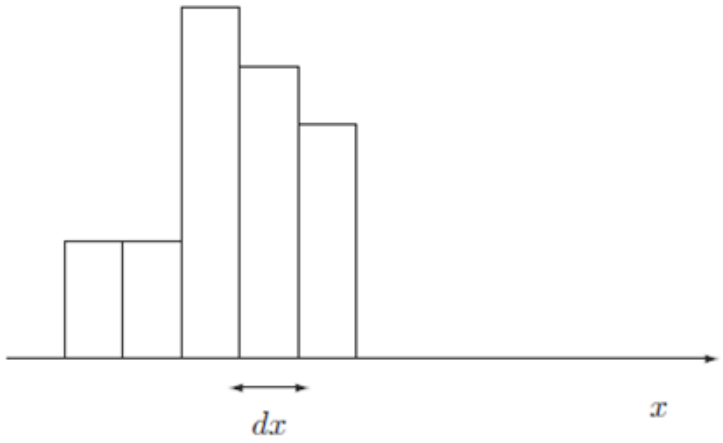
Metod histograma (eng. *Histogram method*) je verovatno najstariji metod procene gustine. To je klasičan pristup po kojem se gustina verovatnoće formira iz skupa uzoraka. U jednoj dimenziji, realna linija je podeljena u određeni broj ćelija identičnih veličina (Slika 4.2.4) i procena gustine u tački x je uzeta kao

$$\hat{p}(x) = \frac{n_j}{\sum_j^N n_j dx}$$

gde je n_j broj uzoraka u ćeliji širine dx koji okružuje tačku x, N je ukupan broj ćelija i dx je veličina ćelije. Ovo se uopštava u

$$\hat{p}(x) = \frac{n_j}{\sum_j n_j dV}$$

za multidimenzionalni prostor posmatranja, gde je dV veličina segmenta j (Webb & Copsey, 2011).



Slika 4.2.4: Histogram.

Iako je ovo veoma jednostavan koncept i lak za implementaciju, i ima prednost što ne zahteva čuvanje tačaka uzoraka, postoji nekoliko problema sa osnovnim pristupom histograma. Kao prvo, retko je praktičan u prostoru većeg broja dimenzija. U jednoj dimenziji, postoji N ćelija; u dve dimenzije N^2 ćelija (ako pretpostavimo da je svaka promenljiva particionisana u N ćelija). Za uzorke podataka $x \in \mathrm{R}^d$ (d-dimenzioni vektor x) postoji N^d ćelija. Ovaj eksponencijalni rast broja ćelija znači da kod većeg broja dimenzija veoma velika količina podataka je potrebna za procenu gustine. Na primer, gde su podaci šesto-dimenzionalni, deljenje skale promenljivih na 10 ćelija (razumno mali broj) daje milion ćelija. Kako bi se sprečilo da procena bude nula na velikom broju regiona, potreban je veliki broj opservacija. Ovo je poznato kao „prokletstvo dimenzionalnosti“. Drugi problem sa pristupom histograma je da su procene gustina diskontinualne i iznenada padaju na nulu na ivicama regiona (Webb & Copsey, 2011). Pristup histograma se može generalizovati da uključi Bajesove mreže (eng. *Bayesan networks*), takođe poznate i kao Kondiciono-probabilističke mreže (eng. Conditional Probabilistic Networks, CPN) (Webb & Copsey, 2011; Evans, 1998).

4.2.2.3 Metode jezgra

Kao što je već rečeno, jedan od problema sa pristupom histograma je da, za fiksne dimenzije ćelija, broj ćelija eksponencijalno raste sa dimenzijom vektora podataka. Ovaj problem se može prevazići donekle pomoću promenljive veličine ćelija. Metod *k*-

najbližih suseda (u svom najprostijem obliku) prevazilazi problem procenjujući gustinu pomoću ćelije u kojoj je broj trening uzoraka fiksan i nalazi veličinu ćelije koja sadrži najbližih *k*. *Metoda jezgra* (eng. *Kernel method*), takođe poznat i kao *Parzen metod* procene gustine) fiksira veličinu ćelije i nalazi broj uzoraka unutar ćelije i koristi to za procenu gustine, pri čemu rešava problem naglih „šiljaka“ (Webb & Copsey, 2011).

Posmatrajmo jednodimenzionalni primer i neka je $\{x_1, ..., x_n\}$ skup opservacija ili uzoraka podataka koje ćemo koristiti za procenu gustine. Možemo jednostavno zapisati procenu kumulativne funkcije raspodele kao

$$\hat{P}(x) = \frac{broj\ opservacija\ \leq x}{n}$$

Funkcija gustine, $p(x)$, je izvodena iz raspodele, ali raspodela je diskontinualna (u vrednostima opservacija; Slika 4.2.5) i njeni izvodi resultiraju skupom „šiljaka“ u tačkama uzoraka, x_i, i nulama na drugim mestima. Međutim, možemo definisati procenu gustine kao

$$\hat{p}(x) = \frac{\hat{P}(x+h) - \hat{P}(x-h)}{2h}$$

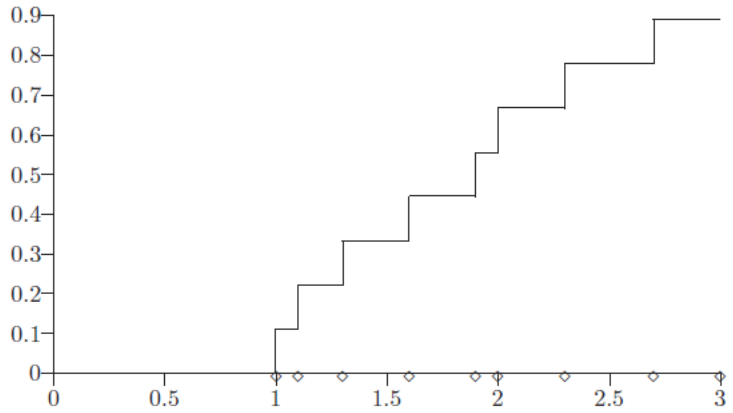
gde je h pozitivan broj. Ovo je odnos opservacija koje upadaju u interval $(x-h, x+h)$ podeljen sa $2h$. To se može zapisati kao

$$\hat{p}(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

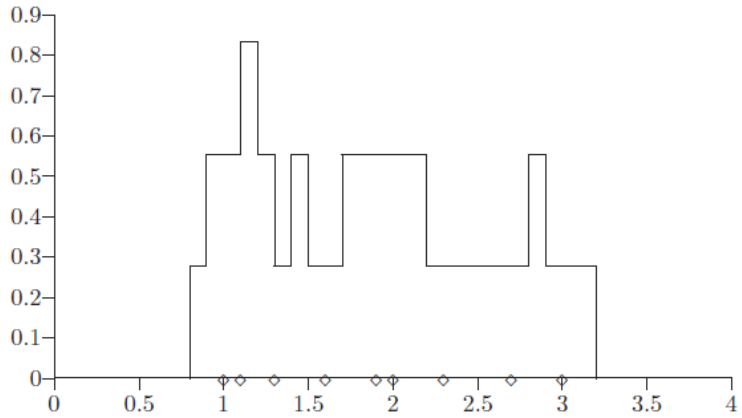
gde je $K(z)$ *pravougaono jezgro* (takođe poznato kao *top hat*)

$$K(z) = \begin{cases} 0 & |z| > 1 \\ \frac{1}{2} & |z| \leq 1 \end{cases}$$

Iz gornje dve jednačine proizilazi prethodna, jer za tačke uzorka x_i, unutar h od x, funkcija jezgra uzima vrednost ½, i time sumiranje daje vrednost ½ od broja opservacija unutar intervala. Slika 4.2.6 prikazuje procenu gustine za podatke koji su dali kumulativnu distribuciju na Slici 4.2.5 (Webb & Copsey, 2011).



Slika 4.2.5: Kumulativna raspodela

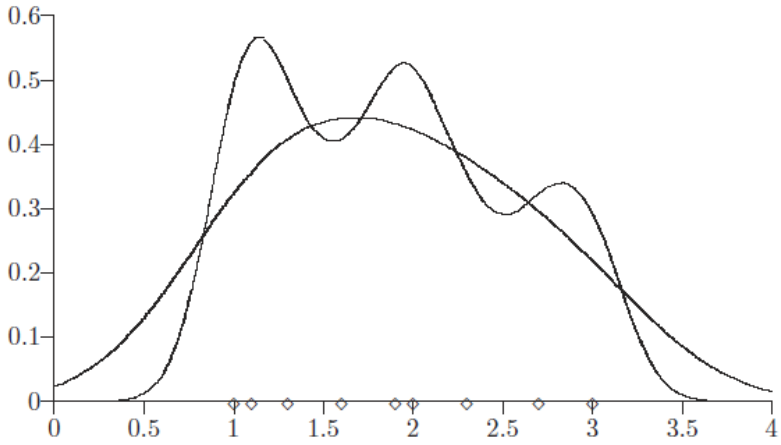


Slika 4.2.6: Procena gustine verovatnoće sa *top hat* jezgrom, h=0.2.

Slika 4.2.6 pokazuje nam da je procena gustine sama po sebi diskontinualna. Ovo dolazi iz činjenice da tačke unutar razlike h od x doprinose vrednosti $\frac{1}{2hn}$ gustini a udaljenije tačke vrednošću nula. Taj skok od $\frac{1}{2hn}$ do nula stvara te diskontinualnosti. Možemo ovo ukloniti i generalizovati estimator, korišćenem težinske funkcije koja je više glatka od prethodno date. Na primer, možemo imati težinsku funkciju $K_1(z)$ koja se smanjuje kako $|z|$ raste. Slika 4.2.7 prikazuje procenu gustine dobijenu sa normalnim jezgrom

$$K_1(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

i vrednostima h=0.2 i h=0.5. Ovo daje više glatku procenu gustine. Naravno, to ne znači da je ova procena obavezno tačnija nego ona sa Slike 4.2.6, ali možemo prestpostaviti da je osnovna gustina glatka funkcija i da traži glatku procenu (Webb & Copsey, 2011).



Slika 4.2.7: Gustina verovatnoće za dati prethodni primer, Gausovim jezgrom i različitim nivoima izglađivanja: h=0.2 i h=0.5.

Dakle, metoda jezgra za procenu gustine se može definisati tako da sa datim skupom opservacija $\{x_1, ..., x_n\}$, procena funkcije gustine, u jednoj dimenziji, je data kao

$$\hat{p}(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Gde je $K(z)$ nazvan funkcija jezgra a h je *paramtar širenja* ili *izglađivanja* (nekad nazivan i *protok*) (Webb & Copsey, 2011). Primeri popularnih univarijantnih funkcija jezra su dati u Tabeli 4.2.1.

Tabela 4.2.1: Popularno korišćene funkcije jezgra za univarijantne podatke.

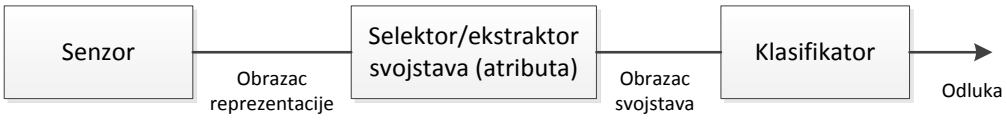
Funkcija jezgra	Analitička forma, $K(x)$
Pravougaona (*top hat*)	½ za $\|x\|<1$, 0 u suprotnom
Trouglasta	$1 - \|x\|$ za $\|x\|<1$, 0 u suprotnom
Dvotežinska (*Quartic*)	$\frac{15}{16}(1-x^2)^2$ za $\|x\|<1$, 0 u suprotnom
Normlna (Gausova)	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$
Bartlett–Epanechnikov	$\frac{3}{4}(1-\frac{x^2}{5})/\sqrt{5}$ za $\|x\| < \sqrt{5}$, 0 u suprotnom

4.2.3 Statističko prepoznavanje obrazaca i klasifikacija

Prepoznavanje obrazaca (eng. *Pattern Recognition*) se kao oblast izučavanja značajno razvilo tokom 1960-ih. Ima veliki broj primena, od klasičnih kao što je automatsko prepoznavanje znakova i medicinske dijagnostike do novijih u data mining-u (npr. bodovanje kredita, analiza prodaje kupcima, itd.) (Webb & Copsey, 2011). Ukratko, prepoznavanje obrazaca je proces uzimanja sirovih podataka i preduzimanje akcija na osnovu kategorije obrasca (Duda, Hart, & Stork, 2000). Poslednjih godina postoje brojni proboji u razvoju metodologije i samoj primeni, koji uključuju, na primer, Bajesove računske metode i *kernel-based* metode (uključujući *metode podržavajućih vektora*) (Webb & Copsey, 2011). Postoji veliko preklapanje oblasti izučavanja statističkog prepoznavanja obrazaca i mašinskog učenja, iako je naglasak mašinskog učenja možda više na strani računarski intenzivnih metoda a manje na statističkom pristupu.

Generalno govoreći, cilj oblasti prepoznavanja obrazaca jeste da automatizuje procese koje obavljaju ljudi. Na primer, automatska analiza i prepoznavanje fotomikrografije ćelija tkiva se može koristiti u testovima krvi, testovima na rak i analizi moždanog tkiva. Drugi primer, koji je od nas od najvećeg značaja, je se odnosi na automatsko prepoznavanje slika daljinskog uzorkovanja, bilo iz satelita ili sa drugih platformi (McLachlan, 2004). Postoji više razloga za razvoj atumatizovanog procesa klasifikacije – umesto da se na klasifikaciju svih obrazaca primenjuje isti postupak kao i kod klasifikacije trening seta, kao što su: da bi se uklonio ljudski rad u procesu prepoznavanja, kako bi proces postao pouzdaniji, da se smanje troškovi i ubrza proces, da se omoguću operativnost u nepristupačnim ili opasnim okruženjima, da se omogući operativnost na daljinu (npr. klasifikacija kod daljinskog uzorkovanja), da se omogući dijagnostika takva da se ne ošteti predmet dijagnoze (npr. u medicini, na živim subjektima, ili kod kompleksih mašina), itd. (Webb & Copsey, 2011).

Projektovanje klasifikatora obrazaca ili pravila za diskriminaciju ili alokaciju zahteva specifikaciju parametara klasifikatora obrazaca (šematski prikazan na slici 4.2.8) tako da daje optimalne (u nekom smislu) odgovore za dati ulazni obrazac (entitet). Taj odgovor je obično procena klase kojoj obrazac pripada. Pretpostavljamo da imamo skup obrazaca poznatih klasa $\{(x_i, z_i), i = 1, ..., n\}$ (trening set, odnosno skup podataka za učenje, trening ili projektovanje) koji smo koristili za projektovanje klasifikatora (za određivanje njegovih unutrašnjih parametara). Kad je ovo urađeno, možemo proceniti pripadnost klasi za obrazac x za koji je klasa nepoznata. Učenje modela na osnovu trening seta je proces *indukcije*, dok je primena naučenog modela na obrasce nepoznatih klasa proces *dedukcije* (Webb & Copsey, 2011).



Slika 4.2.8: Klasifikator obrazaca

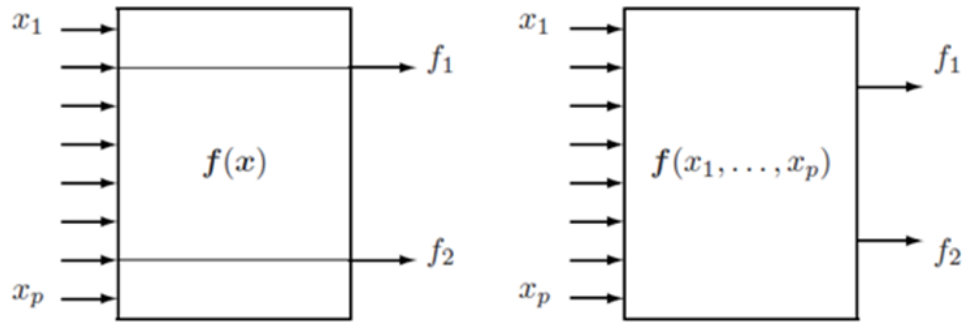
Prema tome, svrha klasifikatora obrazaca je da obezbedi: (a) *deskriptivni model* koji objašnjava razlike između obrazaca različitih klasa u smislu svojstava i njihovih mera, i (b) *prediktivni model* koji predviđa klase neoznačenih obrazaca. Prema (Duda, Hart, & Stork, 2000) pod-problemi klasifikacije obrazaca, koji se moraju rešavati, su: ekstrakcija svojstava, šum u podacima, *overfitting*, odabir modela, priorno znanje, nedostajuća svojstva, segmentacija, kontekst, invarijantnost, troškovi i rizici, kompleksnost računanja, itd. Istraživanje kod prepoznavanje obrazaca se može sastojati od više faza, od kojih ne moraju sve biti prisutne ili neke mogu biti spojene zajedno, a tipično to su (Webb & Copsey, 2011):

1. *Formulisanje problema*: razumevanje ciljeva istraživanja i planiranje;
2. *Prikupljanje podataka*: merenja varijabli i utvrđivanje detalja procedure prikupljanja podataka (*ground truth*);
3. *Inicijalno ispitivanje podataka*: provera podataka i razumevanje strukture;
4. *Selekcija i ekstrakcija svojstava* (atributa);
5. *Nenadgledana klasifikacija* ili *klasterovanje*: eksploratorna analiza podataka kako bi se došlo do zaključaka u analizi. Takođe to može biti predprocesiranje podataka za nagledane procedure klasifikacije;
6. *Primena diksriminacionih (ili regresionih) procedura*: klasifikator se projektuje pomoću trening seta i primera;
7. *Procena rezultata*: primena istreniranog klasifikatora na nezavisne test podatke obrazaca imenovanih klasa. Performanse klasifikacije se često sumiraju u vidu *matrice konfuzije*;
8. *Interpretacija*.

Selekcija svojstava (eng. *feature selection* ili još *feature selection in the measurement space*) je proces odabira onih promenjivih iz izmerenog skupa podataka koji su najprikladniji za zadatak klasifikacije. Kod nekih problema nema automatske selekcije svojstava, već te zadatke vrši „istraživač“ koji poznaje domen problema. Metode selekcije svojstava koje su blisko vezane za sam tip klasifikatora u glavnom daju bolje rezultate u odnosu na metode filtriranja (koje su nezavisne od klasifikatora) (Webb & Copsey, 2011).

Ekstrakcija svojstava (eng. *feature extraction*, ili još *feature selection in the transformed space*) je uzimanje linearnih ili nelinearnih transformacija originalnih promenljivih kako bi se dobile nove promenljive. Jedna od najšire korišćenih tehnika ekstrakcije svojstava je *analiza glavnih komponenti* (eng. *Principal Component Analysis*, PCA), koja sprovodi kako bi se pronašao smanjeni skup u odnosu na osnovne varijable ili komponente koje najviše doprinose varijacijama u osnovnim podacima (Webb & Copsey, 2011). Konceptualne granice između ekstrakcije svojstava i same klasifikacije su donekle proizvoljne: idealan ekstraktor svojstava bi činio posao klasifikatora trivijalnim, dok svemogući klasifikator ne bi trebao pomoć sofisticiranog ekstraktora. Razlika je tu više zbog praktičnih nego teoretskih razloga. Genralno, zadatak ekstraktora svojstava je značajno više zavisan od problema odnosno domena, nego što je sam klasifikator (Duda, Hart, & Stork, 2000).

U oba slučaja, i kod selekcije i kod ekstrakcije, se radi o redukovanju broja korisnih varijabli u odnosu na originalni broj, odnosno do redukcije dimenzionalnosti podataka, što može dovesti boljih performansi klasifikatora i boljeg razumevanja podataka. U prvom slučaju se identifikuju one promenljive koje ne doprinose zadatku klasifikacije, i kod rešavanja problema diskriminacije možemo zanemariti promenljive koje ne doprinose razdvajanju klasa. Drugi pristup je nalaženje transformacije u prostor svojstava niže dimenzije. Transformacija može biti urađena pre projektovanja klasifikatora (pre-filtracija podataka) ili integrisana sa klasifikatorom. Oba slučaja su šematski prikazan na Slici 4.2.9 (Webb & Copsey, 2011).



Slika 4.2.9: Redukcija dimenzionalnosti pomoću (a) selekcije svojstava i (b) ekstrakcije svojstava.

Oblast prepoznavanja obrazaca poznata kao statističko prepoznavanje obrazaca ima čvrste veze sa statističkom teorijom odlučivanja i oblastima multivarijacione analize – naročito diskriminacionom analizom (McLachlan, 2004). Pošto je klasifikacija, u osnovi, zadatak utvrđivanja modela koji je generisao obrasce (entitete), različite tehnike klasifikacije mogu biti korisne zavisno od vrste samih modela koji su kandidati. Kod statističkog prepoznavanja obrazaca se fokusiramo na statističke osobine

obrazaca (generalno izraženih u gustinama verovatnoća) (Duda, Hart, & Stork, 2000). Kako smo već rekli, obrascom se smatra jedan entitet koji je predstavljen vektorom svojstava (atributa) konačne dimenzije. Prema tome, prepoznavanje obrazaca u odnosu na konačni broj predefinisanih grupa obrazaca može biti formulisano u okvirima diskriminacione analize (McLachlan, 2004).

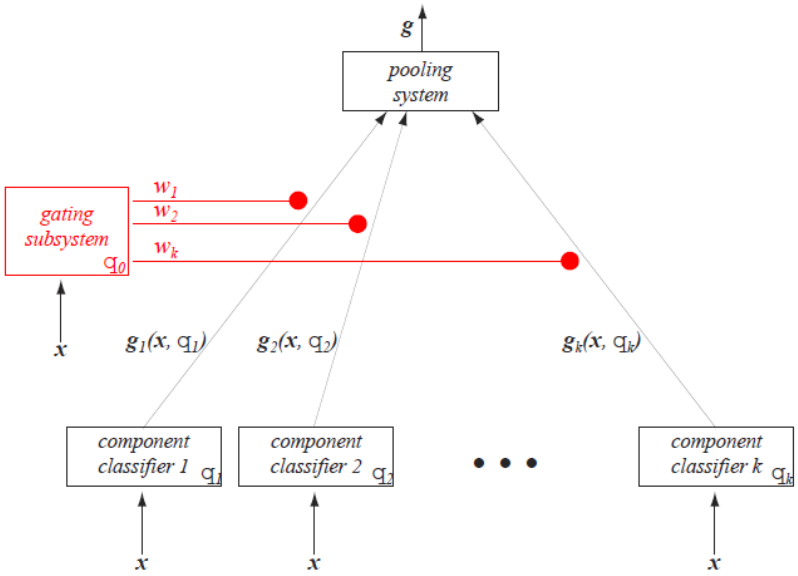
Diskriminaciona analiza ili (statistička) diskriminacija je ovde koristi za probleme povezane za statističku separaciju između različitih klasa ili grupa i za alokaciju entiteta u grupe (ograničenog broja), gde je postojanje grupa poznato a priori i gde, tipično, postoje dostupni *podaci o svojstvima* (eng. *feature data*) entiteta iz različitih grupa koje su u osnovi (McLachlan, 2004). Drugim rečima, podrazumeva veliki broj problema statističkog prepoznavanja obrazaca, gde se obrascom smatra jedan entitet i predstavljen je vektorom svojstava (atributa) obrasca konačne dimenzije (McLachlan, 2004). Ono što je od posebnog značaja za nas u ovom istraživanju je to da je diskriminaciona analiza takođe široko u upotrebi u oblasti prepoznavanja obrazaca koja se primarno bavi slikama (McLachlan, 2004). Međutim, važno je napomenuti da takva klasifikacija obrazaca nije isto što i obrada slika. Kod obrade slika, i ulaz i izlaz iz obrade je slika a koraci obrade slike često uključuju obradu rotiranja, povećanja kontrasta, i druge transformacije koje zadržavaju polazne informacije, dok ekstrakcija svojstava, kao što je nalaženje najviših ili najnižih vrednosti intenziteta, gube informacije (ali u nadi da očuvaju sve što je od značaja za željeni zadatak) (Duda, Hart, & Stork, 2000).

Radi kompletnosti pregleda, spomenućemo još dve metode klasifikacije obrazaca: *stabla odlučivanja* i *zbirne* (*ansambl*) *metode*.

Stabla odlučivanja (eng. *Decision Trees*), još i *stabla klasifikacije*, leže u preseku oblasti statističkog prepoznavanja obrazaca i mašinskog učenja. S jedne strane ona su primer neparametarskog pristupa koji modeluje funkciju klasifikacije/regresije kao težinsku sumu osnovnih funkcija. S druge strane, mogu se koristiti da generišu pravila interpretacije, koja mogu biti korisna u mnogim primenama. Stabla odlučivanja su suštinski veoma jednostavni modeli koji se razlikuju u jednom veoma važnom aspektu od ostalih pristupa klasifikaciji: interpretacija klasifikatora odnsono reprezentacija pravila klasifikacije je data kao stablo, sa čvorovima označenim kao svojstvima, ivicama kao vrednostima (ili skupom vrednosti) tih svojstava, i listovima kao klasama. Stablo odlučivanja je primer višefaznog procesa odlučivanja: umesto da koristi kompletan skup svojstava odjednom za donošenje odluke, zarličiti podskupovi svojstava se koriste na različitim nivoima stabla (Webb & Copsey, 2011).

Zbirne (*ansambl*) *metode* (eng. *Ensemble Methods*, još *mixture of expert models*, *ensemble classifiers*, *modular classifiers* ili ponekad *pooled classifiers*) polaze od pretpostavke da možemo povećati performanse klasifikatora kombinujući izlaze nekoliko (komponentnih) klasifikatora (Webb & Copsey, 2011; Duda, Hart, & Stork, 2000). Takvi klasifikatori su naročito korisni ako je svaki od njegovih komponentnih klasifikatora veoma „istreniran“ (je „ekspert“) u različitoj oblasti prostora svojstava

(Duda, Hart, & Stork, 2000). Njihova arhitektura se sastoji od k komponentnih klasifikatora, od kojih svaki ima parametre za učenje θ_i, i=1,...,k. Za svaki ulazni obrazac x, svaki komponentni klasifikator i daje procenu pripadnosti kategoriji $g_{ir}=P(\omega_r|x,\theta_i)$. Izlazi su ponderisani pomoću *gating* podsistema kojim upravlja parametarski vektor θ_0, i objedinjeni za konačnu klasifikaciju (Slika 4.2.10) (Duda, Hart, & Stork, 2000). Zbirme metode (ili *tehnike kombinovanja klasifikatora*) su oblast od velikog izučavanja poslednjih godina i vezane su za napredak u literaturi fuzije podataka, gde je naročito problem fuzije odluka intenzivno adresiran (Webb & Copsey, 2011).



Slika 4.2.10: Šematski prikaz zbirnog (modularnog) klasifikatora

Na kraju, postavlja se pitanje: ako nas zanimaju generalizovane performanse (ako nema priornih pretpostavki o prirodi zadatka klasifikacije), da li možemo preferirati jedan klasifikator ili algoritam učenja u odnosu na drugi (da li ukupno jedan metod klasifikacije može biti superiorniji ili inferiorniji)? Kako nalaže *No Free Lunch teorema*, odgovor je – ne: po pitanju generalnih performansi, nema kontekstno ili problemski nezavisnih razloga da se favorizuje jedan metod učenja ili klasifikacije u odnosu na drugi (Duda, Hart, & Stork, 2000). Superiornost jednog algoritma koja naizgled može postojati je usled prirode problema koji se istražuje i raspodele podataka. Kada se suočavamo sa praktičnim problemima prepoznavanja obrazaca, potrebno je fokusirati se na najvažnije aspekte – priorne informacije, raspodelu podataka, količina trening podataka i funkciju troška (ili nagrade) (Duda, Hart, & Stork, 2000).

4.3 Diskriminaciona analiza

Diskriminaciona analiza je multivarijaciona metoda koja se bavi razdvajanjem različitih grupa i alokacijom entiteta u predefinisane grupe. Njena dva su osnovna cilja su: (1) *diskriminacija* ili *razdvajanje* između grupa – utvrđivanje da li postoji statistički značajna razlika srednjih vrednosti dve ili više grupa i određuje koja od više posmatranih promenljivih daje najveći doprinos utvrđenoj razlici, i (2) *klasifikacija* ili *alokacija* entiteta – utvrđivanje postupka za klasifikaciju entiteta na osnovu vrednosti nekoliko promenljivih u dve ili više razdvojenih, predefinisanih grupa (Kovačić, 1994). Čest je slučaj preklapanja ovih ciljeva, pa tehnike analize za razdvajanje između grupa istovremeno mogu služiti i za klasifikaciju opservacija u takve predefinisane grupe. Prvi cilj analize – diskriminacija ili razdvajanje između grupa obuhvata metode *deskriptivne diskriminacione analize*, dok drugi cilj – klasifikacija ili alokacija opservacija obuhvata metode *klasifikacije*. Diskriminaciona analiza, sa tehničke strane, formira linearne kombinacije nezavisnih promenljivih kojima će se diskriminacija između predefinisanih grupa izvršiti na taj način da greška pogrešne klasifikacije entiteta bude minimizirana, odnosno da se maksimizira relativan odnos varijansi između i unutar grupa (Kovačić, 1994; Radojičić, 2001; Radojičić, 2007).

Slično kao i kod regresione analize, kod diskriminacione analize na osnovu skupa nezavisnih promenljivih se predviđa ili opisuje ponašanje zavisne promenljive. Skup nezavisnih promenljivih je skup kontinualnih (neprekidnih) promenljivih, međutim za razliku od regresione analize, gde zavisna promenljiva takođe predstavlja kontinualnu promenljivu, u diskriminacionoj analizi je zavisna promenljiva kvalitativna. Ovde će zavisna promenljiva uzeti vrednost predefinisanih kategorija odnosno grupa (na primer, uzeće vrednosti 0 i 1 ako razmatramo problem diskriminacije dve grupe). Za razliku od metode analize varijanse, gde je zavisna promenljiva kvantitativna, a nezavisne su kvalitativne, kod diskriminacione analize zavisna promenljiva je kvalitativna, a nezavisne promenljive su kvantitativne (Kovačić, 1994).

U ovoj disertaciji fokusiraćemo se na tehnike diskriminacione analize koje se koriste da bi se određeni entitet klasifikovao u jednu od dve ili više alternativnih grupa (ili populacija), a u odnosu na niz izvršenih merenja daljinskim uzorkovanjem. Ove tehnike se takođe mogu koristiti za određivanje onih promenljivih koje najviše doprinose klasifikaciji. Prema tome, kao i kod regresione analize, postoje dve uloge diskriminacione analize: *predikcija* (predviđanje) i *deskripcija* (opisivanje).

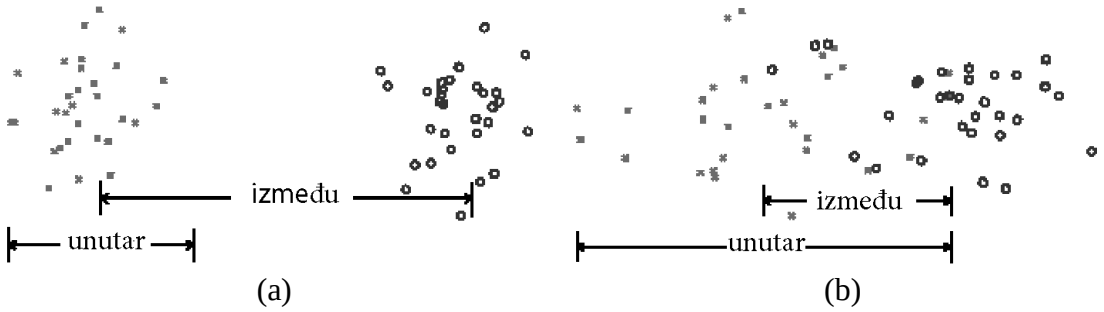
4.3.1 Klasifikacija entiteta metodama diskriminacione analize

Prvi korak u analizi podataka je određivanje deskriptivnih mera za svaku od predefinisanih grupa (Radojičić, 2001). Pretpostavimo da entitete treba klasifikovati u dve grupe na osnovi jedne promenljive X. Da bismo odredili u koju grupu da klasifikujemo svaki od entiteta, moramo odrediti *graničnu tačku klasifikacije* (eng.

Cutting Point). Ako obeležimo ovu tačku sa C, tada će entitet e_i biti klasifikovan u grupu (populaciju) 1, ako važi $X_i \geq C$, gde je X_i vrednost promenljive X, za *i*-ti entitet (Radojičić, 2001). Ukoliko entitet dolazi iz grupe 1 ali je izmereno X manje od C, tada bi se izvršila pogrešna klasifikacija entiteta u grupu (populaciju) 2 i obrnuto. Ukoliko možemo pretpostaviti da dve populacije imaju jednake varijanse, tada je uobičajena vrednost C:

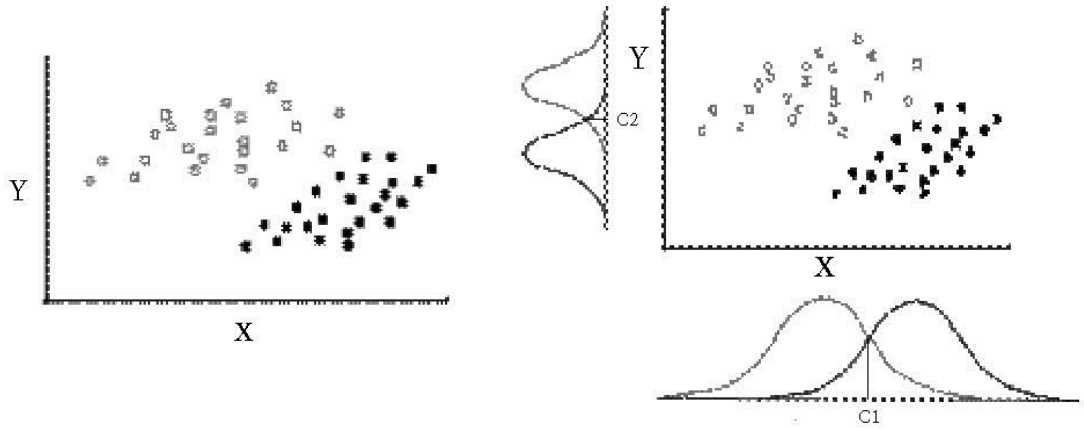
$$C = \frac{\bar{X}_1 + \bar{X}_2}{2}$$

gde su $\bar{X}_1$ i $\bar{X}_2$ prosečne vrednosti za svaku od predefinisanih grupa. Ova vrednost obezbeđuje da verovatnoće obeju grešaka budu jednake. Na Slici 4.3.1 prikazana je separacija entiteta na osnovu jedne promenljive. Idealizovana situacija prikazana na Slici 4.3.1.(a) se retko sreće u praksi. U situacijama u realnom životu nivo preklapanja dve distribucije je vrlo čest, dok su varijanse retko jednake (Slika 4.3.1.(b)).



Slika 4.3.1: (a) „Dobra“ i (b) „loša“ separacija entiteta između dve grupe, kada se posmatra jedna promenljiva

Kombinacija dve ili više promenljivih može doprineti boljoj klasifikaciji entiteta u grupe. Potrebno je napomenuti da broj promenljivih koje se koriste za klasifikaciju mora biti manji od $N_1 + N_2 - 1$, gde je N_1 broj entiteta u prvoj a N_2 broj entiteta u drugoj grupi, da se ne bi ušlo u oblast sa negativnim brojem stepeni slobode (Radojičić, 2001). Posmatrajmo dve promenljive X i Y, i skup entiteta, kao što je dato na Slici 4.3.2.(a). Ukoliko bismo posmatrali separaciju entiteta u grupe na osnovu primenljivih X i Y zasebno, tada bi takva separacija grafički mogla biti prikazana kao što je na Slici 4.3.2.(b).

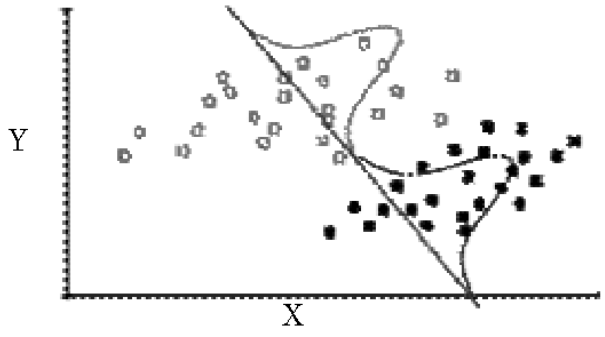


(a) (b)

Slika 4.3.2: (a) Distribucija i (b) separacija entiteta između dve grupe, kada se posmatraju dve promenljive

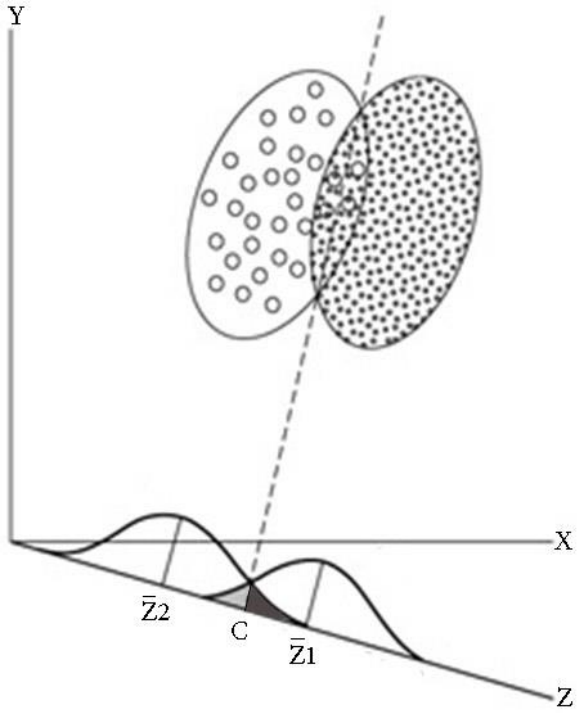
Slika 4.3.2 prikazuje univarijantnu distribuciju X i Y odvojeno. Univarijantna distribucija X je ono što se dobije ako se vrednosti Y zanemare. Baziravši se samo na X i njegovoj odgovarajućoj graničnoj tački C_1, došlo bi do pojave relativno velike količine grešaka za određene entitete (tj. pogrešne klasifikacije). Iako u nešto manjoj količini, slični rezultati se javljaju i za Y i njegovu odgovarajuću graničnu tačku C_2 (Radojičić, 2001).

Da bismo koristili obe varijable simultano, neophodno je da se na neki način podeli ravan X0Y na dva dela, od kojih svaki odgovara jednoj populaciji odnosno grupi, i da se entiteti odgovarajuće klasifikuju u predefinisane grupe. Najbolja moguća separacija za gornji primer prikazana je na Slici 4.3.3.



Slika 4.3.3: Simultana separacija entiteta između dve grupe

Jednostavan način za određivanje ova dva regiona je da se povuče prava linija kroz tačke preseka dve koncetracione elipse, kako je prikazano na Slici 4.3.4.



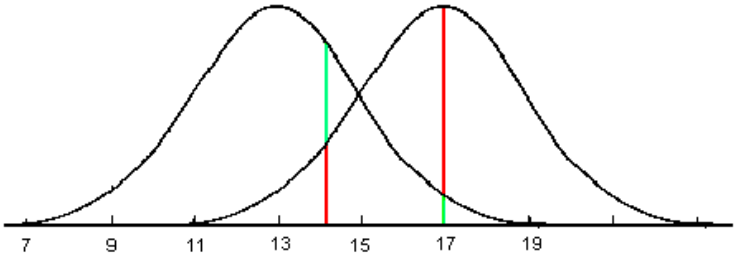
Slika 4.3.4: Simultana separacija entiteta između dve grupe

Procenat entiteta iz druge grupe (populacije) koji su pogrešno klasifikovani u prvu grupu je prikazan u tamno sivoj šrafiranoj oblasti desno od tačke C. Osenčena površina svetlije sivom bojom prikazuju procenat entiteta iz prve grupe (populacije) koje su pogrešno klasifikovane u drugu. Greške koje nastaju pri korišćenju dve promenljive simultano (Slike 4.3.3 i 4.3.4) su znatno manje nego one koje nastaju korišćenjem bilo koje od dve promenljive samostalno (Slika 4.3.2.(b)).

4.3.1.1 Klasifikacija i granična tačka u modelu sa normlanom raspodelom

Promenljive korišćene pri klasifikaciji su X_1, X_2,..., X_p. Standardni model pretpostavlja da svaka promenljiva ima normalnu raspodelu obe grupe u koje se klasifikuju entiteti (Radojičić, 2001). Dalje se pretpostavlja da je matrica kovarijanse jednaka u obe grupe (populacije). Međutim, srednje vrednosti za date promenljive mogu biti različite u dve populacije. Dalje pretpostavke su da imamo slučajan uzorak za svaku od populacija. Veličine uzoraka su predstavljene kao N_1 i N_2. Alternativno, problematiku možemo predstaviti kao dve subpopulacije jedne iste populacije, odnosno, sakupljen je jedan uzorak, a kasnije je ustanovljeno da se sastoji od dve subpopulacije (Radojičić, 2001).

Ranije spomenuti postupak klasifikacije je pridruživao entitete bilo grupi 1 ili grupi 2. Međutim, obzirom da uvek postoji mogućnost pogrešne klasifikacije entiteta, moguće je izračunati verovatnoću da entitet pripada jednoj ili drugoj grupi (Slika 4.3.5).



Slika 4.3.5: Određivanje verovatnoće pripadnosti entiteta

To je moguće izračunati pomoću formule verovatnoće pripadanja populaciji koja, kako su definisali (Truett, Cornfield, & Kannel, 1967), glasi:

$$P_i = \frac{1}{1+e^{(-Z_i+C)}}$$

gde je P_i verovatnoća pripadnosti i-tog entiteta grupi 1, Z_i je vrednost diskriminacione funkcije za *i*-ti entitet, a C je granična tačka. Verovatnoća pripadnosti populaciji 2 je jedan minus verovatnoća pripadnosti populaciji 1.

Ranije je granična tačka C bila korišćena kao tačka koja je uticala na podjednak procenat grešaka oba tipa, tj. verovatnoće pogrešne klasifikacije entiteta iz grupe 1 u grupu 2 i obrnuto. Međutim, izbor vrednosti tačke C može biti takav da rezultat bude bilo koji odnos navedenih verovatnoća grešaka (Radojičić, 2001). Da bi se objasnio način na koji se takav izbor vrši, neophodno je uvesti koncept *verovatnoće apriori*. Obzirom da dve populacije (grupe) zajedno čine ukupnu populaciju, neophodno je ispitati njihovu relativnu veličinu. Verovatnoća apriori prve populacije je verovatnoća da entitet, koji je slučajno izabran, zaista dolazi iz te populacije. Drugim rečima, to je onaj udeo entiteta u ukupnoj populaciji, koji pripada prvoj populaciji (grupi). Ovaj udeo se obeležava sa q_1, a kao što je ranije naglašeno q_1=1-q_1.

Teorija koja stoji iza izbora tačke C je tako postavljena da ukupna verovatnoća pogrešnih klasifikacija bude minimalna (Radojičić, 2001). Ova ukupna verovatnoća je definisana kao q_1 pomnoženo sa verovatnoćom pogrešne klasifikacije entiteta iz populacije 1 u populaciju 2 $P\{2 \text{ dato } 1\}$, plus q_2 pomnoženo sa verovatnoćom pogrešne klasifikacije entiteta iz populacije 2 u populaciju 1 $P\{1 \text{ dato } 2\}$, odnosno:

$$q_1 \cdot P\{2 \text{ dato } 1\} + q_2 \cdot P\{1 \text{ dato } 2\}$$

U slučaju normalnog modela sa više promenljivih, optimalni izbor granične tačke C je:

$$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln \frac{q_2}{q_1}$$

Za podjednake grupe (populacije), odnosno ukoliko je $q_1 = q_2 = ½$, tada je $q_2/q_1 = 1$ i $\ln(q_2/q_1) = 0$. U tom slučaju C je isto kao što je definisano na početku poglavlja:

$$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2}$$

Zbog toga je ranije implicitno podrazumevano da je $q_1 = q_2 = ½$. U praksi, biraju se različite vrednosti za C i za svaku od vrednosti određuju se dve verovatnoće pogrešne klasifikacije. Željena vrednost C će biti postignuta kada bi se postigne balans te dve verovatnoće. Jedna od metoda određivanja težine grešaka ide preko određivanja relativnih gubitaka dva tipa pogrešnih klasifikacija (Radojičić, 2001). Ovi gubici se mogu izraziti kao gubitak $g\{2 \text{ dato } 1\}$ i gubitak $g\{1 \text{ dato } 2\}$.

Granična tačka C tada može biti izabrana tako da minimizira ukupne gubitke pogrešne klasifikacije (Radojičić, 2001), tj.:

$$q_1 \cdot P\{2 \text{ dato } 1\} \cdot g\{2 \text{ dato } 1\} + q_2 \cdot P\{1 \text{ dato } 2\} \cdot g\{1 \text{ dato } 2\}$$

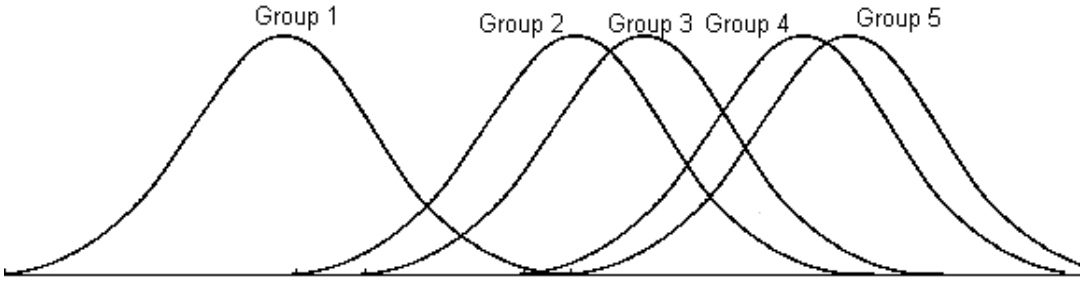
Izbor tačke C koja postiže datu minimizaciju je:

$$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2} + K$$

gde je:

$$K = \ln \frac{q_2 \cdot g\{1 \, dato \, 2\}}{q_1 \cdot g\{2 \, dato \, 1\}}$$

U dosadašnjem tekstu je radi lakšeg razumevanja, razmatran problem klasifikacije na dve grupe, ali analogijom se vrlo jednostavno može doći do rešenja za probleme koje sadrže klasifikaciju u više grupa (Slika 4.3.6).



Slika 4.3.6: Diskriminaciona analiza u više grupa

Dalje, istraživanja su pokazala da, i u slučajevima da promenljive X_1, X_2, ... , X_p nemaju normalnu raspodelu, njihovo korišćenje u linearnoj diskriminacionoj analizi može poboljšati klasifikaciju. Zbog toga se može reći da nije tako isključiv i strog uslov korišćenja promenljivih koje zadovoljavaju uslov normalnosti raspodele (Kovačić, 1994).

4.3.1.2 Veza multivarijacione linearne regresije i diskriminacione analize

Postoji korisna veza između multivarijacione linearne regresije i diskriminacione analize (Radojičić, 2001). Kada se vrši interpretacija regresije, klasifikacione promenljive X_1, X_2, ... , X_p su predstavljene kao nezavisne promenljive. Zavisne promenljive kod diskriminacione analize su proste promenljive koje daju neke indikacije o populaciji za koju je vršena opservacija. Tačnije,

$$Y = \frac{N_2}{N_1 + N_2}$$

ukoliko je vršena opservacija prve grupe (populacije) i

$$Y = -\frac{N_1}{N_1 + N_2}$$

ukoliko je vršena opservacija druge grupe. Kada se vrši uobičajena višestruka regresiona analiza, rezultujući regresioni koeficijenti su proporcionalni koeficijentima kod diskriminacione funkcije $a_1, a_2, ...a_p$ (Lachenbruch, 1975). Vrednost rezultujućeg višestrukog koeficijenta korelacije R je u vezi sa Mahalanobisovim odstojanjem D^2 preko sledeće formule:

$$D^2 = \frac{R^2}{1-R^2} \cdot \frac{[N_1 + N_2] \cdot [N_1 + N_2 - 2]}{N_1 \cdot N_2}$$

Dakle, moguće je iz višestruke regresione analize dobiti koeficijente diskriminacione funkcije $a_1, a_2, ...a_p$ i vrednost odstojanja D^2. $\bar{Z}_1$ i $\bar{Z}_2$ se mogu dobiti množenjem svakog koeficijenta sa srednjom vrednošću odgovarajuće promenljive uzorka $X_1, X_2, ..., X_p$, i dodavanjem rezultata. Granična tačka C se može izračunati kao $C = \left(\bar{Z}_1 + \bar{Z}_2\right)/2$ (Radojičić, 2001).

4.3.2 Diskriminacione funkcije i ravni odlučivanja

U ranijim poglavljima, klasifikacija se postizala primenom Bajesovog pravila odlučivanja. To zahteva poznavanje klasno-uslovnih funkcija gustine, $p(x|\omega_i)$ (kao što je normalna raspodela čiji su parametri procenjeni iz podataka), ili neparametarske metode procene gustine (poput procene gustine metodom jezgra). Ovde, ćemo umesto da pravimo pretpostavke o $p(x|\omega_i)$, praviti pretpostavke o formi *dikriminacionih funkcija* (eng. *discriminant functions*) (Webb & Copsey, 2011). U poglavlju koje govori o klasifikaciji entiteta pomoću diskriminacione analize već smo govorili o pristupu klasifikaciji s nadzorom sa linijom-diskriminacionom funkcijom. U ovom delu ćemo dati više detalja o pristupu korišćenjem *linearne diskriminacione funkcije*.

Dalje, iz ugla klasifikacije obrazaca pomoću klasifikacije s nadzorom, kada je dat skup merenja dobijen kroz opservacije, predstavljen kao vektor obrasca x, želimo da dodelimo obrazac u jedno od C mogućih klasa, ω_i, $i = 1, ..., C$. Pravilo odlučivanja particioniše prostor merenja u C oblasti, Ω_i, i = 1, . . . , C. Ako je vektor opservacije unutar Ω_i onda pretpostavljamo da pripada klasi ω_i (pri čemu svaka oblast klase Ω_i može biti višestruko povezana – može biti sačinjena od više razdvojenih oblasti). Granice između regiona Ω_i su *granice odlučivanja* ili *površine* (*ravni*) *odlučivanja* (Webb & Copsey, 2011). Pretpostavimo da imamo skup trening uzoraka $x_1, ..., x_n$, od kojih je svaki dodeljen jednoj od dve klase, ω_1 ili ω_2. Korišćenjem tog skupa uzoraka tražićemo težinski vektor w i prag w_0 takve da

$$w^T x + w_0 \begin{cases} > 0 \\ < 0 \end{cases} \Rightarrow x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}.$$

Površina odlučivanja (eng. *decision surface*) je hiperravan pretstavljena jednačinom

$$g(x) = w^T x + w_0 = 0$$

koja ima jediničnu normalu u pravcu w, i normalno rastojanje $|w_0|/|w|$ od koordinatnog početka. Udaljenost entiteta x od hiperravni odlučivanja je data sa $|r|$, gde je

$$r = \frac{g(x)}{|w|} = \frac{w^T x + w_0}{|w|}$$

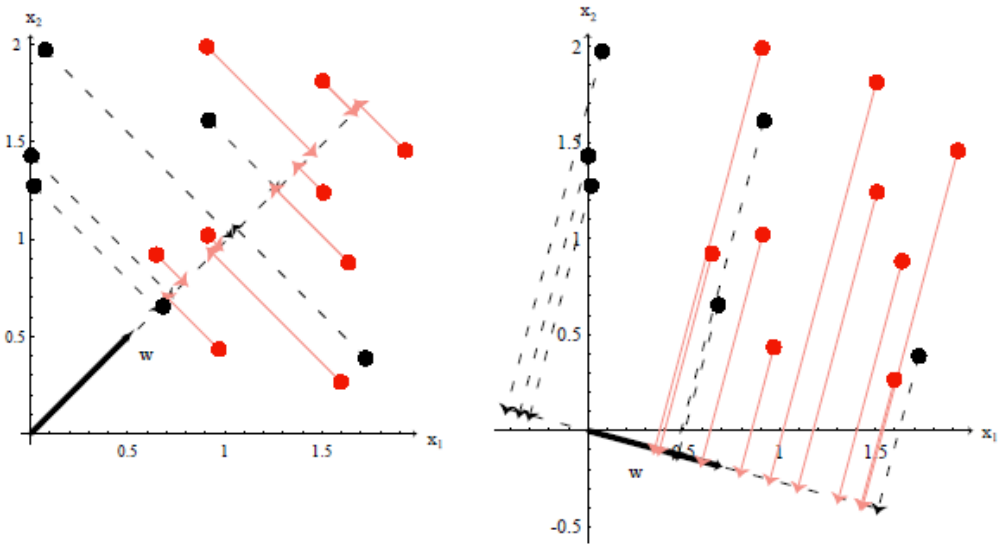
i gde znak od r ukazuje na kojoj strani hiperravni odlučivanja leži entitet, i prema tome kojoj klasi bi trebalo da bude de dodeljen (Webb & Copsey, 2011).

Nelinearna diskriminaciona analiza (eng. *Nonlinear discriminant analysis*) je razvijana primarno u literaturi koja se bavi *neuronskim mrežama* (eng. *neural networks*) i *mašinskim učenjem* (eng. *machine learning*), *mrežama funkcija sa radialnom osnovom* (eng. *radial basis function*, RBF, *network*), *mašine podržavajućih vektora* (eng. *support vector machine*, SVM) i *višeslojnim perceptronima* (eng. *multilayer perceptron*, MLP). Oni pretstavljaju fleksibilne modele za nelinearnu diskriminacionu analizu koja daje dobre performanse na širokom broju problema (Webb & Copsey, 2011).

4.3.2.1 Fišerova linearna diskriminaciona funkcija

Jedan od frekventnih problema koji se javlja u primeni statističkih tehnika za prepoznavanje obrazaca i klasifikaciju entiteta se naziva „prokletstvo dimenzionalnosti“. Procedure koje su analitički ili računarski smislene u niskodimenzionalnom prostoru mogu biti veoma nepraktične u prostoru sa 50 ili 100 dimenzija. Iz tog razloga su razvijene razne tehnike za redukovanje dimenzionalnosti prostora svojstava u nadi da će se doći do problema koji je lakše rešavati (Duda, Hart, & Stork, 2000).

Možemo redukovati dimenzionalnost sa *d* dimenzija na jednu dimenziju ukoliko jednostavno projektujemo *d*-dimenzionalne podatke na jednu liniju. Naravno, čak i ako uzorci formiraju dobro razdvojene, kompaktne klastere u *d*-prostoru, projekcija na proizvoljnu liniju će obično dovesti do mešanja uzoraka iz svih klasa, i tako loše rezultate klasifikacije. Međutim, pomerajući tu liniju možemo naći takvu orijentaciju za koju su projekcije uzoraka dobro razdvojene (Slika 4.3.7). Ovo je upravo i cilj klasične diskriminacione analize (Duda, Hart, & Stork, 2000).



Slika 4.3.7: Projekcija uzoraka na dve različite linije. Slika desno pokazuje veće razdvajanje između crvenih i crnih projektovanih tačaka.

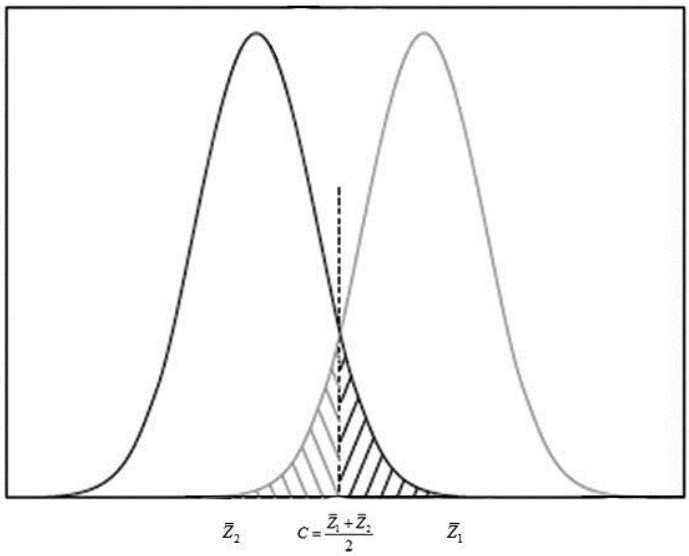
Liniju podele, koja u ovom slučaju služi za klasifikaciju entiteta, je uveo Fišer (Fisher, 1936) kao jednačinu Z=C, gde nova promenljiva Z predstavlja linearna kombinaciju X_1 i X_2, a C je konstanta definisana relacijom

$$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2}$$

gde su $\bar{Z}_1$ srednja vrednost Z u prvoj grupi (populaciji) i $\bar{Z}_2$ srednja vrednost Z u drugoj grupi (populaciji). Linearnu kombinaciju Z ćemo zvati *Fišerovom diskriminacionom funkcijom* (Radojičić, 2001), napisanom kao:

$$Z = a_1 X_1 + a_2 X_2$$

za slučaj dve promenljive. Formule za izračunavanje koeficijenata a_1 i a_2 mogu se naći u radovima poput (Fisher, 1936; Lachenbruch, 1975). Kada se raspodela funkcije Z prikaže na grafikonu, dobija se rezultat kao na Slici 4.3.8. U ovom slučaju, problem klasifikacije zasnovan na dve promenljive X_1 i X_2, redukovan je na situaciju u kojoj imamo samo jednu promenljivu Z (Radojičić, 2001).



Slika 4.3.8: Raspodela Fišerove diskriminacione funkcije za populacije 1 i 2

Kao što smo napomenuli, suština koncepta diskriminacione funkcije je da se takođe koristi u situacijama gde ima više promenljivih X_1, X_2,...,X_p (Radojičić, 2001). Prilikom razvoja diskriminacione linearne funkcije, Fišer (Fisher, 1936) nije morao da pravi nikakve pretpostavke vezane za raspodele promenljivih koje se koriste u procesu klasifikacije. On je postavio diskriminacionu funkciju kao:

$$Z = a_1 X_1 + a_2 X_2 + \mathrm{K} + a_p X_p$$

Interpretacija datih koeficijenata $a_1, a_2, ...a_p$ može pružiti više informacija koje će olakšati tumačenje rezultata. Fišerova diskriminaciona funkcija omogućava određivanje pravca i stepena doprinosa svake promenljive korišćene pri klasifikaciji. Prvo je neophodno ispitati znak svakog koeficijenta: ukoliko je pozitivan, entiteti sa većom vrednošću odgovarajuće promenljive imaju tendenciju pripadanja prvoj grupi (populaciji), i obrnuto. Da bi se kvantifikovala veličina distribucije, potrebno je standardizovati gore navedene koeficijente, radi lakšeg sagledavanja dobijenih rezultata. Kao i u slučaju regresione analize, vrednosti $a_1, a_2, ...a_p$ se ne mogu direktno porediti. Međutim, uticaj relativnog efekta svake promenljive na diskriminacionu funkciju može se dobiti iz *standardizovanih diskriminacionih koeficijenata*. Ove tehnike uključuju korišćenje ukupne (ili unutar-grupne) matrice kovarijansi (Radojičić, 2001).

Posmatraćemo dve srednje vrednosti za Z, $\bar{Z}_1$ i $\bar{Z}_2$. Takođe imamo ukupnu varijansu uzorka Z kao S_Z^2 (ova statistika je slična ukupnoj varijansi korišćenoj u *t* testu za dva nezavisna uzorka). Da bi se izmerilo koliko su „daleko“ dve grupe jedna od druge, u smislu vrednosti Z, računamo:

$$D^2 = \frac{\left(\bar{Z}_1 - \bar{Z}_2\right)^2}{S_Z^2}$$

Fišer je odredio koeficijente $a_1, a_2, ...a_p$ na taj način da D^2 uzme maksimalnu moguću vrednost (Radojičić, 2001). Vrednost D^2 može se interpretirati kao kvadrat rastojanja između srednjih vrednosti standardizovane vrednosti Z. Veća vrednost D^2 dovodi do zaključka da je lakše odlučiti se između dve grupe. Vrednost D^2 se naziva *Mahalanobisovo odstojanje* (eng. *Mahalanobis distance*). Navedeni koeficijenti i odstojanje su funkcije grupnih srednjih vrednosti i ukupne varijanse i kovarijanse promenljivih (Klecka, 1980).

Razmotrimo sada detaljnije osnove za primenu Fišerove diskriminacione funkcije. Uzmimo skup od *n d*-dimenzionalnih uzoraka $x_1,...,x_n$, n_1 u podskupu D_1 označenih kao ω_1 i n_2 u podskupu D_2 označenih sa ω_2. Ako formiramo linearnu kombinaciju komponenata x, dobićemo skalarni proizvod

$$y = \mathrm{w}^t \mathrm{x}$$

i odgovarajući skup od *n* uzoraka $y_1,...,y_n$ podeljenih u podskupove Y_1 i Y_2. Geometrijski, ako je $\|\mathrm{w}\| = 1$, svako y_i je projekcija odgovarajućeg x_i na liniju u pravcu w. Zapravo, intenzitet w i nije od stvarnog značaja, s obzirom da samo skalira *y*. Međutim, usmerenje w je važno (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011).

Sada je potrebno odrediti najbolje takvo usmerenje w, takvo za koje očekujemo da će obezbediti preciznu klasifikaciju. Mera separacije između projekcije tačaka je razlika između srednjih vrednosti uzoraka. Ako je m_i d-dimenziona srednja vrednost uzoraka data sa

$$m_i = \frac{1}{n_i} \sum_{\mathrm{x} \in D_i} \mathrm{x} \, ,$$

onda srednja vrednost uzorka za projekcije tačaka data kao

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in Y_i} \mathrm{w}^t \mathrm{x} = \mathrm{w}^t \mathrm{m}_i \, .$$

i samo je projekcija m_i.

Sledi da je udaljenost između projektovanih srednjih vrednosti

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathrm{w}^t(\mathrm{m}_1 - \mathrm{m}_2)| \, ,$$

i da možemo ovo rastojanje povećati skaliranjem w. Naravno, da bismo dobili dobro razdvajanje projekcije podataka želimo da razlika između srednjih vrednosti bude velika u odnosu na neku meru standardnih devijacija za svaku klasu. Umesto formiranja varijansi uzoraka, definisaćemo *rasipanje* za projekcije uzoraka označenih sa ω_i kao

$$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2 \, .$$

Tako, $(1/n)(\tilde{s}_1^2 + \tilde{s}_2^2)$ je procena varijanse objedinjenih podataka, a $\tilde{s}_1^2 + \tilde{s}_2^2$ se naziva ukupna *rasipanje (raštrkanost) unutar-klasa* za projekcije uzoraka. *Fišerov linearni diskriminant* koristi takvu linearnu funkciju $\mathrm{w}^t\mathrm{x}$ za koju kriterijumska funkcija

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

maksimalna (i nezavisna od $\|\mathrm{w}\|$). Dok w koje maksimizira $J(\cdot)$ vodi do najbolje separacije između dva skupa projekcija (kako je upravo opisano), takođe nam treba kriterijum praga (otklona) da bismo zaista dobili pravi klasifikator. Dakle prvo treba razmatrati kako pronaći optimalno w a zatim i odgovoriti pitanje praga (Duda, Hart, & Stork, 2000).

Za dobijanje $J(\cdot)$ kao eksplicitne funkcije od w, definišemo *matrice rasipanja* S_i i S_W kao

$$\mathrm{S}_i = \sum_{\mathrm{x} \in D_i} (\mathrm{x} - \mathrm{m}_i)(\mathrm{x} - \mathrm{m}_i)^t$$

i

$$\mathrm{S}_W = \mathrm{S}_1 + \mathrm{S}_2 \, .$$

Matricu S_W nazivamo *unutar-klasnu matricu rasipanja*. Pošto faktor skaliranja za w je nevažan, možemo odmah zapisati rešenje za w koje optimizuje $J(\cdot)$:

$$\mathrm{w} = \mathrm{S}_W^{-1}(\mathrm{m}_1 - \mathrm{m}_2).$$

Time smo dobili w za Fišerov linearni diskriminant – linearnu funkciju koja daje maksimalan odnos među-klasnog i unutar-klasnog rasipanja. Time je klasifikacija pretvorena iz *d*-dimanzionalnog problema u očekivano više smisleni jednodimenzionalni. Ovo mapiranje je više-u-jedan, i teoretski ne može smanjiti minimalnu moguću stopu greške ukoliko imamo veoma veliki skup trening podataka. Generalno, često želimo da žrtvujemo deo teoretski mogućih performansi tačnosti za prednost rada u jednoj dimenziji (Duda, Hart, & Stork, 2000). Još preostaje da se nađe prag (pomeraj), npr. tačka duž koje jednodimenzionalni podprostor razdvaja projekcije tačaka.

Kada su uslovne verovatnoće $p(\mathrm{x}|\omega_i)$ multivarijantne normale sa jednakim matricama kovarijansi Σ, možemo izračunati pomeraj direktno. Optimalna granica odluke ima jednačinu

$$\mathrm{w}^t\mathrm{x} + w_0 = 0$$

gde je

$$w = \Sigma^{-1}(\mu_1 - \mu_2),$$

i gde je w_0 konstanta koja uključuje w i priorne verovatnoće. Ukoliko koristimo srednje vrednosti uzoraka i matricu kovarijansi uzoraka da procenimo μ_i i Σ, dobićemo vektor istog usmerenja kao w iz gornje jednačine koje maksimizira $J(\cdot)$. Tako, za normalan, slučaj jednakih kovarijansi, optimalno pravilo odlučivanja je prosto da odlučimo ω_1 ako Fišerov linearni diskriminator prelazi neki prag (pomeraj), i da odlučimo ω_2 u suprotnom.

Računska kompleksnost nalaženja optimalnog w za Fišerov linearni diskriminator je najviše uslovljena računanjem unutar-klasnog ukupnog rasipanja i njegove inverzije, što je $O(d^2n)$ kompleksnost računanja (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011).

Ukoliko se uvedu određene pretpostavke u vezi raspodele promenljivih koje se koriste u procesu klasifikacije, moguće je razviti dalju statističku proceduru u vezi problema klasifikacije. Ove procedure uključuju testiranje hipoteza u vezi korisnosti nekih ili svih promenljivih na osnovu kojih se vrši klasifikacija, kao i metode za određivanje grešaka pri klasifikaciji, što ćemo detaljnije razmotriti kasnije.

4.3.2.2 Klasifikacija linearnim diskriminacionim funkcijama u više kategorija

Gore smo opisali slučaj klasifikacije u jednu od dve klase. Za probleme sa *c* klasa, prirodno uopštenje Fišerovog linearnog diskriminatora uključuje *c*–1 diskriminacionu funkciju. Prema tome, projekcija je iz d-dimenzionalnog prostora na (*c*–1)-dimenzionalni prostor, pri čemu je prećutno podrazumevano da je $d \geq c$ (Duda, Hart, & Stork, 2000).

Postoji više od jednog načina da se osmisli klasifikator za više kategorija korišćenjem linearnih diskriminacionih funkcija. Na primer, možemo svesti problem na *c*–1 dvo-klasnih problema, gde je *i*-ti problem rešen pomoću linearne diskriminacione funkcije koja razdvaja tačke dodeljene ω_i od onih koje nisu dodeljene ω_i. Nešto ekstravagantniji pristup bi bio da se koristi *c*(*c*–1)/2 linearnih diskriminatora, jedan za svaki par klasa. Oba ova pristupa mogu dovesti do oblasti u kojima je klasifikacija nedefinisana. Taj problem se može izbeći ako se usvoji pristup da se definiše *c* linearnih dmiskriminacionih funkcija

$$g_i(\mathrm{x}) = \mathrm{w}^t \mathrm{x}_i + w_{i0} \qquad i = 1, \dots, c\ ,$$

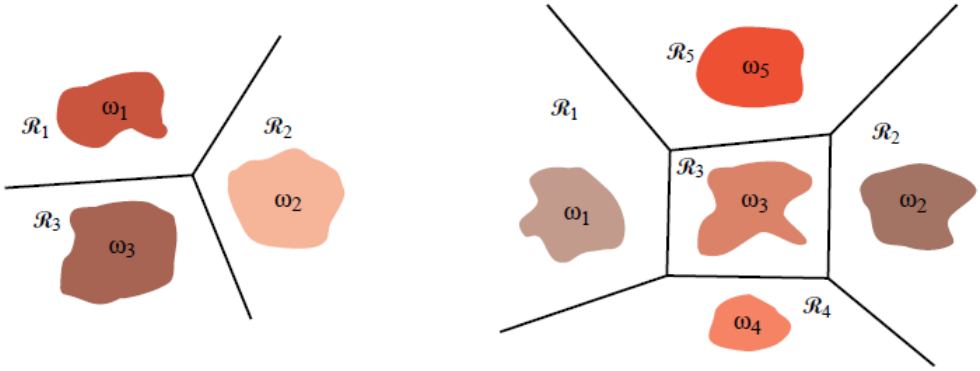
i dodeli x u ω_i ako je $g_i(\mathrm{x}) > g_j(\mathrm{x})$ za svako $j \neq i$; a u „nerešenom“ slučaju, klasifikacija se ostavlja nedefinisanom. Takav klasifikator se naziva *linearna mašina* (eng. *linear machine*) (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011). Linearna mašina deli prostor svojstava u *c* oblasti odlučivanja, sa $g_i(\mathrm{x})$ kao najvećim diskriminatorom ako je x u oblasti R_i. Ako su R_i i R_j granični, granica između njih je deo hiperravni H_{ij} definisanom sa

$$g_i(\mathrm{x}) = g_j(\mathrm{x})$$

ili

$$(\mathrm{w}_i - \mathrm{w}_j)^t \mathrm{x} + \left(w_{i0} - w_{j0}\right) = 0\ .$$

Sledi odjednom da je $\mathrm{w}_i - \mathrm{w}_j$ normalno na H_{ij}, i označena udaljenost od x do H_{ij} je data sa $(g_i - g_j)/\left\|\mathrm{w}_i - \mathrm{w}_j\right\|$. Prema tome, sa linearnom mašinom nisu težinski vektori ti koji su od značaja već su njihove *razlike* od značaja. Pošto ima *c*(*c*–1)/2 parova oblasti, ne moraju sve biti susedne, i ukupan broj segmenata hiperravni koji se javlja u oblastima odlučivanja je često manji od *c*(*c*–1)/2 (Slika 4.3.9) (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011).



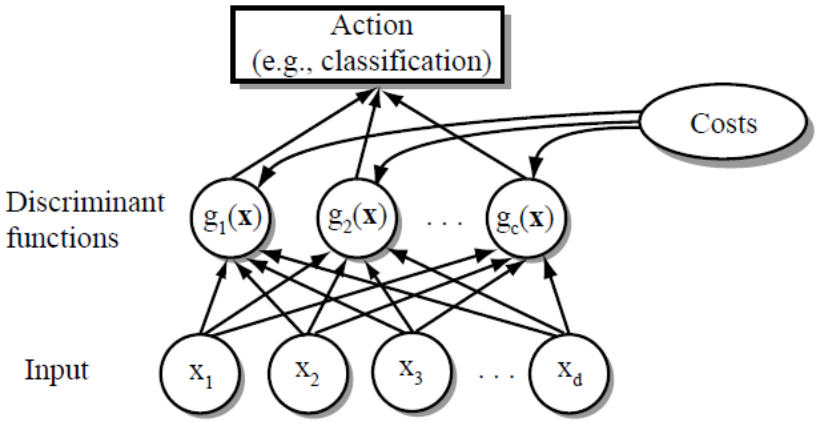
Slika 4.3.9: Granice odlučivanja generisane od strane linearne mašine za probleme sa tri i pet klasa (prostorna diskriminacija prikazana pomoću teritorijalne mape)

Nastavljajući da dodajemo pojmove poput $w_{ijk}x_ix_jx_k$ možemo dobiti klasu *polinomialnih diskriminacionih funkcija*. One se mogu smatrati okrnjenim serijama proširenja proizvoljnog $g(x)$, što predstavlja *uopštenu linearnu diskriminacionu funkciju* (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011)

$$g(x) = \sum_{i=1}^{\hat{d}} a_i y_i(x) \, .$$

4.3.2.3 Klasifikator obrazaca zasnovan na diskriminacionim funkcijama

Već smo videli da postoji više različitih načina da se predstavi klasifikator obrazaca. Jedan od najkorisnijih je pomoću skupa diskriminacionih funkcija $g_i(x)$, $i = 1, ..., c$. Klasifikator će dodeliti vektor svojstava x klasi ω_i ako je $g_i(x) > g_j(x)$ za svako $j \neq i$ (Slika 4.3.10) (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011).

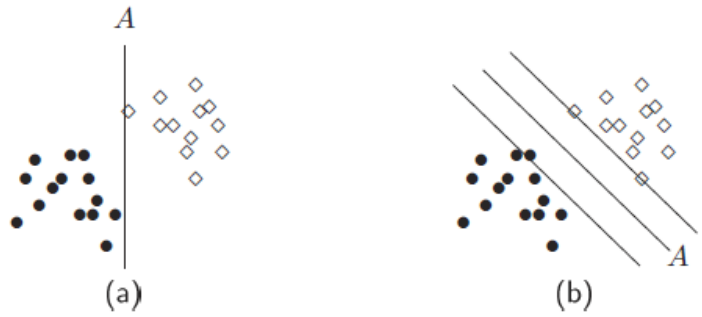


Slika 4.3.10: Funkcionalna struktura opšteg statističkog klasifikatora obrazaca koji uključuje d ulaza i c diskriminacionih funkcija $g_i(x)$.

Na ovaj način se jednostavno i prirodno predstavlja Bajesov klasifikator. Za opšti slučaj, neka je $g_i(\mathrm{x}) = -R(\alpha_i|\mathrm{x})$, s obzirom da će maksimalna diskriminaciona funkcija odgovarati minimalnom uslovnom riziku (pogrešne odluke). Za slučaj najmanje stope greške, možemo dalje uprostiti stvari uzimajući $g_i(\mathrm{x}) = P(\omega_i|\mathrm{x})$, tako da maksimalna diskriminaciona funkcija odgovara maksimumu posteriorne verovatnoće (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011).

4.3.2.4 Mašine podržavajućih vektora (SVM)

Algoritmi za linearne diskriminacione funkcije se mogu primeniti na originalne promenljive ili u trasnformisanom prostoru promenljivih (svojstava) određenom nelinearnim transformacijama originalnih promenljivih. To važi i za *mašine podržavajućih vektora* (eng. *Support Vector Machines*, SVM). One implementiraju prilično jednostavnu ideju – mapiraju vektore obrazaca (entiteta) na višedimenzionalni prostor promenljivih (svojstava) gde je formirana hiperravan koja pruža „najbolju" separaciju (hiperravan sa najvećom marginom) (Slika 4.3.10) (Webb & Copsey, 2011). Na Slici 4.3.10 diskriminaciona hiperravan na 4.3.10.(b) ostavlja najbliže tačke na maksimalnoj udaljenosti. Dve linije na svakoj od strana hiperravni A definišu marginu.



Slika 4.3.10: Dva linearno razdvojiva skupa podataka sa diskriminacionim hiperravnima označenim sa A.

Videli smo ranije kako da istreniramo linearne mašine. Mašine podržavajućih vektora su motivisane većinom istih razmatranja, ali se oslanjaju na pretprocesiranje podataka da da pretstavljaju obrasce u većem broju dimenzija – tipično značajno većem nego što je polazni prostor podataka. Sa odgovarajućim nelinearnim mapiranjem $\varphi()$ na dovoljno veliki broj dimenzija, podaci iz dve kategorije mogu uvek biti razdvojeni pomoću hiperravni. Ovde pretpostavljamo da je svaki entitet x_k transformisan u $\mathrm{y}_k = \varphi(\mathrm{x}_k)$; i postavlja se pitanje izbora $\varphi()$. Za svaki od n entiteta, k=1,2,...,n, neka je z_k=±1, zavisno da li je k iz ω_1 ili ω_2. Linearni diskriminant u uvećanom prostoru y je

$$g(\mathrm{y}) = \mathrm{a}^t\mathrm{y}\,.$$

gde su težinski vektor i trasnformisani vektor podatka uvećani (za a0=w0 i y0=1, respektivno) (Duda, Hart, & Stork, 2000; Webb & Copsey, 2011). Pa diskriminaciona hiperravan osigurava

$$z_k g(\mathrm{y}_k) \geq 1 \quad k = 1, \dots, n\,.$$

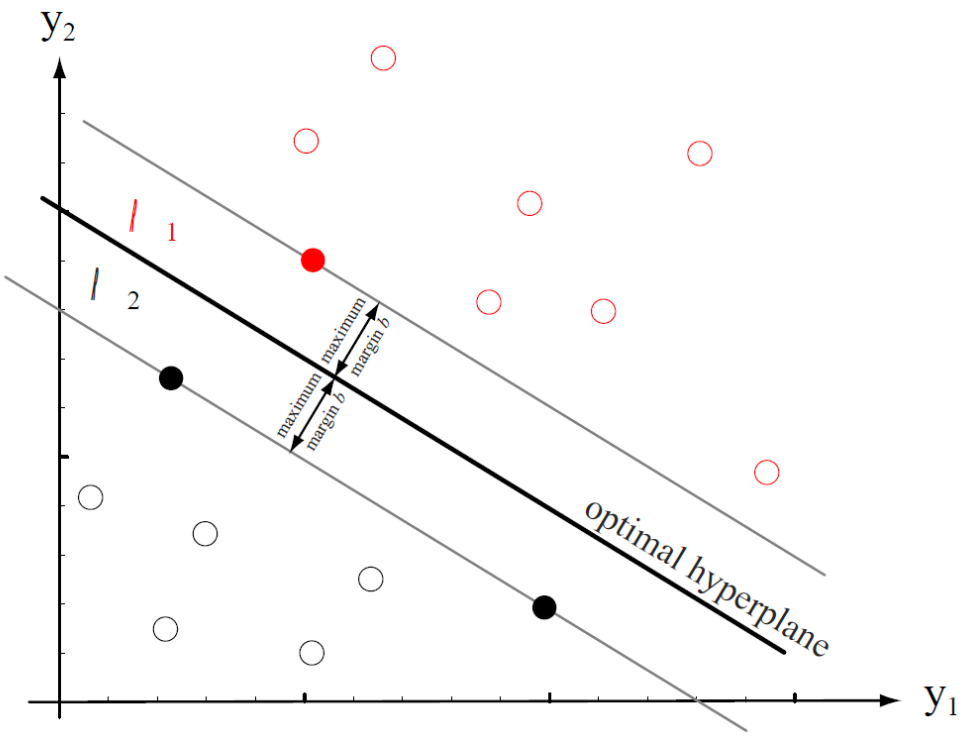
Margina je bilo koja pozitivna udaljenost od hiperravni odluke (diskriminacione hiperravni). Cilj treniranja mašine podržavajućih vektora je nalaženje diskriminacione hiperravni sa najvećom marginom, jer očekujemo da što je veća margina, bolja je generalizacije klasifikatora. Udaljenost hiperravni od (transformisanog) entiteta y je $|g(\mathrm{y})|/\|\mathrm{a}\|$, i pretpostavljajući da postoji pozitivna margina b, prethodna jednačina podrazumeva

$$\frac{z_k g(\mathrm{y}_k)}{\|\mathrm{a}\|} \geq b \quad k = 1, \dots, n;$$

i cilj je naći težinski vektor a koji maksimizira b. Naravno, vektor rešenja se može proizvoljno skalirati dok se i dalje održava hiperravan, te da bi se obezbedila jedinstvenost uvodi se ograničenje $b\|a\| = 1$; tj. tražimo rešenje koje takođe minimizira $\|\mathrm{a}\|^2$ (Duda, Hart, & Stork, 2000).

Podržavajući vektori su (transformisani) trening obrasci za koje prethodna jednačina predstavlja jednakost – podržavajući vektori su (podjednako) blizu hiperravni (Slika 4.3.11). Podržavajući vektori su trening uzorci koji određuju optimalnu diskriminacionu hiperravan i to su entiteti koje je najteže klasifikovati. U informativnom smislu, to su uzorci koji nose najviše informacija za zadatak klasifikacije (Duda, Hart, & Stork, 2000).

Generalno, čitav opisani princip je princip u osnovi mnogih metoda klasifikacije obrazaca: transformisanje ulaznih podataka nelinearno u prostor u kome se linearne metode mogu primeniti (Webb & Copsey, 2011).



Slika 4.3.11: Treniranje Mašine podržavajućih vektora se sastoji od nalaženja optimalne hiperravni, tj. one sa najvećom udaljenošću od najbližih trening uzoraka. Podržavajući vektori su oni najbliži uzorci, sa udaljenošću b od hiperravni. Tri podržavajuća vektora su na slici prikazani obojeni.

4.3.3 Uspešnost klasifikacije

Procena uspešnosti klasifikacije bi trebalo da je deo projektovanja klasifikatora a ne dodatak koji se razmatra odvojeno. Nažalost, često su sofisticirani koraci projektovanja praćeni ne toliko sofisticiranim koracima evaluacije, što često rezultira lošijim pravilima klasifikacije. Kriterijum korišćen za projektovanje klasifikatora je često različit od onoga koji se koristi za njegovu procenu, koji je opet različit od mere performansi pogodne za operativne uslove klasifikatora. Na primer, kod konstrukcije pravila diskriminacije, često se koristi parametri pravila koji optimizuju meru kvadrata greške, dok se za procenu pravila koriste drugačija mera uspešnosti, kao što je stopa greške. Povezani aspekt uspešnosti je onaj koji se odnosi na poređenje performansi više različitih klasifikatora treniranih nad istim skupom podataka. Za praktičnu primenu, možemo implementirati više klasifikatora i želeti da izaberemo najbolji, u smislu merenja stope greške ili možda efikasnosti računanja (Webb & Copsey, 2011).

Ovde ćemo adresirati tri aspekta performansi pravila klasifikacije (Webb & Copsey, 2011):

- *Sposobnost diskriminacije* (eng. *discriminability*) pravila, koji ukazuje koliko dobro pravilo klasifikuje nepoznate podatke, gde se možemo fokusirati na jednom određenom metodu – *stopi greške* (eng. *error rate*) odnosno stopi pogrešne klasifikacije.
- *Pouzdanost* (eng. *reliability*) diskriminacionog pravila (još se naziva i nepreciznost), što pretstavlja meru uspešnosti procene posteriornih verovatnoća pripadnosti klasi od strane datog pravila.
- *Operativne karakteristike prijemnika* (eng. *Receiver Operating Characteristic*, ROC), kao indikator performansi.

Kod projektovanja klasifikatora, često imamo: (a) skup trening podataka koji koristimo za treniranje klasifikatora, (b) skup podataka za validaciju (koje koristimo kao deo procesa treninga) odabira modela ili završetak iteratvinog učenja, i (c) nezavisan skup za testiranje, koji koristimo da merimo performanse generalizacije klasifikatora: sposobnost klasifikatora da bude generički za buduće objekte, tj. koliko dobro klasifikuje nepoznate podatke. Ovi skupovi podataka se često prikupljaju kao deo istog merenja (uzorkovanja) i nije neuobičajeno da su oni, zapravo, različiti delovi istog skupa podataka. U mnogim praktičnim situacijama operativni uslovi se mogu razlikovati od onih koji preovlađuju u momentu prikupljanja testnih podataka, naročito u problemima analize senzorskih podataka. Na primer, karakteristike senzora mogu odstupati vremenom, karakteristike objekta posmatranja se mogu promeniti ili se uslovi okruženja mogu promeniti. Ovi efekti rezultiraju u promeni distribucije iz koje se uzorci izvlače i pojava se naziva *odstupanje (pomeraj) populacije* (eng. *Population drift*) a njhove okolnosti i stepend odstupanja zavise od problema do problema. Sve to predstavlja problem za evaluaciju performansi klasifikatora s obzirom da testni skup podataka možda više nije reprezentativan za operativne uslove. Projektovanje takvih klasifikatora koji se prilagođavaju odstupanju populacije je specifično od problema do problema (Webb & Copsey, 2011).

U primeni diskriminacione analize obično se smatra da su trening entiteti tačno klasifikovani. Klasifikacija trening skupa je često skupa i teška, što se može pokazati na različitim primerima primene diskriminacione analize (npr. kod medicinskih dijagnoza ili klasifikacije podataka sa slika daljinskog uzorkovanja). Veoma važna tačka za razmatranje, pored poteškoća i cene dolaženja do trening opservacija, je da i sama klasifikacija trening podataka može biti podložna greškama. Zaista, koncept tačne dijagnoze je verovatno neprikladan u nekim oblastima medicine, naročito u oblastima poput psihijatrije, ili kod daljinskog uzorkovanja, npr. kod klasifikacije i identifikacije useva, klasifikacija trening piksela se radi vizuelno i može biti podložno greškama. U svakom slučaju ovde govorimo o *pogrešno klasifikovanim trening podacima* (eng. *Misclassified training data*) (McLachlan, 2004).

4.3.3.1 Standardi dobre klasifikacije

U konstrukciji procedure klasifikacije potrebno je minimizirati verovatnoću pogrešne klasifikacije, ili konkretnije, poželjno je minimizirati rezultate loših efekata pogrešne klasifikacije (Radojičić, 2001). Podrazumevajmo ovde slučaj dve kategorije, koji se kasnije može generalizovati. Pretpostavimo da su entiteti iz bilo koje populacije P_1 ili P_2. Klasifikacija entiteta zavisi od vektora promenljivih $X=(x_1,..., x_p)$ koje se odnose na taj entitet. Možemo postaviti pravilo da ukoliko je entiet kategorizovan određenim skupom vrednosti $X_1,...,X_p$ biće klasifikovan kao da dolazi iz populacije P_1, ukoliko ima druge vrednosti biće klasifikovana kao da dolazi iz populacije P_2. Možemo da razmišljamo o entitetu kao tački u *p*-dimenzionalnom prostoru i možemo podeliti prostor na dva regiona: ukoliko opservacija dolazi iz regiona R_1 klasifikovaćemo je kao da dolazi iz populacije P_1, a ako dolazi iz regiona označenog sa R_2 klasifikovaćemo je kao da dolazi iz populacije P_2 (Radojičić, 2001).

U proceduri klasifikacije statističari mogu da naprave dve vrste grešaka klasifikacije i to ukoliko je entitet iz P_1 statističari mogu da je klasifikuju kao da dolazi iz P_2 ili obrnuto. Potrebno je poznavati relativne nepoželjnosti ove dve pogrešne klasifikacije. Neka je sa C(2|1) (>0) označena cena (gubitak) zbog toga što opservacija pripada P_1, a statističari je svrstavaju u P_2 i neka je sa C(1|2) (>0) označena cena (gubitak) zbog toga što opservacija pripada P_2, a statističari je svrstavaju u P_1. Ovi uzroci mogu biti mereni u bilo kom uzorku. Količnik ova dva uzroka je važan, i u slučajevima kada statističari ne znaju za gubitke u ova dva slučaja vrlo često imaju grube ideje o njima. Tabela 4.3.1 prikazuje gubitke (troškove) dobre i loše klasifikacije. Dobra procedura klasifikacije je ona koja minimizira u nekom smislu tu cenu pogrešne klasifikacije (Radojičić, 2001).

Tabela 4.3.1: Klasifikacija u jednu od dve populacije

		Predviđena klasifikacija	
	Klasifikacija	P1	P2
Opservirana klasifikacija	P1	C(1/1)=0	C(1/2)
	P2	C(2/1)	C(2/2)=0

4.3.3.2 Metode provere uspešnosti klasifikacije

Definicija diskriminacionih funkcija kao sredstva za klasifikaciju entiteta bazirana je na minimiziranju ukupne verovatnoće pogrešne klasifikacije (Kovačić, 1994). U ranijim izrazima za oblasti klasifikacije, odnosno pravilima za klasifikaciju, kao poznate veličine figurišu funkcije gustine verovatnoća. U tom slučaju relativno je jednostavno

izračunati minimalnu vrednost ukupne verovatnoće pogrešne klasifikacije. Određivanje verovatnoća pogrešne klasifikacije P(2|1) ili P{*2 dato 1*} i P(1|2) ili P{*1 dato 2*} (verovatnoće da se entitet iz klase 1 klasifikuje pod klasom 2 i obrnuto, tj. šrafirane površine pod krivom na slici 4.3.8), kao elemenata ukupne verovatnoće pogrešne klasifikacije, predstavlja problem u teoriji verovatnoće koji se rutinski rešava. Na osnovu izračunatih vrednosti tih verovatnoća moguće je proceniti kakav je kvalitet diskriminacionih funkcija pri klasifikaciji entiteta. Interesantniji je slučaj kada funkcije gustine verovatnoća nisu poznate. Tada, na osnovu slučajnog uzorka, treba proceniti u kojoj meri diskriminaciona funkcija uspešno klasifikuje entitete u odgovarajuće populacije, a u koliko slučajeva to radi pogrešno.

Dakle, uspešnost klasifikacije entiteta može se meriti pomoću dve verovatnoće za pogrešnu klasifikaciju, verovatnoće P{*2 dato 1*} i verovatnoće P{*1 dato 2*}. Postoje različite metode za utvrđivanje ovih verovatnoća. Jedna od metoda, poznata kao *empirijska metoda*, prikazana je u Tabeli 4.3.2, i predstavlja određeni način provere diskriminacione funkcije. Da bi mogli da izmerimo nivo uspešnosti postupka klasifikacije, potrebno je izbrojati koliko pripadnika svake grupe je korektno a koliko je pogrešno klasifikovano, što je upravo analiza sadržana je u Tabeli 4.3.2.

Tabela 4.3.2: Utvrđivanje valjanosti klasifikacije

Aktuelni status entiteta	Status entiteta dobijen predviđanjem		Ukupno	Procenat korektno klasifikovanih
	Populacija 1	Populacija 2		
Populacija 1	*a*	*b*	*a+b*	*a/(a+b)*100*
Populacija 2	*c*	*d*	*c+d*	*d/(c+d)*100*
Ukupno	*a+c*	*b+d*	*a+b+c+d*	*(a+d)/(a+b+c+d)*100*

Iako je metoda intuitivno jasna i jednostavna, ipak rezultuje izvesnim pristrasnim procenama (Radojičić, 2001), jer rezultujući udeli koji se dobijaju na kraju potcenjuju pravu verovatnoću pogrešne klasifikacije. Ovo se dešava iz razloga što je isti uzorak korišćen za računanje i proveru diskriminacione funkcije. U idealnom slučaju diskriminaciona funkcija se računa iz jednog uzorka, a zatim se primenjuje na drugom, kako bi se utvrdio udeo pogrešno klasifikovanih entiteta. Ovaj postupak se naziva *unakrsna provera*, i, kao rezultat, dobijaju se nepristrasne procene. Moguće je sprovesti unakrsnu proveru tako što se, na slučajan način, originalni uzorak podeli na dva poduzorka: jedan za računanje diskriminacione funkcije i drugi za njenu unakrsnu proveru.

Alternativna metoda koja se koristi u slučaju malog uzorka, a imitira podelu uzorka na poduzorke, naziva se *jackknife* procedura. Ovde se jedan od entiteta isključuje iz prve grupe, i nakon toga se računa diskriminaciona funkcija na osnovu preostalih

entiteta. Zatim se vrši klasifikacija izostavljenog entiteta i ova procedura se ponavlja za svaki entitet iz prve grupe. Udeo pogrešno klasifikovanih entiteta je *jackknife* procena *P{2 dato 1}*. Slična procedura se koristi za utvrđivanje *P{1 dato 2}*. Ova metoda daje približno nepristrasne procene (Radojičić, 2001).

Ukoliko prihvatimo da model ima normalnu raspodelu sa više promenljivih, tada je moguć i teorijski način procene verovatnoća. U tom slučaju zahteva se da je poznato samo Mahalanobisovo odstojanje D^2. Formule su:

$$\frac{K-1/2D^2}{D} \qquad\qquad \frac{K-1/2D^2}{D}$$

procenjeno P{2 dato 1}=površina levo procenjeno P{1 dato 2}=površina levo

gde je:

$$K = \ln \frac{q_2 \cdot g\{1\,dato\,2\}}{q_1 \cdot g\{2\,dato\,1\}}$$

Ukoliko je K=0, ove dve promene su jednake površini koja se nalazi levo od (-D/2) pod uslovom standardizovane normalne raspodele. Ovaj postupak je naročito koristan ukoliko se diskriminaciona funkcija dobija iz regresionog modela, obzirom da D^2 lako može biti proračunato iz R^2.

Ova metoda takođe potcenjuje prave verovatnoće pogrešne klasifikacije. *Nepristrasna procena* populacije Mahalanobis D^2 je (Radojičić, 2001):

$$\text{nepristrasno}\quad D^2 = \frac{N_1+N_2-P-3}{N_1+N_2-2} D^2 - P\left(\frac{1}{N_1}+\frac{1}{N_2}\right)$$

Da bi analizirali kako se određena diskriminaciona funkcija ponaša, moguće je izračunati verovatnoću tačnog predviđanja na osnovu čistog *pogađanja (slučajna klasifikacija)*. Taj postupak je sledeći: pretpostavimo da je apriori verovatnoća pripadanja prvoj grupi poznata i iznosi q_1. Tada je $q_2 = 1 - q_1$. Jedan od načina da se klasifikuju entiteti je korišćenjem samo ovih verovatnoća. Zatim se računa ukupna verovatnoća tačne klasifikacije. Verovatnoća da entitet koja pripada prvoj grupi bude *tačno* klasifikovan iznosi q_1^2. Slično, q_2^2 je verovatnoća da entitet koji pripada drugoj grupi bude tačno klasifikovan. Tako je ukupna verovatnoća tačne klasifikacije korišćenjem samo apriori verovatnoća $q_1^2 + q_2^2$. Može se primetiti da se najniža moguća vrednost ove verovatnoće javlja kada je $q_1 = 0.5$, tj. kada je podjednako verovatno da entitet pripada obema grupama (Radojičić, 2001).

4.3.3.3 Merila uspešnosti klasifikacije

Da bi se definisala mera uspešnosti klasifikacije koja se ne zasniva na poznatom ili pretpostavljenom obliku funkcije gustine verovatnoće, neophodno je kreirati *matricu konfuzije* (eng. *confusion matrix*) koja prikazuje broj ispravno i pogrešno klasifikovanih entiteta po grupama (Webb & Copsey, 2011; Kovačić, 1994). Iz matrice konfuzije moguće je izračunati mnoge mere performansi klasifikacije. Tabela 4.3.3 prikazuje 2×2 matricu konfuzije za klasifikator u dve klase (pozitivnu i negativnu), odakle je uzeta veličina uzorka od po P=TP+FN i N=FP+TN opservacija respektivno (Webb & Copsey, 2011; Kovačić, 1994). Ona prikazuje broj tačnih i netačnih predikcija za svaku od dve klase. *Tačni pozitivi* (eng. *true positives*, TP) je broj članova pozitivne klase koji su tačno predviđeni (klasifikovani) od strane klasifikatora da pripadaju pozitivnoj klasi. *Netačni pozitivi* (eng. *false positivies*, FP) je broj članova negativne klase koji su netačno predviđeni da pripadaju pozitivnoj klasi. Tabela 4.3.4 popisuje neke od uobičajenih mera performansi izvedenih iz matrice konfuzije (Webb & Copsey, 2011). Mnogi od njih su prošireni za probleme sa više klasa.

Tabela 4.3.3: Matrica konfuzije 2×2.

		Tačna klasa	
		Pozitivna	Negativna
Predviđena klasa	Pozitivna	Tačni pozitivi (TP)	Netačni pozitivi (FP)
	Negativna	Netačni negativi (FN)	Tačni negativi (TN)

Tabela 4.3.4: Mere performansi klasifikacije izvedene iz 2×2 matrice konfuzije.

Tačnost (*Accuracy*, *Acc*)	$\frac{TP + TN}{P + N}$
Stopa greške (*Error rate*, *E*)	$1 - Acc$
Stopa netačnih pozitiva (*False positive rate*, *fpr*), Stopa lažnih alarma (*False alarm rate)*	$\frac{FP}{N}$
Stopa tačnih pozitiva (*True positive rate*, *tpr*), Opoziv (*Recall*, *Rec*), Osetljivost (*Sensitivity*, *Sens*)	$\frac{TP}{P}$
Preciznost (*Precision*, *Prec*)	$\frac{TP}{TP + FP}$
Specifičnost (*Specificity*, *Spec*)	$\frac{TN}{N}$
F-mera (*F measure*)	$\frac{2}{\frac{1}{Prec} + \frac{1}{Rec}}$

4.3.3.4 Testiranje hipoteza

Postavlja se pitanje da li je moguće klasifikovati entitete korišćenjem raspoloživih promenljivih bolje nego što bi to bilo slučajnom klasifikacijom. Ovo pitanje može biti formulisano kao problem za testiranje hipoteza. Nulta hipoteza koja se testira može se definisati kao da nijedna od promenljivih neće poboljšati rezultate tzv. slučajne klasifikacije. Ekvivalentna nulta hipoteza je da su srednje vrednosti obe populacije za svaku promenljivu jednake, odnosno da je populaciono Mahalanobisovo odstojanje D^2 nula. Statistički test za nultu hipotezu je Fišerov test:

$$F = \frac{N_1 + N_2 - P - 1}{P(N_1 + N_2 - 2)} \cdot \frac{N_1 \cdot N_2}{N_1 + N_2} \cdot D^2$$

sa stepenima slobode (P), (N_1+N_2–P–1) (Rao, 1973). Takođe, izraz $N_1N_2D^2/(N_1+N_2)$ je poznat kao dvouzorački *Hotelling* T^2, koji je originalno razvijen za testiranje jednakosti dva skupa srednjih vrednosti (Morrison, 1976). Međutim, iz izračunate vrednosti gore navedenog F, može se izračunati D^2 na sledeći način:

$$D^2 = \frac{P \cdot (N_1 + N_2) \cdot (N_1 + N_2 - 2)}{(N_1 \cdot N_2) \cdot (N_1 + N_2 - P - 1)} \cdot F$$

Sledeći koristan test odnosi se na to da li uvođenje dodatne promenljive u model poboljšava diskriminaciju metoda (Radojičić, 2001). Pretpostavimo da je populaciono D^2 bazirano na promenljivim X_1, X_2, ..., X_p predstavljeno preko D^2_P. Suština ovog testa je ispitati da li dodatna promenljiva X_{P+1} može značajno da poveća vrednost D^2_p, tj. testira se nulta hipoteza da je $D^2_{P+1} = D^2_P$. Uzimajući u obzir višedimenzionalni model normalne raspodele, vrednost statistike F testa je:

$$F = \frac{(N_1 + N_2 - P - 2) \cdot (N_1 \cdot N_2) \cdot (D^2_{P+1} - D^2_P)}{(N_1 + N_2) \cdot (N_1 + N_2 - 2) + N_1 \cdot N_2 \cdot D^2_P}$$

sa (1),(N_1+N_2–P–2) stepeni slobode (Rao, 1964).

Sprovođenje *stepwise diskriminacione analize* vrši se na osnovu istih koncepata koji se koriste kod *stepwise regresione analize*. Pri ovoj analizi diskriminacione funkcije, umesto da se testira da li se vrednost višestrukog R^2 menja kada se dodaje (ili oduzima) promenljiva, testira se da li se vrednost D^2_p menja pri dodavanju ili oduzimanju promenljive. Za ovu svrhu koristi se statistički F test; moguće je odrediti vrednosti F-*za-unos* i/ili F-*za-izbacivanje* promenljive (Radojičić, 2001).

5 METODOLOGIJA KLASIFIKACIJE U DALJINSKOM UZORKOVANJU

Klasifikacija u GIS sistemu je tehnika namernog uklanjanja detalja iz skupa ulaznih podataka, kako bi se ustanovili važni obrasci (prostornog rasporeda). To radimo dodeljujući karakterističnu vrednost svakom elementu u ulaznom skupu podataka, koji je obično skup ulaznih svojstava koji mogu biti rasterske ćelije ili tačke, linije ili poligoni. Ako je broj karakterističnih vrednosti mali u odnosu na veličinu ulaznog skupa podataka, to znači da smo klasifikovali ulazni skup (Huisman & de By, 2009). Kod slika prikupljenih daljinskim uzorkovanjem, izmerena vrednost refleksije na slici zavisi od lokalnih karakteristika zemljine površine – postoji veza između zemljišnog pokrivača i izmerene reflektovane vrednosti. Da bi se izvukle informacije sa slike, ta veza se mora pronaći, a proces nalaženja te veze se zove klasifikacija. Klasifikacija se može izvesti korišćenjem jednog spektralnog opsega, npr. u procesu koji se zove *sečenje gustine* (eng. *density slicing*), ili korišćenjem većeg broja opsega (multispektralna klasifikacija) (Levin, 1999).

Možemo za početak konstatovati da klasifikacija u GIS-u može biti kontrolisana od strane korisnika ili automatska klasifikacija. *Klasifikacija kontrolisana od strane korisnika* je takva gde korisnik selektuje atribute koji će biti korišćeni kao parametri klasifikacije i definiše metod klasifikacije. To uključuje određivanje broja klasa kao i kompletnog preslikavanje između starih vrednosti atributa i novih vrednosti klasa. Klasifikacija koju kontroliše korisnik zahteva intenzivnu interakciju sa korisnikom ili unapred pripremljenu tabelu klasifikacije, koja definiše sva preslikavanja (Huisman & de By, 2009). *Automatska klasifikacija* može biti postignuta pomoću GIS softvera, npr. tako da korisnik samo specificira broj klasa u izlaznom skupu podataka. Sistem onda automatski određuje granične tačke između klasa, što se postiže pomoću jedne od dve tehnike: (a) *tehnika jednakih intervala* (ukupan interval vrednosti između minimalne i maksimalne vrednosti se podali na *n* jednakih intervala), što je pogodno za otkrivanje obrazaca distribucije pomoću broja svojstava (atributa) u svakoj od kategorija, i (b) *tehnika jednakih frekvencija* (gde se kreiraju kategorije sa približno jednakim brojem svojstava po kategoriji) (Huisman & de By, 2009).

Reklasifikacija skupa podataka je najčešće jedan od prvih koraka analize rasterskih podataka. Ona je u suštini proces dodeljivanja nove klase ili raspona svim pikselima u skupu podataka, na osnovu njihovih originalnih vrednosti (Slika 5.1). Na primer, koordinatna mreža elevacije praktično sadrži različite vrednosti svih piksela, ali to se može uprostiti agregirajući vrednosti piksela u nekoliko diskretnih klasa (Campbell & Shin, 2011).

Ulazni raster

456	416	364	326	243
448	364	315	276	218
359	325	268	234	164
306	296	201	133	44
274	231	184	65	5

Reklasifikovan raster

5	5	4	4	3
5	4	4	3	3
4	4	3	3	2
4	3	3	2	1
3	3	2	1	1

Slika 5.1: Reklasifikacija rastera

U rešavanju problema merenja i nadgledanja podataka kod geoinformacionih sistema, problem klasifikacije pomoću daljinski uzorkovanih merenja je veoma značajan. Za rešavanje tih problema razvijeni su različiti algoritmi prepoznavanja slika, statističkog odlučivanja i analize grupisanja (Krapivin, Varotsos, & Soldatov, 2015).

5.1 Digitalne slike i njihova obrada u daljinskom uzorkovanju

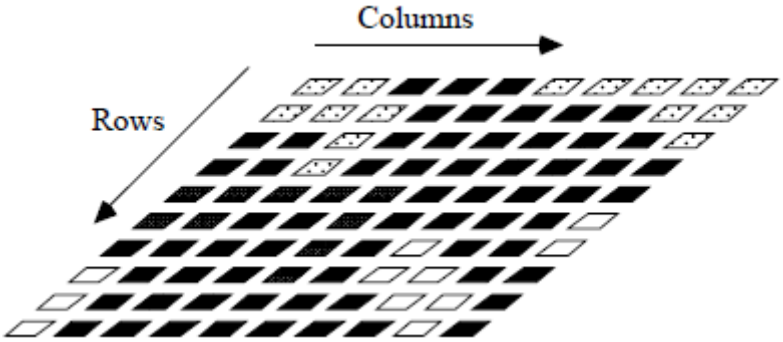
5.1.1 Digitalne slike i njihova reprezentacija

Digitalne slike su fotografije koji su konvertovane u računarski čitljiv (binarni) format. Obično pod slikom smatramo nepomičan vizuelni zapis, koji se ne menja kroz vreme (Bovik, 2009). Slike refleksije beleže zračenje koje se reflektuje od površine objekta, slike emisije beleže zračenje koje emituju samo-osvetljeni objekti, dok slike apsorpcije beleže informacije o unutrašnjoj strukturi objekata. Važno svojstvo digitalnih slika je da su one multidimenzionalni signali, odnosno da su funkcije više od jedne jedinstvene promenljive. Dimenzije signala je broj koordinata potrbnih da se indeksira i da se okarakteriše svaka tačka na slici (Bovik, 2009).

Jedan od osnovnih aspekata obrade slika je reprezentacija slike. U osnovi svih oblika prikaza slike je koncept da je digitalna slika matrica brojeva, odnosno dvodimenzionalan niz piksela. Takođe je bitno interpretirati kako su te vrednosti vezane za fizičku scenu koju slika prikazuje. Slika se može opisati pomoću funkcije f (x, y, λ, t), gde su x i y prostorne koordinate, λ ukazuje na talasnu dužinu zračenja u t predstavlja vreme. Slike se mogu smatrati 2D prostornim raspodelama, ali se više dimenzija može predstaviti prostim proširenjem. Za nas su od značaja slike koje se mogu predstaviti pomoću dve prostorne koordinate i koordinatom talasne dužine (boje), f (x, y, λ). Takva kontinualna slika se sempluje u svim dimenzijama i rezultat je funkcija definisana u diskretnom koordinatnom sistemu, f (m, n, l), što zahteva 3D matricu (Bovik, 2009). Takva slika kod koje su x, y i amplituda vrednosti λ konačne, diskretne vrednosti, se

naziva digitalna slika a oblast obrade digitalnih slika se odnosi na obradu digitalnih slika pomoću digitalnih računara (Gonzalez & Woods, 2002).

Semplovanje (uzimanje uzorka, eng. *Sampling*) je proces konvertovanja signala iz kontinualnog u diskretan prostor. Semplovana slika je niz semplovanih vrednosti koje su obično poređane u formi redova i kolona. Svaki indeksirani element tog niza se obično naziva elementom slike, ili skraćeno piksel (od eng. *picture element*, tj. *pel* ili *pixel*) (Slika 5.1.1), što je najčešće korišćeni termin za nazivanje elemenata digitalne slike (Bovik, 2009; Gonzalez & Woods, 2002).



Slika 5.1.1: Prikaz veoma malog (10x10) dela niza slike.

Kvantizacija (eng. *quantization*) je drugi deo digitalizacije slike, što je proces konvertovanja kontinualnih vrednosti stvarne slike u diskretne vrednosti. Npr. ako je slika vizuelno predstavljena pomoću nijansi sive (kao crno-bela slika), onda su vrednosti piksela nijanse (nivoi) sive, ili slika može imati više vrednosti za svaki piksel (npr. slika u boji). Rezultat semplovanja i kvantizacije je matrica realnih brojeva, koja predstavlja sliku (Bovik, 2009; Gonzalez & Woods, 2002).

Količina podataka kod vizuelnih signala je veoma velika i povećava se geomatrijski sa dimenzionalnošću podataka. Ovo ima značajan uticaj na sve aspekte obrade slike – obim podataka je jedno od najvažnijih pitanja u obradi, čuvanju, prenosu, i prikazu slika (Bovik, 2009).

5.1.2 Obrada digitalnih slika

Operacije obrade digitalnih slika datih kao nijanse sive uključuju tri vrste (Bovik, 2009):

- *Operacije nad tačkama*, ili operacije obrade slike, koje se primenjuju samo na individualne piksele. Osnovni alat za razumevanje, analizu i projektovanje operacija nad tačkama je *histogram* slike.
- *Aritmetičke operacije između slika istih prostornih dimenzija*. Ovo su takođe operacije nad tačkama (bez prostornih informacija), ali se

informacije dele između slika tačku po tačku. Ovo su operacije specijalne namene, poput redukcija šuma ili detekcija promena i kretanja.

- *Geometrijske operacije nad slikom*, koje su komplamentarne operacijama nad tačkama i umesto funkcija intenziteta slike one su u funkciji prostorne pozicije – one menjaju izgled slike menjajući koordinate intenziteta. Npr. translacija i rotacija slike.

Histogram slike (eng. *Image Histogram*) je osnovni alat za operacije nad tačkama kod digitalnih slika (Bovik, 2009). Histogram otkriva radiometrijske osobine digitalne slike i opisuje raspodelu vrednosti piksela. On pokazuje broj piksela koji imaju svaku od vrednosti u rasponu vrednosti, tj. raspodelu frekvencija digitalnih brojeva (DN) (Tempfli, Kerle, Huurneman, & Janssen, 2009). Drugim rečima, histogram H_f digitalne slike f je grafikon frekvencije pojavljivanja svake nijanse (npr. nijanse sive) u f. Histogram digitalne slike sa nijansama sive H_f je diskretna jednodimenzionalna funkcija sa domenom {0, ..., K-1}, što je raspon koji uzimaju različite nijanse sive, i mogućim opsegom od 0 do broja piksela na slici. Histogram je eksplicitno dat sa

$$H_f(r_k) = n_k$$

gde je r_k k-ta nijansa sive i n_k je broj piksela na slici koji imaju nijansu sive r_k, za svako k = 0, ..., K-1 (Bovik, 2009; Gonzalez & Woods, 2002). Uobičajena je praksa da se normalizuje histogram deljenjem svake od vrednosti ukupnim brojem piksela na slici, označenim sa n. Onda je normalizovani histogram dat sa $P_f(r_k) = n_k / n$, za k = 0, ..., K-1. Uopšteno govoreći, $P(r_k)$ daje procenu verovatnoće pojavljivanja nijanse sive r_k (Gonzalez & Woods, 2002). Suma svih komponenti normalizovanog histograma je 1.

Pošto histogram predstavlja redukciju dimenzionalnosti, kod njega dolazi do gubitka informacija (slika ne može biti dedukovana iz histograma, osim u trivijalnim slučajevima). Histogram ne sadrži prostorne informacije o f, već samo opisuje frekvenciju nijansi i ništa više. Ipak, ova informacija je svejedno „bogata“ i histogram pruža korisne statističke podatke slike. Mnoge korisne operacije obrade slika mogu biti izvedene iz histograma slike, npr. kompresija ili segmentacija slike, koja je za nas od naročitog značaja (Bovik, 2009; Gonzalez & Woods, 2002). Ponekad je korisno da možemo specificirati oblik histograma koji želimo da obrađena slika ima. Metod koji se koristi za generisanje obrađene slike koja ima određeni histogram se naziva *podudaranje* ili *specifikacija histograma* (eng. *Histogram Matching* ili *Histogram Specification*) (Gonzalez & Woods, 2002).

Jednostavan ali moćan alat za identifikaciju i označavanje različitih objekata kod binarnih slika je proces koji se naziva *označavanje regiona*, *blob coloring* ili *identifikacija spojenih komponenti*. Taj proces identifikuje spojene grupe piksela na binarnoj slici f, koji imaju istu binarnu vrednost. Koristan je jer omogućuje da se objektima manipuliše odvojeno nakon označavanja (Bovik, 2009). Ciljevi unapređenja

slike uključuju poboljšanja vidljivosti i primetljivosti različitih regiona slike i mogućnosti detekcije svojstava slike unutar tih regiona. To uključuje zadatke poput čišćenje slike od šuma, povećavanje kontrasta, uprošćavanje pomoću izglađivanja ili eliminacije nekih karakteristika ili uzimanja samo nekih karakteristika. Takva unapređenja slike obično prati detekcija svojstava poput ivica, vrhova, ili drugih geometrijskih svojstava. Mnogi problemi uključuju detekciju šablona, što se obično rešava pomoću *uparivanja šablona* (eng. *Template Matching*) (Bovik, 2009).

Neke obrade koriste operacija susedstva da rade sa vrednostima piksela u okruženju (susedstvu) i odgovarajućim vrednostima pod-slike koja ima iste dimenzije kao to susedstvo. Pod-slika se naziva *filter*, *maska*, *jezgro* (eng. *kernel*), *šablon* (eng. *template*) ili *prozor*, pri čemu se prva tri naziva najčešće koriste. Vrednosti u filterskoj pod-slici se češće nazivaju *koeficijentima* nego pikselima. Koncept filtriranja ima osnove u korišćenju Furijeovih transformacija za obradu signala, a može se odnositi na prostorno filtriranje (koje se primenjuje direktno na piksele slike) ili više tradicionalno filtriranje u frekventnom domenu (Gonzalez & Woods, 2002).

Ne postoji usaglašen stav među autorima gde *obrada slika* prestaje i druge povezane oblasti, poput *analize slika* ili *računarske vizije*, počinju. Ponekad se pravi razlika tako što se obrada slika definiše kao disciplina kod koje su i ulaz i izlaz iz procesa obrade slike. Sa druge strane, postoje oblasti poput računarske vizije čiji je krajnji cilj korišćenje računara za emulaciju ljudskog vida, uključujući učenje, što je grana oblasti veštačke inteligencije. Oblast analize slika (takođe nazivana i razumevanje slika) je između obrade slika i računarske vizije (Gonzalez & Woods, 2002). U tim definicijama ne postoje jasne granice u kontinuumu od obrade slika do računarske vizije. Međutim korisna paradigma je da razmatramo tri vrste kompjuterizovanih procesa u tom kontinuumu (Gonzalez & Woods, 2002):

- *Procesi niskog nivoa* uključuju primitivne operacije, poput pretprocesiranja za smanjenje šuma, unapređenje kontrasta, i izoštravanje slika. Kod ovih procesa su i ulaz i izlaz iz procesa slike.
- *Procesi srednjeg nivoa* uključuju zadatke poput segmentacije (particionisanja slike u oblasti ili objekte), opisivanje objekata kako bi se redukovali u formu pogodnu za računarsku obradu, i klasifikacija (prepoznavanje) pojedinačnih objekata. Procesi srednjeg nivoa su okarakterisani činjenicom da su ulazi slike, dok su izlazi atributi ekstrahovani iz tih slika (npr. ivice, konture, i kategorije pojedinačnih objekata).
- *Procesi višeg nivoa* uključuju „davanje smisla“ skupu prepoznatih objekata, poput analize slike, i na samom kraju kontinuuma, obavljanje kognitivnih funkcija obično povezanih sa vidom.

Na osnovu datih definicija, možemo konstatovati da je logično mesto preklapanja između obrade slika i analize slika oblast prepoznavanja pojedinačnih regiona ili objekata na slici (Gonzalez & Woods, 2002). Osnovni koraci u obradi digitalnih slika su: prikupljanje, poboljšanja, restauracija, obrada slika u boji (kolor slika), kompresija, morfološka obrada, segmentacija, reprezentacija i opisivanje, i prepoznavanje (Gonzalez & Woods, 2002).

5.1.3 Klasifikacija digitalnih slika

Klasifikacija slike je jedna od tehnika iz domena interpretacije digitalnih slika. Kod tog procesa ljudski operator daje instukcije računaru da obavi takvu interpretaciju prema određenim uslovima, koje on definiše (Tempfli, Kerle, Huurneman, & Janssen, 2009). Kod klasifikacije slika na osnovu objekata, pikseli se spajaju u objekte, a „grupisanje“ se sprovodi prema složenim kriterijumima koji uključuju oblik, teksturu, osenčenost, veličinu, međusobne veze, i ne samo vrednosti piksela (Jovanović, i drugi, 2015). Prvi preduslov za automatizaciju takvog načina ekstrakcije objekata iz slike je postojanje procedure koja će podeliti sliku u smislene grupe podataka koje mogu predstavljati pojedinačne objekte. Ovaj korak se naziva *segmentacija slike* i on je osnova za klasifikaciju zasnovanu na objektima (Jovanović, i drugi, 2015). Procedure segmentacije particionišu sliku na njene konstitutivne delove ili objekte a autonomna segmentacija je jedan od najtežih zadataka kod obrade digitalnih slika. Generalno, što je segmentacija urađena uspešnije, to je verovatnije da će prepoznavanje i klasifikacija objekata biti uspešni (Gonzalez & Woods, 2002). Naime, hiperspektralne slike kod daljinskog uzorkovanja verovatno imaju visok nivo sličnosti između susednih piksela i korišćenje ove osobine segmentiranjem sličnih oblasti poboljšava performanse klasifikacije (Prabhakar, Gintu, Geetha, & Soman, 2015). Problem segmentacije se može formulisati tako da za sliku predstavljenu kao $x \equiv (x_1, \dots, x_n)$ sa $I \equiv \{1, \dots, n\}$ skupom celobrojnih indeksa i $L \equiv \{1, \dots, K\}$ skupom imenovanih klasa, segmentacija, za razliku od klasifikacije, ima cilj da particioniše dati skup *I* u skupove (oblasti, regione) $R_i \subset I$ gde je $i = 1, \dots, K$ tako da su pikseli u svakoj od oblasti slični na neki način (Prabhakar, Gintu, Geetha, & Soman, 2015). Proces uključivanja svih raspoloživih kontekstualnih informacija poboljšava ukupnu preciznost. Onda je, primera radi, *segmentacija maksimalnim posteriorima* (eng. *maximum a posteriori*, MAP) u okvirima Bajesovog odlučivanja data kao (Prabhakar, Gintu, Geetha, & Soman, 2015)

$$\hat{y} = \arg\max_{y \in L} \left\{ \sum_{i=1}^{n} (\log p(y_i|x_i) - \log p(y_i)) + \log p(y) \right\}$$

Problem segmentacije je takođe usko povezan sa problemom detekcije kontura. Ta dva problema su povezana ali ne identična – detekcija kontura ne daje garanciju da će pronađene konture biti zatvorene, tako da ne moraju dati sliku particionisanu u

regione. S druge strane, na osnovu podele u regione, uvek se mogu izvesti konture (Arbelaez, Maire, Fowlkes, & Malik, 2011). Neki od algoritama segmentacije slika su (Gonzalez & Woods, 2002): detekciju diskontinualnosti (eng. *Detection of Discontinuities*), vezivanje ivica i detekciju granica (eng. *Edge Linking and Boundary Detection*), određivanje granične vrednosti (eng. *Thresholding*), segmentaciju na osnovu regiona (eng. *Region-Based Segmentation*), segmentaciju morfološkim slivovima (eng. *Segmentation by Morphological Watersheds*), i *segmentacija slike pomoću grafa* (eng. *Graph-Based Image Segmentation*) (Felzenszwalb & Huttenlocher, 2004).

Kod procesa obrade slika različitih nivoa, od značaja je i potreba za priornim znanjem ili interakcija između baze znanja i modula za obradu. Znanje o problemskom domenu je sadržano u sistemu za obradu slika u formi baze znanja. Ovo znanje može biti jednostavno poput preciziranje oblasti slike gde se informacija od značaja može nalaziti, kako bi se ograničila oblast pretrage, a može biti i prilično kompleksno, poput međusobno povezanih lista mogućih defekata u problemima inspekcije materijala ili baza slika koja sadrži satelitske snimke visoke rezolucije oblasti povezanih sa primenama detekcije promena (Gonzalez & Woods, 2002).

5.1.4 Klasifikacija multispektralnih digitalnih slika

Digitalna slika je 2D niz piksela. Vrednost piksela, digitalni broj (eng. *Digital Number*, DN), je u slučaju 8-bitnog zapisa u rasponu od 0 do 255. DN odgovara reflektovanoj ili emitovanoj energiji sa tla (osim ukoliko je slika re-semplovana). Prostorna raspodela digitalnih brojeva definiše sliku ili prostor slike. Multispektralni senzor beleži zračenje jedne *ćelije zemaljske rezolucije* (eng. *Ground Resolution Cell*, GRC) u različitim kanalima shodno njegovom razdvajanju spektralnih opsega. Senzor koji snima u tri opsega praktično daje tri piksela sa istim parom red-kolona (*i*, *j*) koji proizilaze iz jedne iste ćelije zemaljske rezolucije (Tempfli, Kerle, Huurneman, & Janssen, 2009).

Klasifikacija multispektralnih slika služi za ekstrakciju tematskih informacija iz satelitskih i drugih snimaka na poluautomatski način. Posmatranjem određenog piksela slike u M opsega simultano, M vrednosti se posmatra u isto vreme. Korišćenjem multispektralnih slika (npr. SPOT slika), gde je M=3, tri vrednosti refleksije su date za svaki piksel. Npr. (34, 25, 117) u jednom pikselu, u drugom (34,24,119) i trećem (11, 77, 51). Ove vrednosti pronađene za jedan piksel u nekoliko opsega se nazivaju *vektorima svojstava* (eng. *feature vectors*). Može se prepoznati da su prva dva skupa vrednosti prilično slične a da je treći različit od ta dva. Prva dva verovatno pripadaju istoj klasi (zemljinog prekrivača) a treći nekoj drugoj. U žargonu klasifikacije je uobičajeno da se ta tri opsega nazivaju „svojstva“, a taj pojam se koristi umesto pojma „opsezi“ zato što je uobičajeno da se primenjuju razne transformacije na sliku, koje prethode klasifikaciji. One se nazivaju „transformacije svojstava“ a njihov rezultat

„izvedena svojstva“ (primeri su: glavne komponente, HIS transformacije, itd.) (Levin, 1999).

U jednom pikselu, vrednosti (tri) svojstva se mogu smatrati komponentama trodimenzionalnog vektora, vektora svojstava. Takav vektor se može prikazati u trodimenzionalnom prostoru, koji se naziva prostor svojstava. Pikseli koji pripadaju istoj klasi (zemljinog prekrivača) i koji imaju slične karakteristike, se nalaze blizu jedan drugome u prostoru svojstava, bez obzira koliko su udaljeni jedan od drugoga na zemljištu i na slici. Svi pikseli koji pripadaju istoj klasi će, nadamo se, formirati klaster u prostoru svojstava. Šta više, nadamo se da će ostali pikseli, koji pripadaju drugim klasama, se naći izvan ovog klastera (ali unutar drugih klastera, pripadajući drugim klsama). Postoji veliki broj metoda klasifikacije (nadgledanih i nenadgledanih), a da bi klasifikator funkcionisao sa tematskim (umesto spektralnim) klasama, određeno „znanje“ o vezi između klasa i vektora svojstava mora postojati (Levin, 1999).

Teoretski, to se može postići iz baze podataka u kojoj su povezanosti između (tematskih) klasa i vektorsa svojstava sačuvane. Može se npr. pretpostaviti da je u prošlosti dovoljno slika od svake vrste senzora analizirano, kako bi bile poznate spektralne karakteristike svake relevantne klase. Npr. to može značiti da piksel sa vektorom svojstava (44, 32, 81) kod multispektralne SPOT slike uvek znači da je u pitanju trava, dok (12, 56, 49) je uvek piksel šume. Nažalost, opaženi vektori svojstava na određenoj slici su afektirani velikim brojem drugih faktora osim vrstom zemljišnog prekrivača, kao što su: atmosferski uslovi, ugao sunca, tipom zemljišta, vlažnošću, fazom rasta kod vegetacije, vetrom (koji utiče na orijentaciju lišća), itd. Problemi sa kojima se srećemo pokušavajući da uzmemo u obzir sve ove uticaje variraju od veoma lakih do praktično nemogućih za rešiti, ili je barem neophodna velika količina dodatnih podataka (Levin, 1999).

5.1.5 Hiperspektralne slike i redukcija dimenzije svojstava

Senzor obično beleži više vrednosti nego što je potrebno za reprezentaciju relevantnih informacija. Npr. sa hiperspektralnim merenjima-snimcima za detekciju stresa u preciznos poljoprivredi, hiperspektralni senzor može beležiti refleksiju na preko sto talasnih dužina, ali relevantne informacije se mogu sumirati u nekoliko svojstava. Redukcija broja svojstava je važan korak kod analize podataka s obzirom da veliki broj nerelevantnih svojstava može značajno umanjiti preciznost predviđanja, što se naziva Hughesovim fenomenom (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

Uobičajeni pristup kod evaluacije hiperspektralnih merenja je računanje *vegetativnih indeksa* (eng. *Vegetation Indices*, VI). Broj svojstava se takođe može redukovati računanjem unapređenog skupa svojstava iz inicijalnog skupa, ko što se radi pomoću *analize glavnih komponenti* (eng. *Principal Component Analysis*, PCA). PCA transformiše podatke u novi ortogonalni prostor svojstava, a bazira na tome da se kod većine skupova podataka veliki deo različitosti u podacima može izraziti pomoću sveka

nekoliko prvih glavnih komponenti, što rezultira konstrukcijom novog skupa svojstava sa približno istim informacijama. Glavna mana PCA je njen fokus na očuvanju varijanse bez uzimanja u obzir labela (imenovanih klasa) i prema tome PCA može ukloniti obrasce koji mogu biti od značaja za određeni zadatak (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

Više specifične metode selekcije svojstava koriste labele klasa za procenu specifičnu relevantnost svojstva ili grupe svojstava za određeni zadatak. Na primer, linearna diskriminaciona analiza (LDA) određuje linearnu kombinaciju svojstava u odnosu na labele klasa. Poslednjih decenija javlja se i niz drugih metoda selekcije svojstava, koje se mogu kategorizovati ili kao metode *filtera*, *omotača* ili *ugrađene metode* (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

5.2 Pristupi i metode klasifikacije u daljinskom uzorkovanju

Metodama daljinskog uzorkovanja se na relativno jednostavan način prikuplja velika količina višedimenzionalnih podataka. Ako postavimo kao glavni cilj klasifikaciju tih podataka, postavlja se pitanje kako obezbediti takvu klasifikaciju kao što više automatizovan proces, sa zadovoljavajućom preciznošću i utroškom resursa. Uzmimo za primer daljinsko uzorkovanje u preciznoj poljoprivredi (eng. *Precision Agriculture*), gde efikasna zaštita useva zahteva ranu i preciznu detekciju različitih pojava, kao što je biotički stres, a poslednjih godina aktuelna su brojna istraživanja na polju detekcije korova, bolesti biljaka i štetočina, kao što su insekti ili glodari u usevima, i sl. Analizu visokodimenzionalnih podataka za identifikaciju ovih pojava je zahtevna, naročito što postoji veliki broj faktora koji utiče na signal iz senzora. Dostignuća i proboji na ovom polju poslednjih godina su vezana i za razvoj neinvazivnih, optičkih, senzora visoke rezolucije i za metode analize podataka koje su sposobne da se nose sa rezolucijom, veličinom i kompleksnošću signala iz tih senzora. Drugim rečima, razvijane su brojne metode mašinskog učenja koje olakšavaju interpretaciju gorepomenutih kompleksnih signala (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015; Rumpf T. , 2012; Singh, Ganapathysubramanian, Singh, & Sarkar, 2016).

Naročito razvoj jeftinih bespilotnih letelica (UAV) i laganih senzora za prikupljanje slika su doveli do povećanja interesovanja za njihovu upotrebu kod primene daljinskog uzorkovanja (Hung, Xu, & Sukkarieh, 2014). Dok se velika pažnja pridaje prikupljanju, kalibrisanju, beleženju i spajanju mozaika podataka prikupljenih pomoću malih bespilotnih letelica, interpretacija ovih podataka u semantički smislene informacije može i dalje biti zahtevan proces. Standardni radni proces prikupljanja i klasifikacije podataka i dalje zahteva značajan ručni rad za segmentaciju, selekciju svojstava i projektovanje klasifikatora na osnovu datih pravila (Hung, Xu, & Sukkarieh, 2014).

Osnovna pretpostavku za klasifikaciju slika je mapiranje specifičnog dela prostora svojstava (eng. *feature space*) u odgovarajuću klasu. To može biti izvedeno poređenjem dijagrama (krivih) spektralne refleksije. Osnovni princip klasifikacije slika je da se piksel dodeljuje klasi na osnovu svog vektora svojstava (atributa) tako što se poredi sa predefinisanim klasterom prostora svojstava (Dutta, Stein, & Patel, 2008; Tempfli, Kerle, Huurneman, & Janssen, 2009). Takve metode koriste samo spektralne informacije, što može biti problematično sa većim prostornim rezolucijama, koje postaju sve veće i veće. Na primer, zgrada koja se sastoji od različitih materijala rezultira u pikselima sa veoma različitim spektralnim karakteristikama, pa su pikseli za trening od male pomoći. Slično, polje može sadržati piksele zdrave vegetacije, obolele vegetacije, kao i samog tla. Iz tih razloga, pored takvog (osnovnog) pristupa klasifikaciji (na osnovu piksela), noviji pristupi uključuju i objektno orijentisani pristup klasifikaciji (eng. *Object-oriented Analysis, OOA*) (Dutta, Stein, & Patel, 2008; Tempfli, Kerle, Huurneman, & Janssen, 2009; Khorram, Wiele, Koch, Nelson, & Potts, 2016).

Objektno orijentisana analiza, koja se još naziva *analiza na osnovu segmentacije*, omogućuje da se, umesto pokušaja da se klasifikuje svaki piksel posebno i samo na osnovu spektralnih informacija, slika razloži na spektralno homogene segmente koji odgovaraju poljima, stablima, zgradama, itd. Slično kognitivnom pristupu vizuelne interpretacije slika, gde posmatramo svaki element iz ugla njegovog spektralnog prikaza ali takođe iz ugla njegovog oblika i teksture, kao i okruženja, kod OOA možemo precizirati kontektustalne relacije i složenije karakteristike segmenata kako bismo klasifikovali objekte dobijene u procesu segmentacije. Na primer, možemo koristiti teksturu objekta da razlikujemo dve slične vrste šume, ili da razlikujemo bazen od jezera razmatranjem njegovog oblika i možda betona koji ga okružuje, umesto zemlje i vegetacije (Tempfli, Kerle, Huurneman, & Janssen, 2009; Khorram, Wiele, Koch, Nelson, & Potts, 2016). Iako je klasifikacija digitalnih slika doprinela većoj objektivnosti procesu klasifikacije, još uvek postoje brojna ograničenja da bi se dobio adekvatan nivo tačnosti. Zato je poboljšanje tačnosti klasifikacije ključna komponenta savremenih istraživanja na polju (Oommen, i drugi, 2008).

Za klasifikaciju slika u daljinskom uzorkovanju su u upotrebi tehnike klasifikacije i sa i bez nadzora (nadgledani i nenadgledani algoritmi klasifikacije), i to su dobro prihvaćeni standardi kod daljinskog uzorkovanja (Deekshatulu, i drugi, 1995; Khorram, Wiele, Koch, Nelson, & Potts, 2016; Oommen, i drugi, 2008; Rumpf, i drugi, 2010; Tempfli, Kerle, Huurneman, & Janssen, 2009). Uopšteno govoreći, statističko modelovanje se može podeliti na *generativne* i *diskriminativne* modele, pa time i nenadgledano i nadgledano učenje može biti implementirano pomoću takvih reprezentacija modela (Bishop, 2006; Rumpf T. , 2012; Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015). *Generativni modeli* (eng. *Generative models*) su u celosti probabilistički svih promenljivih, dok diskriminativni modeli (eng. *Discriminative models*) samo daju model za ciljne promenljive u zavisnosti od opservacija. I jedni i drugi modeli određuju maksimalnu posteriornu verovatnoću $p(y|x)$ za date opservacije x,

kako bi se dodelila jedna od klasa *y* za svako novo *x* (Bishop, 2006). *Generativni modeli* nastaju iz pretpostavke o raspodeli verovatnoća podataka (P_X), koja omogućuje generisanje novih, simuliranih podataka iz modela. U slučaju generativne klasifikacije, opservacija se obično klasifikuje pomoću mere sličnosti koja uključuje raspodelu klasa. *Diskriminativni modeli* direktno opisuju zavisnost između podataka i klasa. Oni se direktno fokusiraju na posteriore klasa $p(y|x)$ bez eksplicitnog modelovanja marginalnih $p(x)$ (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015; Rumpf T. , 2012). Kada su informacije o podacima u osnovi ograničene, što je obično slučaj kod daljinskog uzorkovanja u preciznoj poljoprivredi, pojednostavljenje koje donose diskriminativni modeli je korisno. Zaista, u tom slučaju nam nije potreban tačan posterior dok god možemo koristiti trening podatke da nađemo diskriminacionu funkciju koja mapira svako *x* direktno na imenovanu klasu. Time su faze izvođenja i zaključivanja kombinovane u jedinstven problem učenja (Rumpf T. , 2012).

Među metodama nadgledane klasifikacije, veliku popularnost je stekla metoda *klasifikacije maksimalne izglednosti* (eng. *Maximum Likelihood Classification*, MLC). Takođe, kako bi se uzele u obzir spektralno-prostorne informacije, i za veću preciznost, *relaksirajuće obeležavanje* (eng. *Relaxation Labelling*) i teksturni pristupi sve više privlače istraživače daljinskog uzorkovanja. Tehnike relaksacije su uspešno primenjene u problemima analize slika poput uklanjanja šuma, detekcije i naglašavanja ivica, prepoznavanje oblika, klasifikacije/segmentacije, itd. (Deekshatulu, i drugi, 1995). *K-means* je jedna od najšire korišćenih metoda za nenadgledano učenje a *Rk-means* se dodatno uvodi zbog analize uticaja izbora centroida. *K-nn* (*k*-najbližih suseda) se koristi jer se predviđanje klasifikacije zasniva na relaciji udaljenosti (Pérez-Ortiz, i drugi, 2015; Rumpf, i drugi, 2010).

Za rešavanje problema klasifikacije tradicionalna statistika procenjuje posteriorne verovatnoće klasa preko raspodela verovatnoća (Rumpf T. , 2012). Takve tradicionalne metode analize podataka, poput linearne regresije i linearne diskriminacione analize, su zasnovane na predefinisanim raspodelama i pretpostavkama modela. One su primenljive bez gubitka preciznosti za podatke koji su usaglašeni sa tim uslovima, npr. Fišerov linearni diskriminant zahteva da se klase mogu razdvojiti linearno i da se svaka od dve klase može predstaviti unimodalnom Gausovom distribucijom (Bishop, 2006; Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015). Time oblast primene može biti ograničena za tu vrstu metoda analize.

Metode mašinskog učenja sa robustnim algoritmima za trening igraju ključnu ulogu kod rešavanja nelinearnih problema klasifikacije i one pružaju algoritme za optimizaciju preciznosti predviđanja pomoću diskriminacionih funkcija (Rumpf T. , 2012). Jedna od najpopularnijih metoda koja se koristi za veliki broj primena kod daljinskog uzorkovanja je *metoda podržavajućih vektora* (SVM), koja je generalno jedna od najuspešnijih metoda mašinskog učenja (Pérez-Ortiz, i drugi, 2015; Rumpf, i drugi, 2010). Ostale metode mašinskog učenja korišćene za daljinsku detekciju u preciznoj poljoprivredi uključuju: (a) za klasifikaciju, pored SVM i *neuronske mreže*

(NN) (nadgledano učenje), i (b) za klasterovanje, *K-means* i samoorganizujuće mape (nenadgledano učenje). Ovakvi napredniji pristupi mašinskog učenja zahtevaju manje priornih informacija i primenljive su na širi spektar zadataka, s obzirom da one izvode distribucije i pretpostavke modela implicitno kroz učenje. One su u stanju da računaju i linearne i nelinearne modele, zahtevaju manje statističkih pretpostavki i prilagođavaju svoju fleksibilnost na veliki broj karakteristika podataka. Uspešne primene uključuju ranu detekciju bolesti biljaka na osnovu spektralnih svojstava i detekciju korova na osnovu oblika, korišćenjem metoda nadgledanog ili nenadgledanog učenja (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

Osnovna pretpostavka kod mašinskog učenja je da su računari u stanju da detektuju i kvantifikuju obrasce u podacima korišćenjem specifičnih algoritama (Witten & Frank, 2005). To omogućuje mašini da automatski koristi nove informacije – što znači da se program unapređuje automatskim učenjem, a to dalje omogućuje detekciju obrazaca i ekstrakciju informacija iz sirovih podataka čak i kada je model u osnovi nepoznat (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015). Može se reći da su takvi pristupi klasifikaciji razvijeni istraživanjima u oblastima veštačke inteligencije i prepoznavanja obrazaca. Neki pristupi se primenjuju već neko vreme, pa se, na primer, veštačke neuronske mreže primenjuju za klasifikaciju slika daljinskog uzorkovanja još od ranih 1990-ih (Khorram, Wiele, Koch, Nelson, & Potts, 2016; Rumpf, i drugi, 2010; Rumpf T. , 2012). Mašinsko učenje se, kao pod-disciplina veštačke inteligencije, bavi automatskim učenjem redovnih obrazaca u podacima. U slučaju zadataka dihotomne klasifikacije trening podaci se sastoje od opservacija $x \in R^2$ i imenovanih klasa $y \in \{+1, -1\}$. Cilj je da na osnovu trening podataka izvučemo zaključke kako klasifikovati neimenovane opservacije. Rezultat takve generalizacije, koju analitičar mora izvesti, je diskriminaciona funkcija koja se može koristiti za interpretaciju novih podataka (Rumpf T. , 2012).

Dakle, postoje dva osnovna zadatka metoda mašinskog učenja (Witten & Frank, 2005). Nenadgledano učenje počinje od neimenovanih (*unlabeled*) podataka i pokušava da pronađe obrasce koji daju novu, kompaktniju i sveobuhvatniju, reprezentaciju sadržanih informacija (Duda, Hart, & Stork, 1995). Najistaknutiji primer nenadgledanog učenja je grupisanje (klasterovanje) (Bishop, 2006; Rumpf, i drugi, 2010). Ukoliko su date ciljne promenljive, imenovane klase y_i, primenljive su metode nadgledanog učenja (Bishop, 2006). Kod njih se traži identifikacija obrazaca u podacima vezanih za imenovane klase (što dozvoljava predviđanje za nove, nepoznate podatke). Ukoliko su date diskretne imenovane klase (npr. „da“ ili „ne“), onda je prikladan metod klasifikacije, dok za kontinualne imenovane klase prkladan je regresioni metod (Bishop, 2006; Witten & Frank, 2005). Zadatak nadgledanog učenja se sastoji u dve faze: prvo se model (F) uči pomoću specifičnih algoritama mašinskog učenja na osnovu imenovanih trening podataka, a zatim se on primenjuje na nove, neimenovane podatke čime se predviđa najbolja klasa za y za testne uzorke X pomoću $y_i = F(x_i)$. Generalno, kvalitet modela je definisan sličnošću stvarnih i predviđenih klasa testnih podataka. Kod

daljinskog uzorkovanja, npr. u oblasti precizne poljoprivrede, oznake klasa (y_i) mogu biti zdravstveno stanje biljke (npr. bolesna ili zdrava) a uzorci podataka (x_i) može biti hiperspektralna refleksija površine lista osmotrena kao jedan piksel hiperspektralne slike (Rumpf, i drugi, 2010; Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

Zbog kompletnosti pregleda koji daje ovo poglavlje, spomenućemo još neke pristupe. Nadgledane i nenadgledane tehnike klasifikacije su klasifikatori „tvrde logike“ (eng. *hard-logic classifiers*) koji raspoređuju elemente slike (piksele) u međusobno isključive kategorije. Za mnoge slučajeve klasifikacije slika to je pristup prihvatljive preciznosti (Khorram, Wiele, Koch, Nelson, & Potts, 2016). Međutim, realni uslovi ne uvažavaju uvek jasne granice između klasa od interesa. Pošto se klasifikacija vrši na osnovu podataka spektralnog odziva, i uprkos naporima analitičara da se klase definišu jednoznačno, slučajevi preklapanja su ponekad neizbežni. To je naročito tačno kod nižih i srednjih prostornih rezolucija multispektralnih slika, gde se vrednosti piksela slike mogu generisati pomoću više od jednog fenomena sa tla, što rezultuje „mešanim pikselima“. Jedan od načina da se zaobiđu konačne granice između klasa je pomoću klasifikatora „mekane logike“ (*fuzzy logic*) (Fisher & Pathirana, 1990; Khorram, Wiele, Koch, Nelson, & Potts, 2016). U vezi sa temom su takođe metaheuristički pristupi, kao što su klasifikatori *simuliranim kaljenjem* (eng. *Simulated Annealing*, SA), koji nude alternativu za opšteprihvaćnu *K-means* metodu klasterovanja (Khorram, Wiele, Koch, Nelson, & Potts, 2016), a paralelno sa pisanjem ove disertacije autor se bavi temom potvrde primene *pretrage promenljivih okolina* (eng. *Variable Neighborhood Search*, VNS) za klasterovanje kod daljinskog uzorkovanja. Takođe se kod daljinskog uzorkovanja koriste i metode *stabla odlučivanja* (eng. *Decision Tree*) (Rumpf, i drugi, 2010).

Izbor optimalne metode analize podataka veoma zavisi od specifičnog problema. Prema tome, nije moguće dati opšte preporuke, ali neki kriterijumi mogu pomoći u identifikaciji primenljivih algoritama. Postoje neke osnovne osobine koje su od značaja za izbor metoda: broj svojstava (dimenzija vektora podataka), veličina uzoraka za trening, vrsta imenovanja klasa i dostupne informacije o raspodelama podataka. Pretpostavljajući da je dostupan podatak o imenovanim klasama, trebalo bi koristiti nadgledane metode. Bez imenovanih klasa, samo nenadgledane metode se mogu koristiti. Diskretne vrednosti oznaka klasa zahtevaju klasifikaciju, kontinualne regresiju (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

U nastavku disertacije, za nas će od najvećeg značaja za istraživanje biti problemi koji se rešavaju nadgledanim metodama, kao problemi klasifikacije – biće poznate imenovane klase koje uzimaju diskretne vrednosti.

5.2.1 Proces klasifikacije u daljinskom uzorkovanju

Prema (Gong & Howarth, 1990), opšta procedura klasifikacije slika podrazumeva:

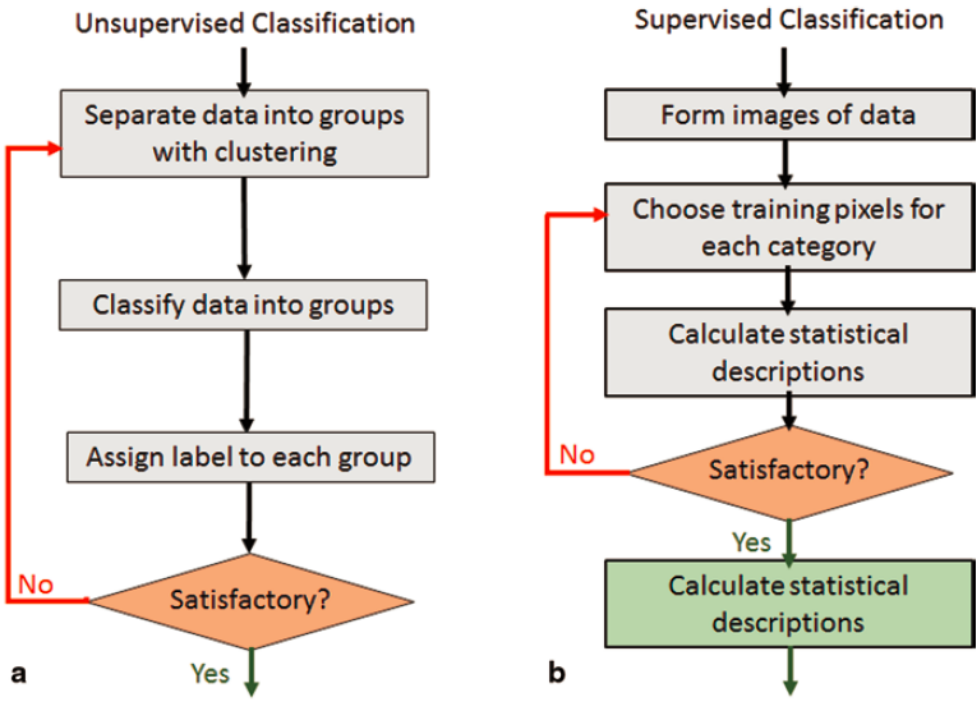
1. Projektovanje šeme klasifikacije slike: obično su klase informacija poput urbane, poljoprivredne, šumske oblasti, itd. Sprovođenje terenskih studija i utvrđivanje informacija sa terena (ground truth) i drugih pomoćnih podataka posmatrane oblasti.
2. Preprocesiranje slike, što uključuje radiometrijske, atmosferske, geometrijske i topografske korekcije, poboljšanja slike, i inicijalno klasterovanje slike.
3. Odabir reprezentativnih oblasti na slici i analiza inicijalnih rezultata klasterovanja ili generisanje trening podataka.
4. Klasifikacija slike:
    a. Nadgledane tehnike: korišćenje trening skupa podataka:
        i. Slika
        ii. Nadgledani trening
        iii. Označavanje piksela
        iv. Procena tačnosti
    b. Nenadgledane tehnike: klasterovanje slike i grupisanje klastera:
        i. Slika
        ii. Klasterovanje i klaster analiza
        iii. Klasterovanje i grupisanje klastera
        iv. Procena tačnosti
5. Post-procesiranje: potpuna geometrijska korekcija i filtriranje i označavanje (dekoracija) klasifikacije.
6. Procena tačnosti: poređenje rezultata klasifikacije sa terenskim istraživanjima.

Prema (Tempfli, Kerle, Huurneman, & Janssen, 2009) proces klasifikacije slika obično uključuje pet koraka:

1. Selekcija i priprema slika daljinskog uzorkovanja. Zavisno od vrste zemljišnog prekrivača ili šta već je potrebno klasifikovati, iznor najprikladnijeg senzora, najprikladnijeg datuma prikupljanja i najprikladnijeg opsega talasnih dužina.
2. Definicija klastera (klasa) u prostoru svojstava. Ovde su moguća dva pristupa: nadgledana i nenadgledana klasifikacija.

3. Izbor algoritma klasifikacije. Nakon što su spektralne klase definisane i prostoru svojstava, operator mora odlučiti kako će pikseli (na osnovu svojih vektora svojstava) biti dodeljeni tim klasama.
4. Pokretanje same klasifikacije. Nakon što su ustanovljeni trening podaci i odabran algoritam klasifikacije, sprovodi se sama klasifikacija. To znači da, na osnovu svojih DN-ova (digitalnih brojeva), svaki „višekanalni piksel“ (ćelija) na slici se dodeljuje jednoj od predefinisanih klasa.
5. Validacija rezultata. Nakon što se dobije klasifikovana slika, potrebno je proceniti njen kvalitet poređenjem sa referentnim podacima (*ground truth*). Ovo zahteva izbor tehnike uzorkovanja, generisanje matrice grešaka (eng. *error matrix*) i izračunavanje parametara greške.

Sam proces klasifikacije, nenadgledane i nadgledane, što je opisano pod tačkom 4 u obe gornje liste, može se šematski predstaviti kao na Slici 5.3.1 (Khorram, Wiele, Koch, Nelson, & Potts, 2016):

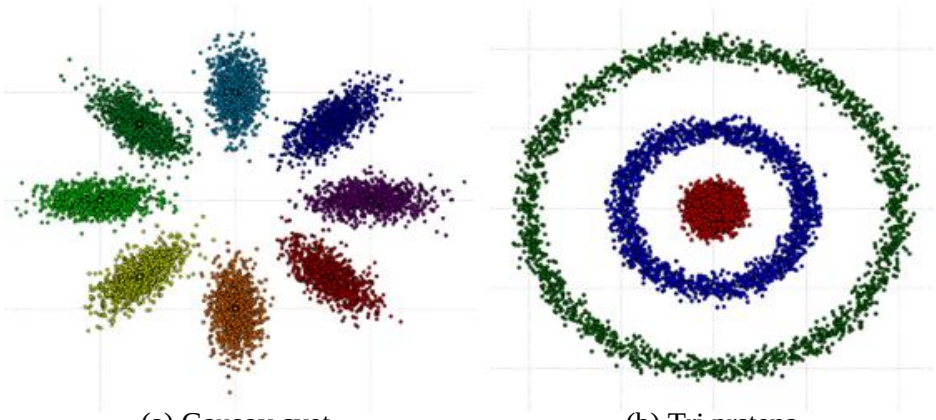


Slika 5.3.1: opšti dijagram toka (a) nenadgledanog i (b) nagledanog metoda klasifikacije.

5.2.2 Klasifikacija bez nadzora (grupisanje)

Klasifikacija s nadzorom zahteva dovoljno znanja o posmatranim oblastima i ukoliko ovo znanje nije dovoljno ili klase od interesa nisu još definisane, može se primeniti *klasifikacija bez nadzora* (*nenadgledana klasifikacija*). Kod ove klasifikacije, algoritam klasterovanja (grupisanja) se koriste da particionišu prostor svojstava u određeni broj klastera – spektralnih klasa. Drugim rečima osnovna svrha ovih metoda je da definišu spektralna grupisanja na osnovu određenih spektralnih sličnosti (Tempfli, Kerle, Huurneman, & Janssen, 2009).

U opštem problemu grupisanja, za dati broj klastera K i n objekata, x_1, ..., x_n, želimo da particionišemo n objekata u K klastera. Opšte pravilo za klaserovanje je da želimo da slični objekti pripadaju istom klasteru. Na primer, ako posmatramo „Gausov cvet“ (slika 5.2.1.(a)), očekujemo da klasterovanje tačaka za K=8 grupiše zajedno tačke jednog „oblačića“. Slično, kod tri pristena (slika 5.2.1.(b)), bismo dodelili sve tačke jednog pristena jednom klasteru (Signal and Information Processing Laboratory, 2013).



(a) Gausov cvet (b) Tri prstena

Slika 5.2.1: Primeri grupisanih (klasterovanih) tačaka.

U klasterovanju kod daljinskog uzorkovanja, prilikom tog procesa se pretpostavlja da ne postoji znanje o „tematici“ zemljišnog prekrivača i njegovim klasama (npr. gradovi, putevi, usevi, itd.). Sve što ove metode mogu uraditi jeste da ustanove da naizgled ima npr. 16 različitih „stvari“ na slici i da im dodeli brojeve, od 1 do 16. Svaka od tih identifikovanih „stvari“ se naziva *spektralna klasa*. Rezultat može biti rasterska mapa, na kojoj svaki od piksela ima dodeljenu klasu (od 1 do 16), shodno grupi (klasteru) kome vektor svojstava datog piksela sa slike pripada. Nakon što se proces klasifikacije završi, na korisniku je da pronađe relaciju između spektralnih i tematskih klasa (Levin, 1999).

Spektralne klase koje se dobiju nenadgledanom klasifikacijom baziraju isključivo na „prirodnom“ grupisanju vrednosti na slici – slike su klasifikovane automatski u

različite spektralne klase na osnovu razlikovanja spektralnih potpisa svakog piksela. Slabosti ovih metoda su u tome da više klasa od velikog značaja za korisnika može imati veoma male spektralne razlike, čime klase identifikovane automatski često nisu od velikog značaja za korisnika (Dutta, Stein, & Patel, 2008; Oommen, i drugi, 2008). Veoma je moguće ustanoviti da se jedna tematska klasa deli na više spektralnih, ili, u gorem slučaju, da više tematskih klasa završi u istom klasteru. Prirodne površine su retko kad sastavljene od samo jednog, uniformnog materijala, a mogu se javiti i slučajevi spektralnog mešanja kada se materijali sa različitim spektralnim svojstvima predstavljaju sa jednim (istim) pikselom (Levin, 1999).

Postoje brojne metode klaster analize, odnosno algoritmi nenadgledane klasifikacije (klasterovanja). Na primer, neki od njih su *DB-SCAN*, *K-means*, *X-means*, *spektralno klasterovanje*, *SVM-klasterovanje*, itd. (Rumpf, i drugi, 2010; Dobrota, Delibašić, & Delias, 2016). Obično algoritmi nisu u potpunosti automatski, već korisnik mora precizirati neke parametre poput broja klastera koji želi dobiti, maksimalne veličine klastera (u prostoru svojstava), najmanje rastojanje (takođe u prostoru svojstava) koje je dozvoljeno između različitih klastera, itd. (Levin, 1999; Tempfli, Kerle, Huurneman, & Janssen, 2009).

5.2.2.1 K-means klasterovanje

K-means metod dodeljuje svaku opservaciju u skupu podataka (u našem slučaju piksel daljinski uzorkovane slike) jednom od *k* klastera na osnovu toga koji centroid, ili sredina, klastera je najbliži opservaciji u višedimenzionalnom Euklidskom prostoru (Khorram, Wiele, Koch, Nelson, & Potts, 2016). Takav algoritam klasterovanja je nenadgledani pristup za struktuiranje podataka na osnovu geometrijske raspodele podataka u okviru prostora svojstava. Uzorci x_j se dodeljuju predefinisanom broju klastera *k* koji su u celosti definisani svojim centralnim tačkama μ_i. U iterativnom procesu, dodele i pozicije centralnih tačaka klastera se optimizuju iterativnim minimiziranjem funkcije

$$j = \sum_{i=1}^{k} \sum_{j \in S_i} \|x_j - \mu_i\|^2$$

gde S_i označava skup svih tačaka podataka dodeljenih klasteru *i*. Drugim rečima, prvo se inicijalizuju centroidi, zatim se klasifikuju uzorci prema najbližem centroidu μ_i, pa se ponovo računaju μ_i i proces se ponavlja sve dok nema promena centroida μ_i. Kompleksnost računanja je $O(ndkT)$, gde je *n* broj uzoraka, *d* broj svojstava a *T* broj iteracija. U praksi je broj iteracija generalno mnogo manji od broja uzoraka (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015; Duda, Hart, & Stork, 1995).

Ova nekonveksna optimizacija rezultuje klasterovanjem koje najbolje odgovara ukoliko raspodela podataka podržava sferične klastere sa sličnim radijusima. Neki od nedostataka ove metode su nedostatak fizičkog značenja povezanog sa svakim od

klastera i zavisnost rezultata od inicijalizacije. Za prevazilaženje nekih od ograničenja, razvijena su neka poboljšanja kao što su *fuzzy K-means* (verovatnoće klastera) i *K-means++* (poboljšana incijalizacija) (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

5.2.2.2 Spektralno klasterovanje

Spektralno klasterovanje (eng. *Spectral Clustering*) postaje jedan od najpopularnijih savremenih algoritama klasterovanja, najviše počev (Shi & Malik, 2000). Jednostavan je za implementaciju, može se efikasno rešiti standardnim softverom linearne algebre i veoma često nadmašuje rezultate dobijene tradicionalnim algoritmima klasterovanja, poput *K*-means (Luxburg, 2007; Dobrota, Delibašić, & Delias, 2016).

Algoritmi spektralnog klasterovanja imaju čvrstu vezu sa teorijom grafova. Kompletan, težinski neusmeren graf se kreira za dati skup podataka, gde čvorovi pretstavljaju podatke a ivice definišu relaciju susednosti među podacima (Filippone, Camastra, Masulli, & Rovetta, 2008). To znači da ako ne raspolažemo sa više informacija nego što je sličnost između tačaka podataka, pogodan način pretstavljanja podataka je u obliku grafa sličnosti $G = (V, E)$. Postoji nekoliko popularnih načina za transformisanje datog skupa podataka $x_1, \dots, x_n$ sa sličnošću parova s_{ij} ili udaljenosti parova d_{ij} u graf: graf ε-suseda, graf k-najbližih suseda ili potpuno povezani graf. Međutim, konstrukcija grafa sličnosti za spektralno klasterovanje nije trivijalni zadatak. Potrebno je uzeti u obzir samu funkciju sličnosti, koju vrstu grafa sličnosti koristiti, parametre grafa sličnosti, računanje svojstvenih vektora, broj klastera, *K*-means korak, koju Laplasovu transformaciju grafa koristiti (Luxburg, 2007).

U okviru definisanog, klasterovanje je problem sečenja grafa kako bi se odvojili skupovi čvorova. Metod spektralnog klasterovanja relaksira kompleksnost optimizacionog problema sečenja grafa pomoću spektralne dekompozicije Laplasove matrice datog skupa podataka. Ona daje korisne informacije o osobinama grafa, pri čemu se određeni svojstveni vektori Laplasove matrice vezuju za sečenje grafa i odgovarajući svojstveni vektor može grupisati slične podatke (Filippone, Camastra, Masulli, & Rovetta, 2008). Prema tome, za brzi pregled spektralnog klasterovanja možemo koristitćemo pogled iz ugla sečenja grafa: ako je dat graf sličnosti sa matricom povezanosti $W = (w_{ij})$, najjednostavniji i najdirektniji način za particionisanje takvog grafa je rešavanje problema *minimalnog sečenja* (*mincut*) (Luxburg, 2007). Za njegovo definisanje, ovedimo notaciju $W(A, B) = \sum_{i \in A, j \in B} w_{ij}$ i $\bar{A}$ za komplement od A. Za dati broj k podskupova, mincut pristup se jednostavno sastoji od biranja takvog particionisanja grafa $A_1, \dots, A_k$ koje minimizira:

$$cut(A_1, \dots, A_k) \coloneqq \frac{1}{2} \sum_{i=1}^{k} W(A_i, \bar{A}_i)$$

Pošto bi klasteri trebali da su razumno velike grupe podataka, moramo se pobrinuti da su skupovi $A_1, \dots, A_k$ razumno veliki. Dve najčešće funkcije cilja koje sadrže to su $RatioCut$ (Hagen & Kahng, 1992) i normalizovano sečenje $Ncut$ (Shi & Malik, 2000). Kod *RatioCut*-a, veličina podskupa A grafa se meri pomoću broja njegovih ivica $|A|$, dok kog *Ncut*-a se veličina meri težinama ivica $vol(A) = \sum_{i \in A} d_i$, gde je $d_i = \sum_{j=1}^{n} w_{ij}$. Funkcije cilja tada postaju:

$$RatioCut(A_1, \dots, A_k) \coloneqq \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \bar{A}_i)}{|A_i|} = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{|A_i|}$$

$$Ncut(A_1, \dots, A_k) \coloneqq \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \bar{A}_i)}{vol(A_i)} = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}$$

Nažalost, kompleksnost izračunavanja optimuma ovih funkcija cilja je veoma velika je uvođenje uslova za balans veličine klastera čini da prethodno jednostavan problem za rešavanje, *mincut*, postane NP-težak (Luxburg, 2007). Iz tog razloga se koristi njegova relaksacija pomoću spektralnih koncepata analize grafova (Filippone, Camastra, Masulli, & Rovetta, 2008). Spektralno klasterovanje je način rešavanja relaksiranih verzija *RatioCut* i *Ncut* problema, pri čemu relaksiranje *Ncut* problema vodi normalizovanom spektralnom klasterovanju a relaksiranje *RatioCut*-a vodi nenormalizovanom spektralnom klasterovanju. Ova relaksacija se može formulisati uvođenjem Laplasove matrice (Luxburg, 2007).

5.2.3 Klasifikacija s nadzorom

Kod *nadgledane klasifikacije* analitičar koji analizira sliku nagleda proces kategorizacije piksela precizirajući, računarskom algoritmu, numerički opis različitih vrsta npr. zemljišnog prekrivača koje se javljaju na slici (Dutta, Stein, & Patel, 2008; Oommen, i drugi, 2008; Rumpf, i drugi, 2010; Levin, 1999):

- Prvo se u *fazi treninga* (*učenja, uzorkovanja*) definišu uzorci za trening koji opisuju za ograničen broj piksela tipičan spektralni obrazac zemljišnog prekrivača, tj. definiše se ključ interpretacije koji statistički opisuje spektralni potpis za svaku klasu od interesa za korisnika. Tokom faze treninga se prvo definišu klase koje će se koristiti i o svakoj klasi je potrebna neka „istina sa zemlje“ (eng. *ground truth*): određeni broj mesta na slici za koje se zna da pripadaju toj klasi. Ako je ground truth poznata, trening uzorci (male oblasti pojedinačnih piksela), takođe nazivani *oblasti interesa* (eng. Regions of Interest, ROI), su označeni na slici i odgovarajući naziv klase se zadaje.

- Zatim se u *fazi odlučivanja* (*klasifikacije*) „nepoznati“ pikseli slike numerički porede sa uzorkom za trening i imenuju se klasama zemljišnog prekrivača koje imaju slične karakteristike. Zadatak je algoritma odlučivanja da particioniše prostor svojstava na osnovu trening uzoraka. Za svaki mogući vektor svojstava, program mora odlučiti kom skupu trening piksela je tak vektor svojstava najsličniji. Nakon toga, program kreira mapu gde je svaki piksel dodeljen imenovanoj klasi na osnovu takvog particionisanja (ili, kod nekih algoritama, pikseli se mogu označiti kao „nepoznate“ klase).

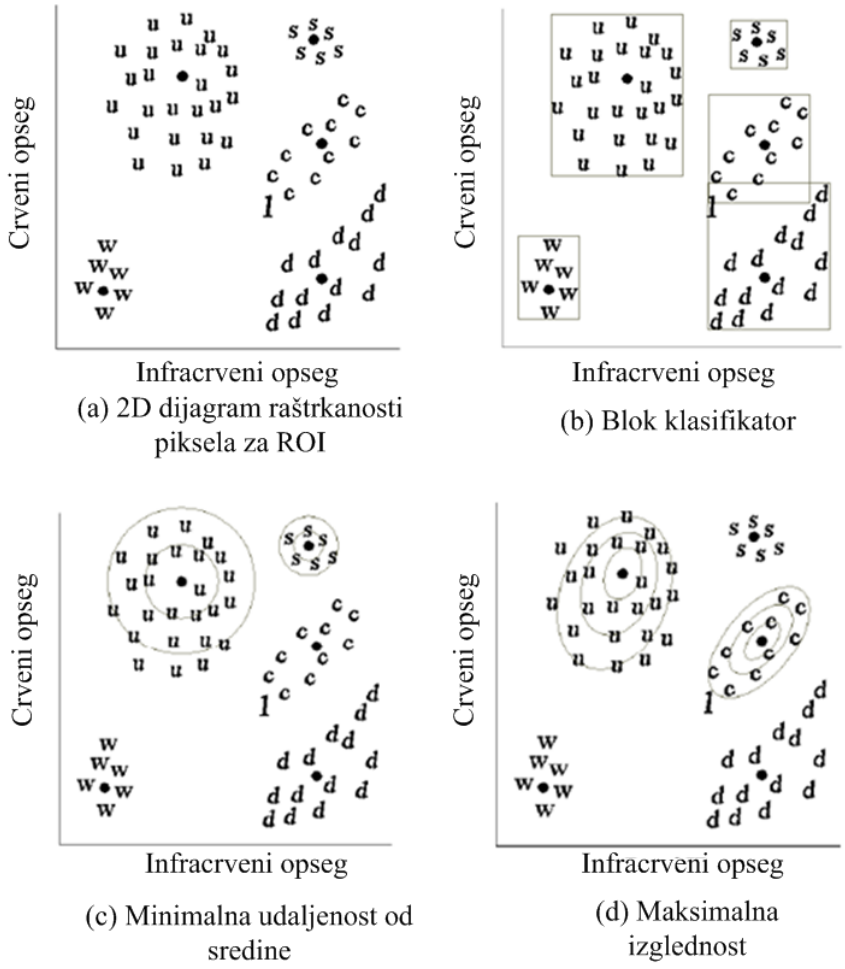
Nasuprot nenadgledanoj klasifikaciji, gde se klase biraju samo na osnovu kontrasta spektralnih karakteristika, ove motede se oslanjaju na znanje i iskustvo korisnika u izboru trening podataka, kako bi se klasifikacija izvršila što efikasnije i tačnije (Oommen, i drugi, 2008).

Nalaženje veze između klasa i vektora svojstava nije trivijalno i različiti algoritmi odlučivanja se koriste. Takvi algoritmi se razlikuju u načinu na koji particionišu prostor svojstava. Neki primeri su (Levin, 1999; Tempfli, Kerle, Huurneman, & Janssen, 2009):

- *Blok klasifikator* (eng. *Box Classifier*): najjednostavniji metod, kod kojeg se pravougaonici (u 2D prostoru) kreiraju oko trening vektora svake klase. Pozicija i veličina blokova se kreira oko vektora svojstava (Min-Max metoda), ili na osnovu središnjeg vektora (koji je centar bloka) i standardne devijacije za svako od svojstava (što određuje veličinu bloka). U slučaju preklapanja blokova, kada vektor spada u preklopljeni deo, obično se prioritet daje manjem bloku, a vektori koji ne upadnu ni u jedan blok se označavaju kao „nepoznati“.

- *Klasifikator minimalnog rastojanja do sredine* (eng. *Minimum Distance-to-mean classifier*): prvo računa za svaku klasu središnji vektor za trening vektore svojstava, a onda se prostor svojstava particioniše tako što se svakom vektoru daje klasa najbližeg središnjeg vektora, na osnovu Euklidske metrike. Rezultat može biti „nepoznat“ ukoliko je rastojanje veće od zadatog praga (maksimuma).

- *Klasifikator Gausove maksimalne izglednosti* (eng. *Gaussian Maximum Likelihood classifiers*): pretpostavlja da su vektori svojstava svake klase (statistički) raspodeljeni prema multivarijacionoj normalnoj funkciji gustine verovatnoća. Trening uzorci se koriste za procenu parametara raspodele. Granice između različitih particija prostora svojstava se nalaze gde se odluka menja iz jedne klase u drugu i nazivaju se *granicama odlučivanja* (eng. *decision boundaries*).



Slika 5.2.2: Primer dijagrama raštrkanosti klasa (C-kukuruz, D-listopadna šuma, S-pesak, U-urbana zona, W-voda, •-sredina oblasti od interesa, 1-piksel za klasifikaciju) i različitih pristupa klasifikaciji.

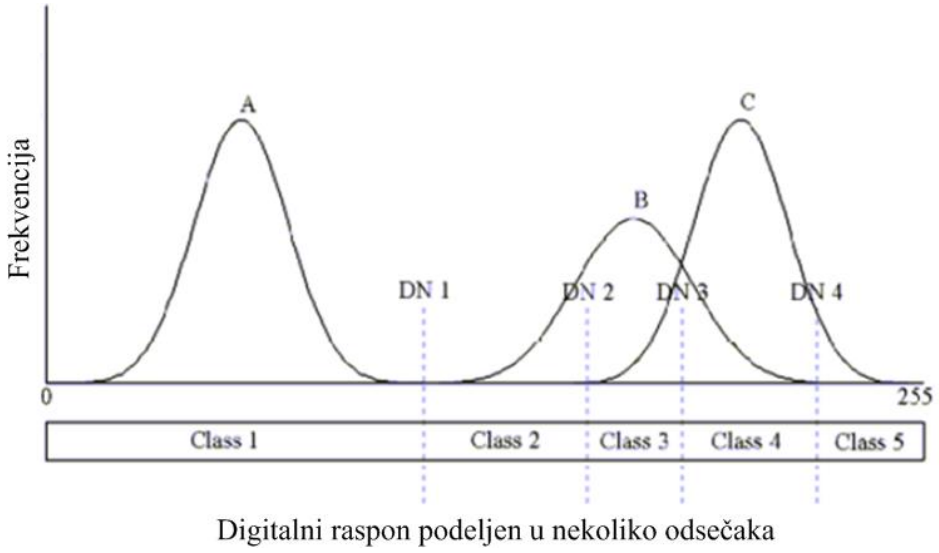
Deterministički pristup klasifikaciji se može opisati pomoću sledećeg problema: neka je slika predstavljena kao $x \equiv (x_1, \dots, x_n)$, gde n piksela predstavljeno kao d-dimenzionalni vektori svojstava. Dalje, neka je $I \equiv \{1, \dots, n\}$ skup celih brojeva koji indeksira n piksela hiperspektralne slike (ili n vektora svojstava) i $L \equiv \{1, \dots, K\}$ skup K imenovanih klasa. Za $i \in I$, neka je slika klasa predstavljena kao $y \equiv (y_1, \dots, y_n)$, gde $y_i \in L$ označava oznaku klase za i-ti piksel. Cilj klasifikacije hiperspektralne slike je zaključiti $y_i \in L$ za svako i, na osnovu vektora svojstava x_i i generisati 2D sliku oznaka klasa koja predstavlja klasnu informaciju originalne d-dimenzionalne hiperspektralne slike (Prabhakar, Gintu, Geetha, & Soman, 2015).

Postoji nekoliko konvecionalnih metoda nadgledane klasifikacije, od kojih je najopštije prihvaćen *Klasifikator maksimalne izglednosti* (MLC), dok su neki ostali konvencionalni algoritmi označavanja piksela kod nadgledane klasifikacije *Multidimenzionalno određivanje granične vrednosti* (eng. *Multidimensional*

thresholding) i *Klasifikacija minimalne udaljenosti* (eng. *Minimum-Distance Classification*). Uprkos svojoj popularnosti, MLC ima i nekoliko ograničenja (Oommen, i drugi, 2008). Relativno novi opšteprihvaćeni metod nadgledane klasifikacije slika je *Klasifikacija podržavajućim vektorima* (SVC), za koju je utvrđeno da postiže veći nivo preciznosti nego savremene konvencionalne metode klasifikacije. SVC je nova generacija metoda nadgledanog učenja koje baziraju na principu teorije statističkog učenja, koja je projektovana da smanjuje neizvesnost u strukturi modela i prigodnosti podataka (Oommen, i drugi, 2008).

5.2.3.1 Sečenje gustine

Spomenimo prvo jedan od najjednostavniji oblik klasifikacije kod daljinskog uzorkovanja. Teoretski, moguće je bazirati klasifikaciju samo na jednom spektralnom opsegu daljinski uzorkovane slike, pomoću *Sečenja gustine* (eng. *Density Slicing*). To je tehnika gde se *digitalni brojevi* (DN, npr. jedna boja ili nijansa sive), distribuirani duž horinzotalne ose histograma slike, dele u skup korisnički određenih intervala ili odsečaka. Broj odsečaka i granice između njih zavise od različitih zemljišnih prekrivača posmatrane oblasti. Svi DN-ovi koji upadaju u određeni interval se klasifikuju u klasu koju taj interval predstavlja. Prvo se određuju rasponi DN-ova za jedan tip zemljišnog prekrivača, a zatim se određuje domenske grupe sa granicama odsečaka i nazivima. Ova tehnika će dati razumne rezultate jedino ako se DN-ovi klasa ne preklapaju značajno (Slika 5.2.3). Na slici, postoji jasna distinkcija između zemljišnog prekrivača A i B/C, međutim kada se posmatraju raspodele za B i C, postoji značajno preklapanje, čime će uvek dolaziti do velike greške klasifikacije za vrednosti koje upadnu u raspon između DN2 i DN4 (Levin, 1999).

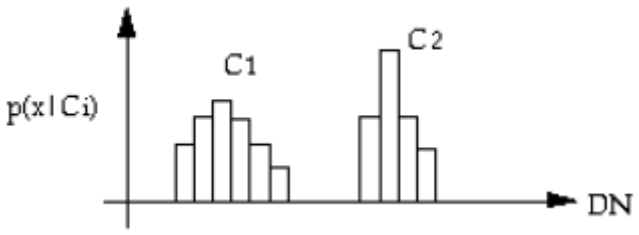


Slika 5.2.3: Klsifikacija sečenjem gustine (frekvencije) piksela

5.2.3.2 Klasifikacija maksimalnom izglednošću (MLC)

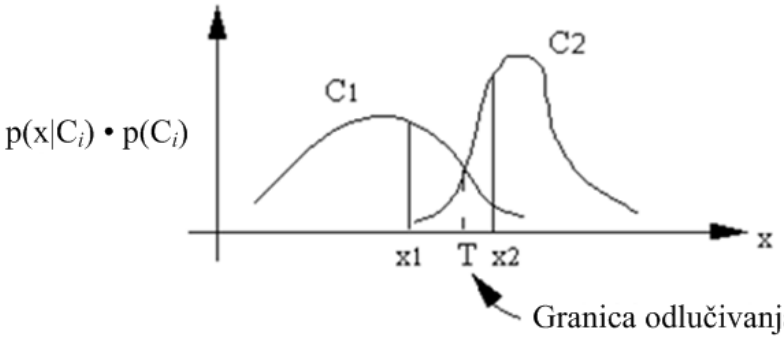
Klasifikacija maksimalnom izglednošću (eng. *Maximum Likelihood Classification*, MLC) je tradicionalno korišćena kao osnov i najpopularnija nadgledana tehnika za klasifikaciju podataka daljinski uzorkovanih slika, naročito kod poljoprivrede, šumarstva i analize korišćenja zemljišta (Evans, 1998; Deekshatulu, i drugi, 1995). Ona je zasnovana na pretpostavci da postoje statistički modeli koji opisuju raspodelu klasa u prostoru atributa (svojstava). Uzimajući te modele u obzir, klasa novog objekta je određena računanjem koji od tih modela najizglednije opisuje taj objekat. Drugim rečima, bira se model sa maksimalnom izglednošću. Klasifikacija maksimalnom izglednošću obično pretpostavlja multivarijantne normalne (Gausove) modele (Evans, 1998).

MLC se zasniva na verovatnoći da svaki piksel slike pripada određenoj klasi. Verovatnoće mogu biti iste za sve klase a varijacija te pretpostavke je, poznata kao *Bajesovo pravilo odlučivanja* (eng. *Bayesan Decision Rule*), gde verovatnoće različitih klasa moraju biti precizirane (Oommen, i drugi, 2008). Drugim rečima, zasniva se na Bajesovom pravilu odlučivanja. Za jednodimenzionalni slučaj, generisanjem trening statistika za dve klase, možemo dobiti njihove raspodele verovatnoća (Slika 5.2.4).



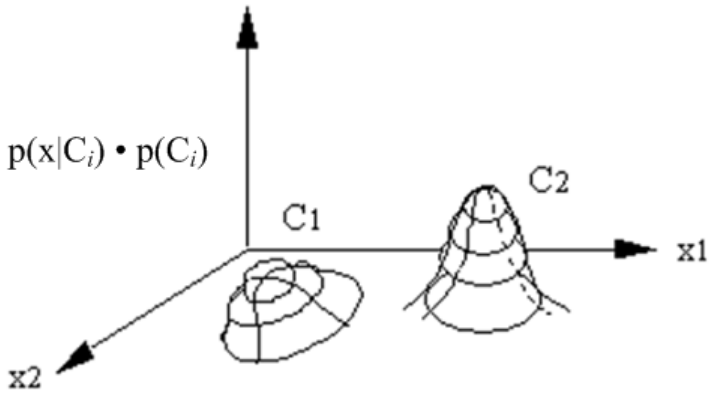
Slika 5.2.4: Statističke raspodele verovatnoća za dve klase, C1 i C2.

Interpretacija klasifikatora maksimalne izglednosti je prikazana na Slici 5.2.5. Na njoj je x klasifikovan u odnosu na maksimum $p(x|C_i) \bullet p(C_i)$: x_1 je klasifikovan u C_1, x_2 u C_2. Granica klasa je određena tačkom jednakih verovatnoća.



Slika 5.2.5: Interpretacija klasifikatora maksimalne izglednosti (MLC).

U dvodimenzionalnom prostoru, granice klasa se ne mogu lako odrediti. Zato se kod maksimalne izglednosti ne koriste granice (kao kod *sečenja gustine*) već se porede verovatnoće (Slika 5.2.6).



Slika 5.2.6: Interpretacija klasifikatora maksimalne izglednosti u dvodimenzionalnom prostoru.

Za klasifikovanje obrasca (piksela) x, računar pri korišćenju pravila odlučivanja maksimalnom izglednosti računa proizvod $p(x|C_i) \bullet p(C_i)$ za svaku klasu i dodeljuje obrazac (piksel) klasi koja ima najveću vrednost, gde je $p(x|C_i)$ funkcija gustine verovatnoće povezana sa pripadanjem izmerenog vektora x klasi C_i (Deekshatulu, i drugi, 1995).

Pravilo odlučivanja maksimalnom izglednošću je: odlučiti $x \in C_i$ ako i samo ako $p(x|C_i) \bullet p(C_i) \geq p(x|C_j) \bullet p(C_j)$ za svako j=1,2,...,m klasa, gde su $p(C_i)$ priorne verovatnoće. Ukoliko su funkcije verovatnoće povezane sa klasama obrazaca multivarijantne normalne funkcije gustine, onda diskriminaciona funkcija postaje:

$$D_i(\mathrm{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left[-(1/2)[(\mathrm{x} - M_i)^T \Sigma_i^{-1}(\mathrm{x} - M_i)]\right]$$

gde su i klasa od interesa, Σ_i matrica kovarijansi (još označavana sa Cov_i) piksela u uzorku klase i iz trening podataka, x vektor svojstva odnosno vektor spektralnog potpisa piksela koji se testira, M_i srednji vektor uzorka za i-tu klasu iz trening podataka, $|\Sigma_i|$ determinanta Σ_i, Σ_i^{-1} inverzna matrica kovarijansi Σ_i i T je funkcija transponovanja (Deekshatulu, i drugi, 1995; Oommen, i drugi, 2008).

Ekstenzija pristupa maksimalne izglednosti je Bajesov klasifikator. Klasifikacija maksimalnom izglednošću se vrši pretpostavljajući jednake verovatnoće za svaku od klasa, dok kod Bajesovog klasifikatora analitičar unosi priornu verovatnoću za klase. Jednačina koja implementira klasifikator maksimalne izglednosti, odnosno Bajesovo pravilo odlučivanja, je

$$D = \ln(a_c) - \left[\frac{1}{2}\ln(|\Sigma_i|)\right] - \left[\frac{1}{2}(\mathrm{x} - M_i)^T(\Sigma_i^{-1})(\mathrm{x} - M_i)\right]$$

gde je D težinska udaljenost (izglednost), a_c procenat verovatnoće da piksel pripada klasi i (podatak o o priornoj verovatnoći – mogu se po potrebi pretpostaviti jednake verovatnoće, odnosno može se pretpostaviti kao 1), i ln je funkcija prirodnog logaritma (Deekshatulu, i drugi, 1995; Oommen, i drugi, 2008). Piksel koji se testira se dodeljuje klasi i koja ima najveću izglednost (D) ili najmanju težinsku udaljenost (Deekshatulu, i drugi, 1995; Oommen, i drugi, 2008).

Koraci implementacije su: kvantitativno evaluirati varijansu i kovarijansu kategorije kada se klasifikuje nepoznati piksel. Pretpostavlja se da su trening podaci Gausovski (normalno) raspoređeni. Ova raspodela je pretstavljena preko srednjeg vektora i matrice kovarijansi. Verikalna osa je verovatnoća piksela da pripada jednoj od klasa. Nakon što se izračunaju verovatnoće za piksel (koji se klasifikuje) da pripada svakoj od klasa, piksel se dodeljuje klasi koja ima najveću vrednost verovatnoće. Ako je vrednost verovatnoće ispod nekog praga (koji određuje analitičar) klasifikuje se kao „nepoznato“ (Deekshatulu, i drugi, 1995).

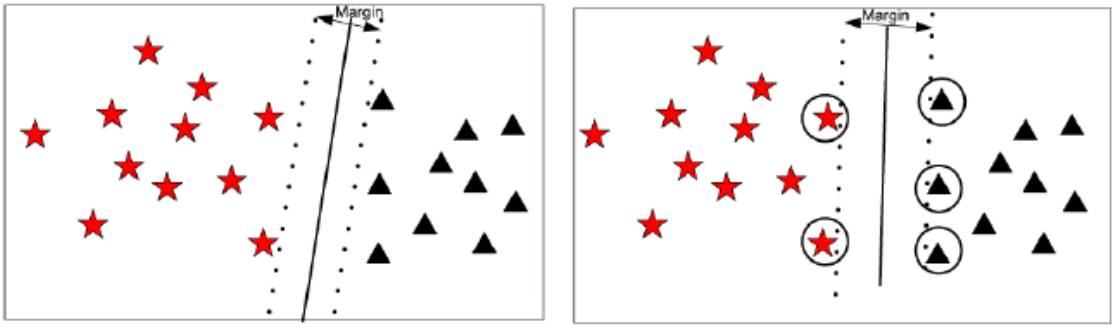
Snaga MLC-a je u tome da uzima u obzir varijabilnost unutar svake klase pomoću matrice kovarijansi za klasifikaciju piksela. Neke od mana su te da se metod oslanja na pretpostavku da su podaci u svakom ulaznom opsegu normalno raspodeljeni i potrebno je da postoji dovoljno uzorkovanih trening piksela kako bi se dobila reprezentativna estimacija srednjeg vektora i matrice varijansi-kovarijansi (Oommen, i drugi, 2008).

Da bi se obezbedila maksimalna separacija između trening klasa može se koristiti Kanonična analiza varijansi, koja pruža metod za transformisanje ulaznih podataka atributa. Kanonična analiza varijansi se može smatrati kao dvofazna rotacija podataka atributa: prva faza se sastoji od analize glavnih komponenti (PCA) podataka atributa, a druga se sastoji od eigen-analize srednjih vrednosti grupa za glavne komponente iz prvog koraka. Na ovaj način se razlike između klasa maksimizuju u odnosu na razlike unutar klasa. To je naročito važno za primene kod daljinskog uzorkovanja gde se trening podaci sastoje od regiona sa mnogo piksela, s obzirom da spektralne vrednosti piksela koje pripadaju istoj trening klasi mogu uključivati širok raspon vrednosti (Evans, 1998). Za unapređenje efikasnosti klasifikacije maksimalnom izglednošću, mogu se koristiti implementacijom *lookup* tabela, redukcija dimenzionalnosti podataka i stabla odlučivanja (Deekshatulu, i drugi, 1995).

5.2.3.3 Klasifikacija podržavajućim vektorima (SVC)

Klasifikacija podržavajućim vektorima (eng. *Support Vector Classification*, SVC) spada u multi-namenske metode koje se nazviaju *Mašine podržavajućih vektora* (SVM). SVM je nova generacija sistema nadgledanog učenja koji bazira na principima teorije statističkog učenja (Oommen, i drugi, 2008). SVM se koriste za nadgledanu klasifikaciju i regresiju pružajući linearne i nelinearne diskriminacione funkcije (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015). Kod klasifikacije SVM funkcioniše tako što nalazi hiperravan u prostoru mogućih inputa. Osnovni pristup je

nalaženje takve hiperravni koja daje optimalnu separaciju između dve klase. Obično se hiperravan razvija korišćenjem podskupa podataka koji se nazivaju trening podaci i sposobnost generalizacije dobijene hiperravni se validira korišćenjem nezavisnog skupa podataka koji se nazivaju testnim podacima. Za klasifikaciju podataka sa N dimenzija (N-1)-dimenzionalna hiperravan se kreira. Uzimajući u obzir binarni slučaju linearno razdvojivih podataka (Slika 5.2.7), može se primetiti da može postojati beskonačan broj hiperravni koje razdvajaju podatke. Međutim, postoji samo jedna hiperravan sa maksimalnom marginom. Ovo je optimalna hiperravan i vektori (tačke) koje ograničavaju širinu margine su podržavajući vektori (Oommen, i drugi, 2008).



Slika 5.2.7: Binarni slučaj linearno razdvojivih podataka: A (levo) – s obzirom da margina nije maksimalna, hiperravan nije optimalna; B (desno) – optimalna hiperravan sa maksimalnom marginom. Podržavajući vektori su zaokruženi.

Sa slike 5.2.7.B se može primetiti da podržavajući vektori leže na dve hiperravni koje su paralelne optimalnoj hiperravni. One su definisane funkcijom $\omega \mathrm{x}_i + b = \pm 1$, gde je x tačka na hiperravni, ω je normalno na hiperravan, i b je otklon. Prema tome, margina između ovih ravni je $2/\|\omega\|$. Cilj maksimiziranja margine je problem optimizacije dat sa

$$\min\left[\frac{1}{2}\|\omega\|^2\right],$$

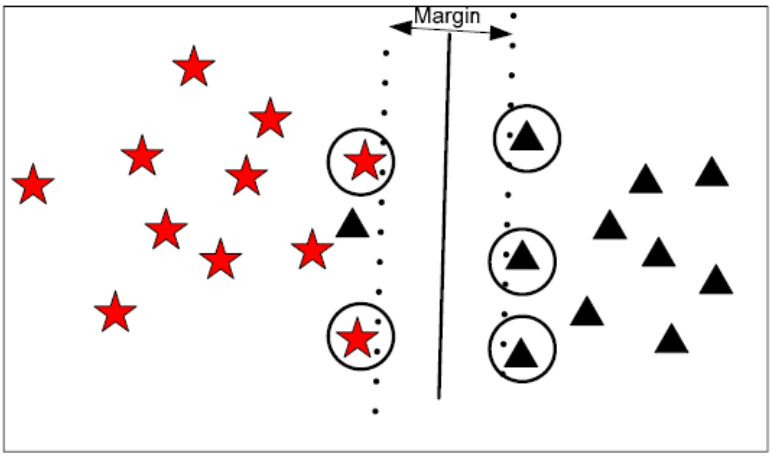
pri ograničenjima $y_i(\omega \cdot x_i + b) - 1 \geq 0$ i $y_i \in \{1, -1\}$ (Oommen, i drugi, 2008).

Međutim, većina problema klasifikacije nije linearno razdvojivo (primer na Slici 5.2.8). Kako bi se nosili sa ovim situacijama, podaci su mapirani u višedimenzionalni prostor svojstava. Iako je tradicionalno ovo postignuto pomoću nelinearnih transformacija, to pati od „prokletstva dimenzionalnosti“ (Hugheov efekat). Kod SVM-a metoda jezgra se koristi da omogući zamenu nelinearne transformacije sa unutrašnjim proizvodom koji se može definisati funkcijom jezgra ϕ. Unutrašnji proizvod ne zahteva evaluaciju prostora svojstava i time adresira probleme dimenzionalnosti a funkcija jezgra omogućuje podacima da se rasprše na način da linearna hiperravan može biti

prikladna. Time optimizacioni problem za maksimiziranje margine postaje kombinacija dva kriterijuma, maksimizacija margine i minimizacija greške. Definisan je sa

$$\min\left[\frac{\|\omega\|^2}{2} + C\sum_{i=1}^{r}\xi_i\right]$$

pri ograničenjima $y_i(\omega \cdot \phi(x_i) + b) - 1 \geq 1 - \xi_i$ i $\xi_i \geq 0, i = 1, \ldots, N$, gde je C vrednost „penala“ koji kontroliše jačinu penala povezanog sa klasifikacijom trening uzorka na pogrešnoj strani hiperravni, i ona uvodi balans između kompetitivnih kriterijuma maksimizacije margine i minimizacije greške. Promenjiva ξ_i ukazuje na razdaljinu pogrešno klasifikovanih tačaka od optimalne hiperravni (Oommen, i drugi, 2008).



Slika 5.2.8: Linearno nerazdvojivi skupovi podataka.

Dakle, SVM-ovi sadrže funkcije jezgra koje mapiraju podatke koji zahtevaju nelinearne diskriminacione funkcije iz prostora posmatranja u potencijalno višedimenzionalni prostor svojstava. U ovom prostorju svojstava, podaci su linearno razdvojivi. Najpoznatije vrste jezgara su linearno jezgro za linearnu diskriminacionu funkciju i jezgro funkcije radijalne osnove, koje omogućuju upotrebu nelinearnih diskriminacionih funkcija (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

Prilikom računanja, podaci se obično uređuju u matricu (X_{ij}), gde je svaki red (x_i=[x_{i1}, x_{i2}, ..., x_{ij}]) uzorak sastavljen od j svojstava (atributa). Kod nekih primena, dodatne ciljne varijable (nazivi ili labele klasa, y_i) se dodaju uzorcima (x_i). Sam algoritam izvlači iz tih trening podataka odgovarajući model, gde se kod SVM klasifikatora primenjuje princip maksimalne margine (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

5.2.4 Klasifikacija ostalim metodama mašinskog učenja

Ovde ćemo samo kratko dati pregled klasifikacije pomoću metoda Neuronskih mreža i tehnika Relaksiranog označavanja.

5.2.4.1 Neuronske mreže

Veštačle neuronske mreže ili neuronske mreže su kompaktna reprezentacija modela za analizu visokodimenzionalnih podataka. Opažanja (ulazi) i rezultati (izlazi) su pretstavljeni čvorovima, koji su inspirisani nervnim sistemom, i nazivaju se neuronima. Neuronske mreže su primenljive i u nenadgledanom i nadgledanom kontekstu. Najpoznatiji tip neuronskim mreža su *feedforward* mreže, ali se takođe koriste i *rekurentne mreže* i *samo-organizujuće mape* (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).

Modeli klasifikacije sa neuronskim mrežama imaju kao prednost nad statističkim metodama to da su neutralni od raspodela i nije potrebno priorno znanje o statističkim raspodelama klasa u izvorima podataka, kako bi se metoda primenila za klasifikaciju (Benediktsson, Swain, & Ersoy, 1990).

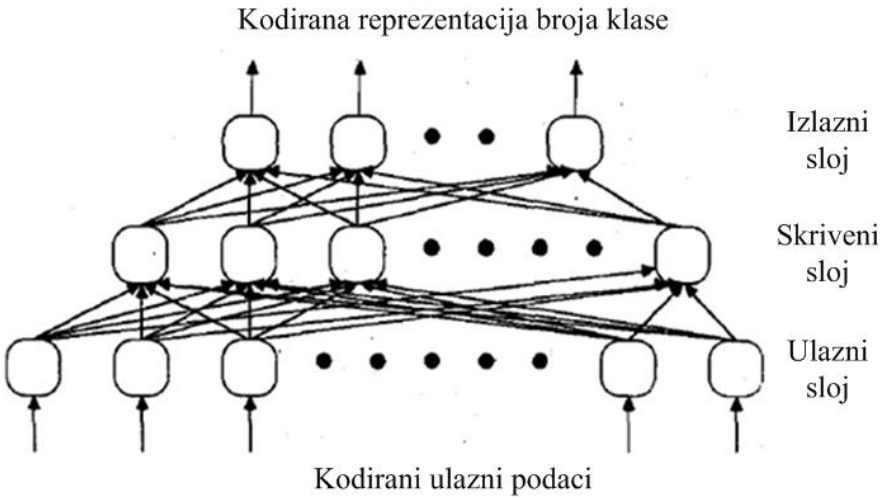
Neuronska mreža je mreža neurona gde se neuron može objasniti na sledeći način: neuron ima više (kontinuirano-vrednovanih) ulaznih signala x_j, j = 1, 2, ..., N, koji pretstavljaju aktivnost na ulazu ili trenutnu frekvenciju ili neuralne impulse koje tom ulazu daje drugi neuron. U najprostijem formalnom obliku neurona, izlazna vrednost ili frekvencija neurona, o, se često apriksimira funkcijom

$$o = K\phi\left(\sum_{j=1}^{N} w_j x_j - \theta\right)$$

gde je K konstanta a ϕ je nelinearna funkcija koja uzima vrednost 1 za pozitivne argumente i -1 (ili 0) za negativne argumente. Vrednosti w_j se nazivaju sinaptičke efikasnosti ili težine, i θ je prag (Benediktsson, Swain, & Ersoy, 1990).

Kod pristupa prepoznavanja obrazaca pomoću neuronskih mreža, neuronska mreža funkcioniše kao crna kutija koja prima skup ulaznih vektora x (opaženih signala) i daje odzive o_i iz svojih izlaznih jedinica i (i=1, ..., L, gde L zavisi od broja klsa informacija). Generalna ideja koju sledi teorija neuronskih mreža je da su izlazi ili o_i=1, ukoliko je neuron i aktivan za trenutni ulazni vektor x, ili o_i=-1, ukoliko je neaktivan. To znači da su vrednosti signala kodirane kao binarni vektori, i za određeni ulazni vektor x, izlaz daje binarnu reprezentaciju broja njegove klase. Onda je proces da se nauče težine kroz adaptivnu (iterativnu) proceduru treninga. Procedura treninga se okončava kada se mreža stabilizuje, tj. kada se težine ne menjaju iz jedne iteracije u sledeću ili kada se menjaju manje od zadatog praga vrednosti. Onda se podaci šalju mreži da izvrši klasifikaciju i mreža daje na izlazu broj klase svakog od piksela

(Benediktsson, Swain, & Ersoy, 1990). Šematski dijagram klasifikatora pomoću troslojne neuronske mreže je prikazan na Slici 5.2.9.



Slika 5.2.9: šematski dijagram neuronske mreže za klasifikaciju podataka slike.

Višeslojne neuronske mreže otklanjaju neke nedostatke dvoslojnih i troslojnih mreža. One, u principu, mogu pružati optimalna rešenja za proizvoljne probleme klasifikacije. U osnovi višeslojne neuronske mreže implementiraju *linearne diskriminatore,* ali u prostoru gde su ulazi mapirani nelinearno. Ključna snaga koju pružaju takve mreže je u tome da one omogućuju korišćenje prilično jednostavnih algoritama, pri čemu se oblik nelinearnosti može naučiti iz trening podataka. Modeli su time veoma moćni, imaju zgodne teoretske osobine, i primenjuju se dobro na veliki broj primena u realnog sveta (Duda, Hart, & Stork, 2000).

5.2.4.2 Tehnike relaksirajućeg obeležavanja

Tehnike relaksirajućeg obeležavanja (eng. *Relaxation Labelling techniques*) su uspešno primenjene na različite probleme analiza slika, uključujući klasifikaciju piksela. Metode relaksirajućeg obeležavanja su paralelne, interaktivne, ažurirajuće tehnike kojim se nedvosmisleno i konzistentno obeležavanje (klasifikacija, imenovanje) dobija korišćenjem kontekstualnih informacija. Postoje brojne metode relaksirajućeg obeležavanja, a neke od njih su: diskretne, *fuzzy* (rasplinute), linearne probabilističke metode i nelinearne probabilističke metode (Deekshatulu, i drugi, 1995).

Prikažimo ukratko koncept: ako je data slika veličine *M*x*N*, možemo definisati:

1. Inicijalne verovatnoće su date sa

$$P_i^0(\lambda), \qquad \lambda = 1, 2, \dots, L$$

gde je i=piksel, λ=klasa, L=broj klasa, i izračunati su ili pomoću klasifikacije maksimalnom izglednosti ili tehnikama klasterovanja.

2. Koeficijente komparabilnosti

$$P_{i,j}(\lambda, \lambda'), \qquad \lambda, \lambda' = 1, 2, \dots, L$$

Nekoliko statističkih modela se može koristiti za procenu koeficijenata komparabilnosti, poput:

(a) Koeficijent korelacije:

$$r_{ij}(\lambda, \lambda') = \frac{\sum\{p_{x,y}(\lambda) - p(\bar{\lambda})\}\{p_{x+i,y+j}(\lambda') - p(\bar{\lambda'})\}}{\sigma(\lambda) * \sigma(\lambda')}$$

(b) Zajednička entropija:

$$r_{i,j}(\lambda, \lambda') = \log\left\{\frac{\sum p_{x,y}(\lambda) * p_{x+i,y+j}(\lambda')}{p_{x,y}(\lambda) * p_{x+i,y+j}(\lambda')}\right\}$$

3. „Podrška“ susednih piksela J

$q_{i,j}(\lambda);\ \lambda = 1, 2, \dots, L$, gde je jedna formula

$$q_{i,j}(\lambda) = \sum_{\lambda} P_i^0(\lambda) P_j^0(\lambda')\, r_{ij}(\lambda, \lambda')$$

4. Ukupna „podrška“ suseda

$q_i(\lambda), \lambda = 1, 2, \dots, L$ gde je $q_i(\lambda) = \sum_j q_{i,j}(\lambda)$

5. Ažurirane verovatnoće klasa (jedno od različitih pravila ažuriranja):

$$P_i^1(\lambda) = \frac{\{P_i^0(\lambda) * q_i^0(\lambda)\}}{\{\sum_\lambda P_i^0(\lambda') * q_i^0(\lambda')\}}$$

i

$$P_i^n(\lambda) = \frac{\{P_i^{n-1}(\lambda) * q_i^{n-1}(\lambda)\}}{\{\sum_\lambda P_i^{n-1}(\lambda') * q_i^{n-1}(\lambda')\}}$$

Broj iteracija može biti proizvoljno određen na, npr., 30 ili proverom *n*-tog koraka da li je neka element verovatnoće vektora $P_i^n(\lambda)$ je veoma veliki, npr. 0.9 ili 0.95 (Deekshatulu, i drugi, 1995).

5.3 Tačnost klasifikacije u daljinskom uzorkovanju

Kod daljinskog uzorkvoanja procena tačnosti dobijenih rezultata (npr. tematskih mapa) je važan element i mora se sprovesti pre nego se one počnu koristiti i dalje primenjivati (Deekshatulu, i drugi, 1995; Khorram, Wiele, Koch, Nelson, & Potts, 2016). Ključno je da istraživači i korisnici daljinski uzorkovanih podataka imaju čvrsto znanje kako o faktorima tako i o tehnikama koje se koriste od procene tačnosti. Ukoliko nedostaje poznavanje takvih tehnika, to može značajno ograničiti efikasnu upotrebu daljinski uzorkovanih podataka (Congalton, 1991).

Tačnost se može definisati kao stepen (često kao procenat) podudaranja između opažanja i stvarnosti (Levin, 1999). (Campbell & Shin, 2011) navode dva osnovna atributa koji karakterišu kvalitet podataka, a to su *tačnost*, koja opisuje koliko je jedno merenje blizu stvarnim vrednostima i često je izražena kao verovatnoća (npr. 80% svih tačaka je u +/- 5 metara od svoje stvarne lokacije), i *preciznost*, koja se odnosi na odstupanje vrednosti kada se izvode ponovljena merenja. (Huisman & de By, 2009) takođe navode razliku između tačnosti i preciznosti: tačno merenje neke pojave ima srednju vrednost blizu stvarne vrednosti, dok precizno merenje ima dovoljno malu varijansu. Preciznost se još navodi kao izraz najmanje jedinice merenja koja može zabeležiti podatke.

Kod daljinskog uzorkovanja (Campbell & Shin, 2011) navode nekoliko tačnosti podataka o kojima je potrebno voditi računa, a to su: a) poziciona tačnost (podataka u odnosu na njihovu stvarnu lokaciju na Zemlji), b) tačnost atributa, odnosno svojstava (zabeležene informacije u odnosu na svojstva realnog sveta koja ona predstavljaju), c) temporalna tačnost (vezana za starost ili pravovremenost podataka), d) logička konzistentnost (osobina topološke ispravnosti), i e) kompletnost podataka (osobina podataka da uključuju sva potrebna svojstva). Mi ćemo se u nastavku praktično baviti samo tačnošću atributa (svojstava), budući da je u nastavku istraživanja od značaja tačnost klasifikacije podataka, odnosno klasifikovanje svojstava u određene klase.

Prema (Levin, 1999) nekoliko vrsta grešaka smanjuje tačnost identifikacije svojstava i raspodele kategorija a većina nastaje prilikom merenja ili u uzorkovanju. Najčešće greške koje tom prilikom nastaju (Levin, 1999) navodi kao: greške prikupljanja podataka (uključuju performanse senzora i platforme), greške obrade podataka (pogrešna registracija piksela u različiti opseg) i greške koje zavise od konkretne scene koja se posmatra (kako su klase definisane i određene). Prema (Khorram, Wiele, Koch, Nelson, & Potts, 2016) izvori grešaka uključenih u procenu tačnosti uključuju registraciju razlike između referentnih i daljinski uzorkovanih podataka, greške ocrtavanja (delineacije) prilikom digitalizacije, greške unosa podataka, greške kod klasifikacije i delineacije slika i greške pri uzorkovanju (semplovanju), prikupljanju i interpretaciji referentnih podataka.

5.3.1 Merenje tačnosti klasifikacije u daljinskom uzorkovanju

Kod daljinskog uzorkovanja mogu biti potrebni značajni napori za procenu tačnosti procedure klasifikacije (Huisman & de By, 2009). Kao opšte pravilo, nivo tačnosti koji se postiže klasifikacijom kod daljinskog uzorkovanja zavisi od niza faktora kao što je prikladnost trening skupa, veličina, oblik, raspodela i frekvencija pojavljivanja pojedinačnih oblasti dodeljenih svakoj od klasa, performansi senzora i rezolucije, i metoda kojim se radi klasifikacija, i drugi (Levin, 1999). Klasifikacija slike rezultira rasterskim fajlom u kome su pojedinačni elementi rastera označeni klasama, pa se provera kvaliteta rezultata obično radi pristupom uzorkovanja kod kog se izabere određeni broj rasterskih elemenata rezultata, i onda se za njih porede rezultati klasifikacije sa stvarnom klasom. Drugim rečima, provera tačnosti se sprovodi proverama nad određenim brojem uzorkovanih tačaka (Tempfli, Kerle, Huurneman, & Janssen, 2009; Khorram, Wiele, Koch, Nelson, & Potts, 2016; Huisman & de By, 2009).

U praksi, tačnost klasifikacije se može testirati na jedan od četiri načina (Levin, 1999; Tempfli, Kerle, Huurneman, & Janssen, 2009):

1. Terenska provera na izabranim tačkama, izabranim ili nasumično ili nekom mrežom tačaka (nerigorozna i subjektivna provera), čime se utvrđuje stvarna klasa;
2. Procena podudaranja između mape klasa i referentne mape (npr. slike veće rezolucije, koje se smatraju da su veće tačnosti), preklapanjem jedne preko druge (nerigorozno provera);
3. Statistička analiza numeričkih podataka dobijenih uzorkovanjem, merenjem i obradom podataka, korišćenjem testova poput srednjih kvadrata, standardne greške, analize varijanse, koeficijenta korelacije, linearna ili višestruka regresiona analiza, i Hi-kvadrat test (rigorozne provere), i
4. Računanjem matrice grešaka (konfuzije) (rigorozne provere).

Najčešći i najefikasniji način predstavljanja tačnosti klasifikacije daljinski uzorkovanih podataka je u formi *matrice grešaka* (eng. *Error matrix*), još poznata kao *matrica konfuzije* (eng. *Confusion matrix*) ili *matrica pogrešnih klasifikacija* (eng. *Misclassification matrix*), iz koje se kasnije računaju mere tačnosti (Congalton, 1991; Dutta, Stein, & Patel, 2008; Khorram, Wiele, Koch, Nelson, & Potts, 2016; Levin, 1999; Tempfli, Kerle, Huurneman, & Janssen, 2009; Prabhakar, Gintu, Geetha, & Soman, 2015; Huisman & de By, 2009). Matrica grešaka je kvadradna matrica brojeva koji u redovima i kolonama daju broj jedinica uzorka (npr. piksela, klastera, ili poligona) dodeljenih određenoj kategoriji u odnosu na stvarnu kategoriju koja je ustanovljena proverom sa terena. Kolone obično predstavljaju referentne podatke a redovi klase generisane klasifikacijom iz daljinski uzorkovanih podataka (Tabela 5.3.1).

Elementi na dijagonali su pikseli koji su tačno klasifikovani za svaku klasu informacija (Congalton, 1991; Dutta, Stein, & Patel, 2008). Matrica grešaka je efektivan način prikaza jer se tačnost svake klasifikovane kategorije prikazuje zajedno sa greškama inkluzije (greška dodele ili uključivanja) i greškama ekskluzije (greška isključenja ili izostavljanja) koje se javljaju u klasifikaciji. Greške uključivanja nastaju kada je piksel jedne klase netačno identifikovan da pripada drugoj klasi, ili od nepravilnog razdvajanja jedne klase u dve ili više klasa, a greške isključivanja nastaju kada pikseli koji bi trebali biti dodeljeni određenoj klasi se jednostavno ne nađu u njoj (Congalton, 1991; Levin, 1999).

Tabela 5.3.1: Primer matrice grešaka (konfuzije); Data ukupna tačnost iznosi (62+18+12)/100 = 92%.

Klasifikovani podaci	Referentni podaci		
	Šuma	Poljoprivreda	Urbana zona
Šuma	62	5	0
Poljoprivreda	2	18	0
Urbana zona	0	1	12

Polazeći od takve matrice grešaka, naredni korak može biti njena normalizacija ili standardizacija. Ova tehnika koristi iterativno proporcionalno podešavanje koje dovodi svaki red i svaku kolonu matrice do toga da u zbiru daju 1. Na ovaj način razlike u veličini uzoraka koji generišu matricu su eliminisane i pojedinačne vrednosti ćelija time postaju direktno uporedive (Congalton, 1991).

Matrica grešaka može koristiti kao polazna tačka niza deksriptivnih i analitičkih statističkih tehnika. Jedna od najprostijih deskriptivnih statistika je *ukupna tačnost*, koja se računa deljenjem ukupno tačnih (suma glavne dijagonale) sa ukupnim brojem piksela u matrici grešaka, a i tačnosti pojedinačnih kategorija se mogu računati na sličan način (Congalton, 1991). Pored deskriptivnih tehnika, matrica grešaka je dobra polazna tačka za mnoge analitičke statističke tehnike, što je naročito tačno za diskretne multivarijacione statističke tehnike (procedure), koje se koriste za izvođenje statističkih testova za tačnost klasifikacije kod digitalnih daljinski uzorkovanih podataka (Congalton, 1991; Khorram, Wiele, Koch, Nelson, & Potts, 2016). Jedna od najčešćih diskretnih multivarijacionih tehnika koja se koristi kod analize tačnosti je *Kappa analiza*. Rezultat Kappa analize je K_{HAT} (ili $\widehat{K}$) statistika (procena za Kappa), koja predstavlja jednu meru tačnosti (Congalton, 1991; Khorram, Wiele, Koch, Nelson, & Potts, 2016).

Sada možemo sumirati mere tačnosti koje se najčešće koriste u daljinskom uzorkovanju za procenu tačnosti klasifikovanja slike, a to su (Congalton, 1991; Dutta,

Stein, & Patel, 2008; Khorram, Wiele, Koch, Nelson, & Potts, 2016; Tempfli, Kerle, Huurneman, & Janssen, 2009; Prabhakar, Gintu, Geetha, & Soman, 2015) (primer dat u tabeli 5.3.2):

- *Ukupna tačnost* (eng. *Overall accuracy*, OA), još nazivana i *procenat ispravno klasifikovanih* (eng. *Proportion Correctly Classified*, PCC): odnos ukupno tačnih (suma glavne dijagonale) sa ukupnim brojem piksela u matrici grešaka.

$$OA = \frac{\sum_{i=1}^{r} x_{ii}}{N}$$

- *Korisnička tačnost* (eng. *User's accuracy*, UA), koja odgovara greškama dodele (uključivanja, ili greške tipa I), još se naziva i *pouzdanost* (eng. *Reliability*). Predstavlja procenat tačno klasifikovanih piksela u odnosu na sve piksele klasifikovane kao ta klasa tokom klasifikacije slike. Korisnička tačnost je verovatnoća da je određena referentna klasa takođe klasifikovana kao ta klasa. Kao ukupan indikator koristi se i *prosečna pouzdanost* (eng. *Average reliability*) koja se računa kao prosečna pouzdanost, odnosno kao suma pozdanosti podeljena sa brojem klasa.

$$UA_i = \frac{x_{ii}}{x_{i+}}$$

- *Proizvođačka tačnost* (eng. *Producer's accuracy*, PA), koja odgovara greškama izostavljanja (isključivanja, ili greške tipa II), još se naziva i samo *tačnost* (eng. *Accuracy*). Predstavlja procenat tačno klasifikovanih piksela u odnosu na sve piksele u datoj referentnoj (*ground truth*) klasi. Proizvođačka tačnost je verovatnoća da je uzorkovana tačka na mapi ta određena klasa. Kao ukupan indikator koristi se i *prosečna tačnost* (eng. *Average accuracy*) koja se računa kao prosečna proizvođačka tačnost, odnosno kao suma tačnosti podeljena brojem klasa.

$$PA_j = \frac{x_{jj}}{x_{+j}}$$

- *Kappa statistika*. Kapa statistika uzima u obzir činjenicu da čak i nasumična (slučajna) dodela klasa rezultuje određenim stepenom tačnosti. Na osnovu Kappa statistike može se testirati da li dva skupa podataka imaju različitu tačnost. Ova vrsta testiranja se koristi da evaluira različite podatke daljinskog uzorkovanja ili metode za generisanje prostornih podataka. Računa se iz matrice grešaka kao:

$$\widehat{K} = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} * x_{+i})}$$

gde je *r* broj redova u matrici, x_{ii} je broj opservacija u redu *i* i koloni *i*, x_{i+} i x_{+i} su totali reda *i* i kolone *i*, respektivno, i *N* je ukupan broj opservacija (npr. piksela).

Tabela 5.3.2: Primer matrice grešaka sa izračunatim greškama i tačnostima. Ukupna tačnost iznosi 53% a ukupna Kappa 0.32.

Rezultat klasifikovanja	Referentne (stvarne) klase				Ukupno	Greška dodele %	Korisnička tačnost %
	Ku-kuruz	Soja	Pšeni-ca	Šuma			
Kukuruz	35	14	11	1	61	43%	57%
Soja	4	11	3	0	18	39%	61%
Pšenica	12	9	38	4	63	40%	60%
Šuma	2	5	12	2	21	90%	10%
Ukupno	53	39	64	7	163		
Greška izostavljanja %	34%	72%	41%	71%			
Proizvođačka tačnost %	66%	28%	59%	29%			

Kod procene tačnosti i navedenih tehnika analize glavna pretpostavka je da su referentni podaci (stvarni podaci, sa zemlje) tačni (Dutta, Stein, & Patel, 2008). Međutim, različitih tehnika analize postoje i drugi faktori koje je potrebno uzeti u obzir kada se radi procena tačnosti. U realnosti, tehnike ne znače mnogo ako se ti faktori ne uzmu u obzir jer je ključna pretpostavka analiza opisanih gore ta da je matrica grešaka zaista tačan prikaz svih klasifikacija. Ukoliko bi matrica bila nepravilno formirana, onda bi sve analiza bile besmislene. Prema tome, potrebno je uzeti u razmatranje sledeće faktore: prikupljanje stvarnih podataka sa zemlje, šemu klasifikacije, prostorne autokorelacije, veličine uzoraka, način (šemu) uzorkovanja (Congalton, 1991).

Što se tiče seme uzorkovanja, ne preporučuje se korišćenje istog uzorka i za klasifikaciju (kao trening seta) i za procenu tačnosti, jer ovo može dati pokazatelje koji su previše optimistični (Levin, 1999). Postoji više načina uzorkovanja za odabir piksela. Potrebno je izabrati strategiju uzorkovanja, broj (veličinu) uzorka i oblasti uzorkovanja. Npr. u kontekstu podataka o zemljinom prekrivaču, preporučuju se strategije uzorkovanja jednostavni slučajni uzorak ili stratifikovani slučajni uzorak. Teorija uzorkovanja se koristi za određivanje veličine uzorka, itd. Jedinica uzorka može biti

tačka ili oblast neke veličine, može biti pojedinačni rasterski element ili može uključiti i susedne. Između ostalog, optimalna veličina oblasti uzorka zavisi od heterogenosti klasa (Tempfli, Kerle, Huurneman, & Janssen, 2009). Neki autori poput (Prabhakar, Gintu, Geetha, & Soman, 2015) koriste Monte Carlo (MC) metod za procenu tačnosti, tako što nakon određenog broja pokretanja simulacija računaju prosečnu ukupnu tačnost.

5.3.2 Analiza osetljivosti u daljinskom uzorkovanju

Analize neizvesnosti i osetljivosti su studije o tome kako se neizvesnost u ulazima modela reflektuje na neizvesnost u njegovim izlazima, kao i o kvantifikovanju neizvesnosti u izlazima modela (Dobrota & Dobrota, 2016). Primene daljinskog uzorkovanja kod kojih se odluke donose na osnovu informacija ekstrahovanih iz slika imaju veliku potrebu za analizom osetljivosti. Često ekstrahovanje informacija zahteva odabir većeg broja (ručno podešenih) parametara i veoma je korisno koristiti brzu analizu osetljivosti za identifikovanje onih parametara koji imaju najsnažniji uticaj na rezultat. To je naročito važno kod obrade slika u analizi, ali i kod drugih primena (Patz & Preusser, 2012). Za nas je ona od značaja u ovoj disertaciji kako bismo bili u prilici ispitati stabilnost procesa merenja i klasifikacije, te time dodatno stabilnost i učinkovitost razmatrane metode klasifikacije.

Prema svojoj osnovnoj definiciji, daljinsko uzorkovanje je jednostavno indirektno merenje. U velikom broju različitih praktičnih primena ne možemo meriti određene veličine direktno, ali možemo meriti neke druge veličine koje su povezane sa veličinama od interesa nekim poznatim relacijama (Ustinov, 2015). Jedna od bitnih pretpostavka smislenog eksperimenta daljinskog uzorkovanja je raspoloživost kvantitativnog okvira koji omogućava simulaciju izmerenih veličina za važeće procene veličina od interesa. Ukratko, takvi modeli pružaju kvantitativni opis uzročno-posledičnih veza između vrednosti od interesa, koji specificiraju objekat koji se proučava i rezultate merenja pomoću instrumenata daljinskog uzorkovanja. U tom smislu se ti modeli mogu nazvati direktnim, jer simuliraju odgovarajuću uzročnu vezu unapred: od vrednosti od interesa do izmerenih podataka. S druge strane, cilj daljinskog uzorkovanja je sleđenje uzročne veze u suprotnom smeru: od izmerenih podataka do vrednosti od interesa. Prema analogiji sa nekom nelinearnom jednačinom $f(x) = a$, to bi bilo poput nalaženja argumenta x na osnovu poznate vrednosti a, za koje važi data jednakost (Ustinov, 2015). Šira oblast teoretske osnove daljinskog uzorkovanja leži u istraživanju i razvoju različitih kvantitativnih metoda za nalaženje (procene) vrednosti od interesa iz izmerenih podataka. Po analogiji sa prethodnom jednačinom koja predstavlja problem, metod njegovih rešavanja je analogan Njutn-Raphson iterativnom metodu rešavanja te jednačine. Prema tom metodu rešavanja, zajedno sa mogućnosti simulacije merenih podataka, moramo biti u stanju izračunati izvode simuliranih

podataka u odnosu na vrednosti od interesa. Drugim rečima, traži se *osetljivost* (eng. *sensitivity*) simuliranih podataka u odnosu na te vrednosti (Ustinov, 2015).

Postoje tri načina implementacije analize osetljivosti kvantitativnih modela. Najjednostavniji je pristup *konačne razlike* (eng. *finite-difference*, FD) koji zahteva višestruka ponovna pokretanja direktnog modela u zavisnosti od broja ulaznih parametara modela, pri čemu se većina direktnih problema u daljinskom uzorkovanju može rešiti korišćenjem samo numeričkih metoda. Iako ovaj pristup koristi direktan model bez modifikacija i ne zahteva analitičke napore, njegova primena može rezultovati veoma računski-zahtevnim algoritmima, što može biti ograničavajući faktor u praktičnoj primeni za modele sa velikim brojem ulaznih parametara. Dva druga pristupa analizi osetljivosti su pristup *linearizacije* (eng. *linearization*) i *adjont* pristup i oni su značajno više računski efikasni i zahtevaju samo jedno pokretanje modela izvedenog iz inicijalnog, osnovnog modela (Ustinov, 2015).

Primene analize osetljivosti u daljinskom uzorkovanju se odnose na tri glavne oblasti: analiza grešaka ulaznih i izlaznih parametara (vrednosti) i rešenje inverznog problema. Analiza grešaka izlaznih vrednosti sa datom greškom ulaznih vrednosti je najdirektnija. Kada se planira eksperiment daljinskog uzorkovanja, važno je proceniti neizvesnosti izlaznih vrednosti na osnovu na osnovu definisanih grešaka ulaznih vrednosti, koje treba dobiti. Kada se dobijaju ulazne vrednosti iz merenja izlaznih vrednosti, važno je proceniti neizvesnosti dobijenih vrednosti na osnovu datih grešaka merenja. Inverzno modelovanje predstavlja suštinu daljinskog uzorkovanja. Merenja izlaznih vrednosti (vidljivih), pružaju informaciju koja se koristi da ograniči ulazne vrednosti modela, vrednosti od interesa, npr. za dobijanje njihovih procena, zajedno sa procenama njihove neizvesnosti (Ustinov, 2015).

Ovde ćemo još kratko navesti par primera iz literature, studija u kojima se koristi analiza osetljivosti kod daljinskog uzorkovanja. (Patz & Preusser, 2012) predstavljaju novi pristup za analizu osetljivosti kod obrade slika. Bazira na pojmu *stohastičkih slika* koji je korišćen u studiji za analizu osetljivosti, a originalno je predstavljen za propagiranje neizvesnosti u procesu prikupljanja slika. Analiza osetljivosti iz te studije je pokazala da se koncept stohastičkih slika može koristiti za modelovanje problema koji nastaju iz neizvesnosti informacija u obradi slika (Patz & Preusser, 2012). (Morris, Kottas, Taddy, Furfaro, & Ganapol, 2008) predstavljaju statistički pristup analizi osetljivosti *modela prenosa zračenja* (eng. *Radiative Transfer Models*, RTM). Fokus stavljaju na opštu analizu osetljivosti, proučavajući kako se izlazi RTM-a menjaju pri kontinualnoj promeni ulaza na osnovu raspodela verovatnoća kroz prostor ulaza. Uticaj svake promenljive ulaza je beležen kroz „glavne efekte“ i „pokazatelje osetljivost“ (Morris, Kottas, Taddy, Furfaro, & Ganapol, 2008). (Lechner, Langford, Bekessy, & Jones, 2012) istražuju nivo do koga studije pejzažne ekologije koje koriste prostorne podatke adresiraju neizvesnost u sprovođenju analiza. Autori identifikuju tri široke kategorije prostorne neizvesnosti koje su važne za određivanje karakterizacije pejzažnih šablona i utiču na izlaz analize: (a) neizvesnost šeme klasifikacije, (b) prostorna skala

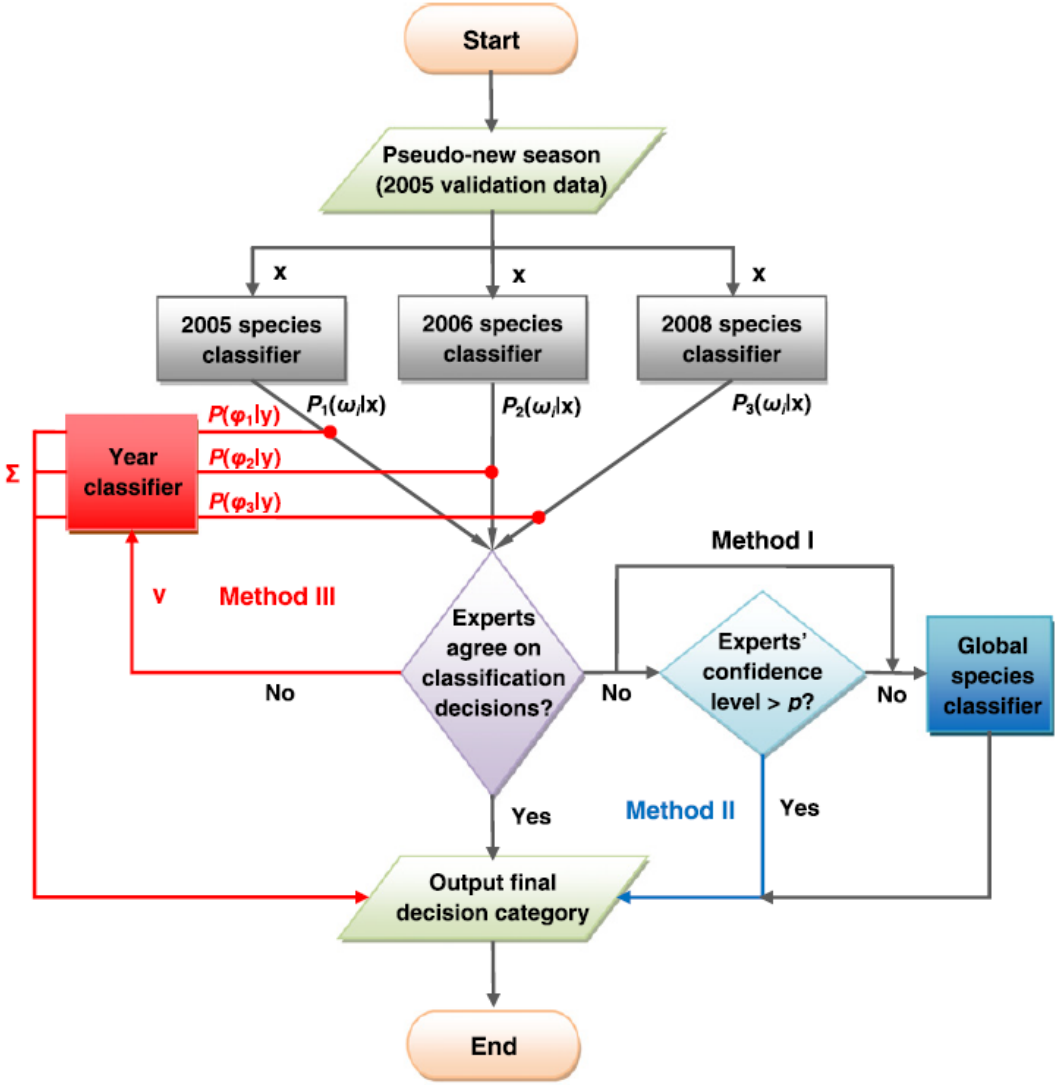
(veličina piksela, najmanja jedinica mapiranja, izglađivanje, tematska rezolucija i opseg), i (c) greške klasifikacije. S obzirom da stvarna količina i priroda greške usled neizvesnosti je nepoznata, istraživači moraju vršiti analizu osetljivosti ishoda koja dolazi iz niza prostornih neizvenosti i grešaka ulaza (Lechner, Langford, Bekessy, & Jones, 2012).

5.4 Pregled studija iz literature

U ovom poglavlju predstavićemo pregledno radove koji su u vezi sa temom ovog istraživanja i koji se na različite načine bave problemima klasifikacije u daljinskom uzorkovanju. U tim radovima autori predlažu i/ili razrađuju azličite, uglavnom problemski specifične, metode za klasifikaciju podataka iz višestrukih izvora. Predlagane metode su statističke, heurističke, i iz domena mašinskog učenja, korišćenjem tehnika nadgledanog i nenadgledanog učenja. U nastavku se kratko navodi tematika kojom se bavi svaki od pregledanih radova i studija slučaja.

(Benediktsson, Swain, & Ersoy, 1990) su se bavili razvojem opštije metode koja se može primeniti za klasifikaciju bilo koje vrste podataka, i u tom smislu razmatraju dva pristupa: statistički (parametarski) pristup i pristup neuronskih mreža (pristup bez raspodela). Kod statističkog pristupa pažnju su usmerili na multivarijacionu statističku analizu u smislu metoda koji bazira na Bajesovoj teoriji klasifikacije (Bajesovoj klasifikaciji), a kojeg proširuju uzimajući u obzir relativnu pouzdanost podataka na izvoru, uključenih u klasifikaciju (Benediktsson, Swain, & Ersoy, 1990).

(Zhang, Slaughter, & Staab, 2012) istražuju robustnost prepoznavanja biljaka na osnovu hiperspektralnih slika u prirodnom okruženju i u kontekstu automatske kontrole korova unutar redova useva. Razvijen je sistem mašinske vizije pomoću CCD kamere integrisane sa linijskim spektrografom za detekciju i mapiranje korova iz malih blizina. Razvijena su tri kanonična Bajesova klasifikatora korišćenjem refleksije listova (400-795 nm) na osnovu podataka prikupljenih za paradajz tokom tri sezone. Performanse tri sezonski-specifičnih klasifikatora su testirane promenom uslova okruženja. Pre klasifikacije hiperspektralnih slika, pozadinsko zemljište je segmentirano od piksela biljnog prekrivača pomoću vegetativnog indeksa modifikovanog odnose crvene (MRVI). U studiji je primenjena multivarijaciona diskriminaciona analiza na osnovu Bajesove teorije odlučivanja (Duda, Hart, & Stork, 2000), kako bi razvio klasifikator biljnih vrsta pomoću refleksije biljnog prekrivača. Primenjeni algoritam je šematski prikazan na Slici 5.4.1 (Zhang, Slaughter, & Staab, 2012).



Slika 5.4.1: Šema mašinskog učenja multi-klasifikatorskog sistema za robustno multi-sezonsko prepoznavanje biljaka.

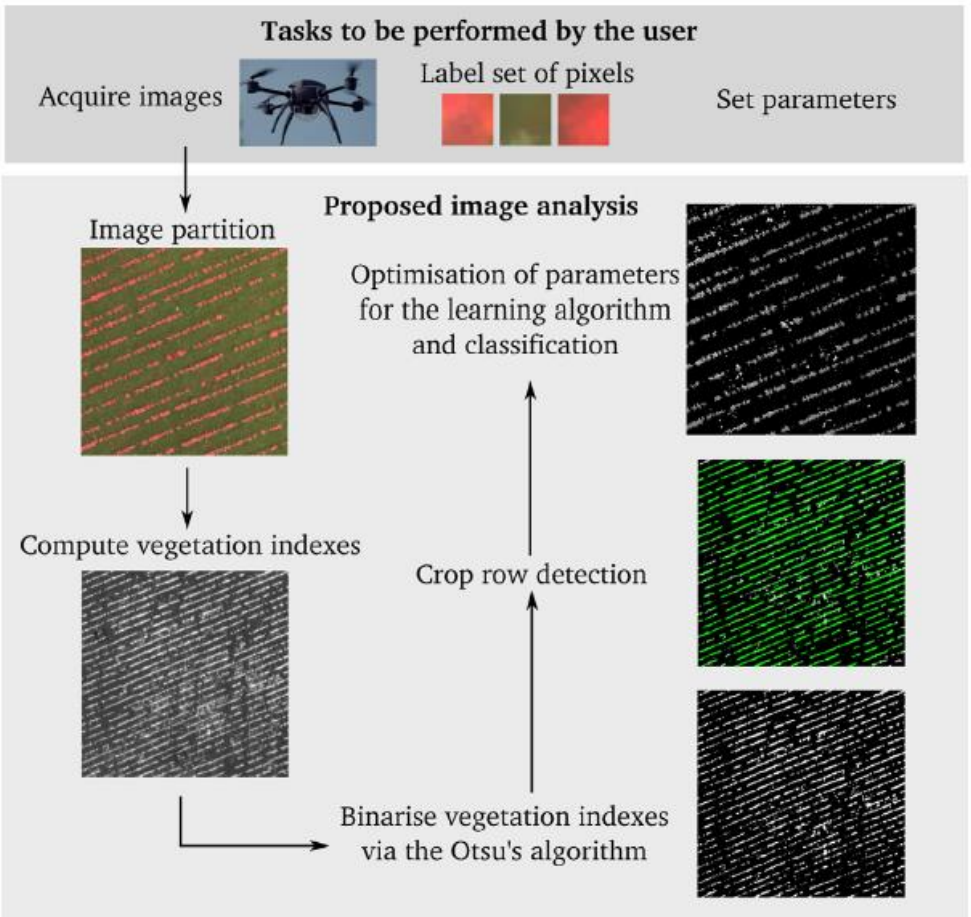
(Larsolle & Muhammed, 2007) se bave sličnom studijom kao i (Zhang, Slaughter, & Staab, 2012). Korišćeni su podaci iz dva terenska eksperimenta, procena nivoa bolesti kod pšenice i merenje gustine biljaka. Metod analize se sastoji iz dva koraka: pretprocesiranja gde su podaci normalizovani i klasifikacije za procenu useva. Za trening uzorak korišćeno je 12% podataka. Za klasifikaciju, deo terenskih merenja je korišćen za dobijanje referentnih podataka. Taj referentni skup podataka je korišćen zajedno sa klasifikatorom najbližih suseda nad novim, neklasifikovanim, hiperspektralnim podacima (Larsolle & Muhammed, 2007).

(Jannoura, Brinkmann, Uteau, Bruns, & Joergensen, 2015) se bave primenom daljinskog uzorkovanja u poljoprivredi pomoću bespilotnih letelica. Oni navode kako poslednjih godina razvoj lakih bespilotnih letelica omogućava novo rešenje i primenu za upravljanje i nadzor useva. Vazdušne fotografije su prikladna tehnika za nadzor biljaka, pružajući kvantitativne informacije o stanju useva i prostornom varijabilitetu. Vegetatitivni indeksi dobijeni iz tih slika se mogu koristiti za procenu promena u stanju useva, biomase i koncentracije hlorofila. Neki vegetativni indeksi se dobijaju kombinovanjem refleksije u vidljivom i infracrvenom delu spektra (npr. NDVI), dok se drugi mogu dobiti samo korišćenjem vidljivog dela spektra (GRVI, NGRDI). Autori primenjuju ove pokazatelje za praćenje fenologije, određivanje biomase i statusa hraniva za lokalizovano suzbijanje korova (Jannoura, Brinkmann, Uteau, Bruns, & Joergensen, 2015).

(Candiago, Remondino, Giglio, Dubbini, & Gattelli, 2015) se bave sličnom studijom kao (Jannoura, Brinkmann, Uteau, Bruns, & Joergensen, 2015) i predstavljaju analizu useva (vinograda i paradajza) pomoću multispektralnih podataka, prikupljenih pomoću mutlispektralne kamere na bespilotnoj letelici. Na osnovu ortoslika dobijeni su različiti vegetacioni indeksi (VI), poput NDVI, GNDVI i SAVI, koji ukazuju na stanje biljaka. Rad istražuje: planiranje fotogrametrije i obradu multispektralnih podataka prikupljenih pomoću UAV platforme, generisanje ortofotografija visoke rezolucije iz multispektralnih slika, identifikaciju i maskiranje pozadinskog zemljišta, generisanje i evaluaciju različitih VI i na kraju statističku analizu vegetacionih indeksa (VI). Krajnji cilj je diskriminacija vegetacije istih useva sa različitim VI bez korišćenja radimetrijskih merenja sa zemlje, međutim opisi poljoprivrednih karakteristika koje su dali lokalni farmeri su korišćeni za validaciju zaključaka. Autori navode veliki potencijal podataka visoke rezolucije prikupljenih pomoću dronova i predlažu da ti instrumenti predstavljaju brz, pouzdan, i jeftin resurs u proceni stanja useva za primene u preciznoj poljoprivredi (Candiago, Remondino, Giglio, Dubbini, & Gattelli, 2015).

(Pérez-Ortiz, i drugi, 2015) opisuju sistem za mapiranje korova na osnovu slika prikupljenih bespilotnim letelicama (UAV-ova, odnosno dronova). Kontrola korova u preciznoj poljoprivredi bazira na projektovanju lokacijsko-specifičnih kontrolisanih tretmana na osnovu zastupljenosti korova. Za to, tradicionalne platforme daljinskog uzorkovanja (avioni i sateliti) nisu pogodni zbog svoje niske prostorne i temporalne rezolucije. Predloženi metod za mapiranje korova particioniše sliku i dopunjuje spektralne informacije drugim izvorima informacija. Sistem uključuje tehnike klasifikacije za karakterisanje piksela kao usev, zemljište i korov. Autori porede različite paradigme mašinskog učenja kako bi identifikovali strategije sa najboljim performansama, uključujući nenadgledane, polu-nadgledane i nadgledane tehnike. Ciljevi studije i razrade sistema su navedeni kao: 1) proučiti kako kombinovati slike iz

dronova sa metodama detektovanja useva u redovima, kako bi se unapredile performanse, i 2) analiza potencijala različitih metoda mašinskog učenja u razvoju algoritma korišćenjem minimalnog skupa informacija, kako bi se dobila uglavnom nenadgledana analiza koja bi se mogla lako koristiti u realnim situacijama na terenu. Metodologija se zasniva na sledećim koracima: particionisanje eksperimentalnog polja u različite pod-slike, računanje vegetacijskih indeksa (VI) i njihova binarizacija, detekcija redova useva (na osnovu Houghove transformacije, HT), i na kraju treniranje modela klasifikacije na osnovu obezbeđenih podataka (gde se postavlja i pitanje automatske optimizacije različitih parametara). Slika 5.4.2 prikazuje korake algoritma obrade slika koji predlažu autori. Raličite paradigme učenja su testirane u eksperimentima kako bi se sprovela detaljna analiza tehnika i ustanovilo koja pruža najrobustnije rezlutate za svrhu mapiranja korova. Autori koriste šest metoda: dva algoritma koji koriste navođenje centroida kao metode nenadgledanog učenja (*k*-means i ponovljeni *k*-means, tj. *Rk*-means), polu-nadgledani SVM metod (SS-SVM) i tri nagledane tehnike (*k*-nn, tj. *k*-najbližih suseda, linearnu verziju SVM-a, tj. LinSVM i verziju SVM-a sa jezgrima). Kod SVM tehnike autori testiraju i linearnu i verziju zasnovanu na jezgrima kako bi uporedili rezultate i analizirali nelinearnost problema učenja (Pérez-Ortiz, i drugi, 2015).



Slika 5.4.2: Koraci predloženog sistema za mapiranje korova.

(Hague, Tillett, & Wheeler, 2006) proučavaju sličan problem kao (Pérez-Ortiz, i drugi, 2015). Oni segmentiraju transformisanu sliku u delove zemljišta i vegetacije korišćenjem jednog fiksiranog praga vrednosti. Primenjuju algoritam za robustno lociranje redova useva. Zone procene nivoa useva se automatski pozicioniraju, za useve direktno iznad redova useva, a za korov između redova. Odnos prebrojanih piksela useva i korova je poređen sa ručnom procenom gustine oblasti, na osnovu drugih slika veće rezolucije. Korelacija ručnog i automatskog merenja je utvrđivana i korišćena za kalibraciju automatskog pristupa (Hague, Tillett, & Wheeler, 2006).

(Prabhakar, Gintu, Geetha, & Soman, 2015) predstavljaju brz, pouzdan i efikasan metod za unapređenje klasifikacije hiperspektralnih slika, pripomognut segmentacijom. Oni proširuju algoritam multinomialne logističke regresije (MLR) na polu-nadgledano učenje posteriornih raspodela klasa korišćenjem neklasifikovanih uzoraka aktivno biranih iz skupa podataka. Rezultati klsifikacije dobijeni kroz regresioni model su unapređeni segmentacijom maksimalnih posteriora, što uključuje prostorne informacije hiperspektralne slike. Multinomialna logistička regresija (MLR) je takav deterministički pristup koji direktno uči i modeluje posteriorne verovatnoće klasa u Bajesovim okvirima bez učenja klasno-uslovnih gustina, što olakšava klasifikaciju sa manjim skupom trening podataka i manje kompleksnosti (Prabhakar, Gintu, Geetha, & Soman, 2015).

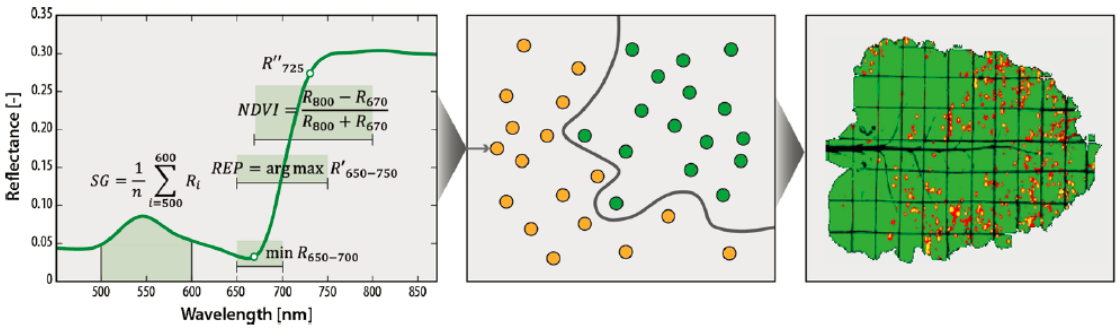
(Oommen, i drugi, 2008) prikazuju komparativnu analizu metoda SVC i klasifikacije maksimalnom izglednošću (MLC), koji navode kao najpopularniju konvencionalnu tehniku klasifikacije nagledanim učenjem. SVC je tehnika optimizacije kod koje se tačnost klasifikacije oslanja na identifikaciji optimalnih parametara. Pomoću studije slučaja, autori proveravaju metod dobijanja tih optimalnih parametara, kako bi se SVC mogao efikasno primeniti. Autori koriste multispektralne i hiperspektralne slike za dobijanje tematskih klasa poznatih litoloških jedinica za poređenje tačnosti klasifikacije oba metoda (Oommen, i drugi, 2008).

(Bauer, Korč, & Forstner, 2011) koriste stereo slike istih listova šećerne repe pomoću dve kamere (RGB i multispektralne) u laboratorijskim uslovima, pri čemu su listovi bili ili zaraženi tačkastim patogenom ili zdravi. Autori primenjuju tri metode klasifikacije: dve metode klasifikacije piksela, gde su klase piksela određene piksel po piksel i dodeljene nezavisno jedan od drugog, i globalni pristup, gde su klase piksela određene i dodeljene simultano. Za klasifikaciju na osnovu pojedinačnih piksela, autori razmatraju potencijal dve metode: *k*-najbližih suseda (kNN), zbog svoje jednostavnosti i nezavisnosti od raspodele podataka, i adaptivne Bajesove klasifikacije najmanjim

rizikom, pretpostavljajući mešani Gausov model (GMM), koji je sofisticirana verzija klasifikacije maksimalnim posteriorima (MAP) koji omogućuje pojedinačno ponderisanje različitih klasa. Kod globalno pristupa, autori koriste model uslovnih slučajnih polja (eng. *Conditional Random Field*, CRF) (Bauer, Korč, & Forstner, 2011).

(Webb & Copsey, 2011) se bave problemom klasifikacije zemljišnog prekrivača pomoću hiperspektralnih satelitskih daljinksi uzorkovanih slika. Oni porede konvencionalnu tehniku klasifikacije (linearni spektralni mešani modeli), i sa teoretske i praktične strane, sa SVM-ima. Performanse SVM klasifikatora su poređene sa RBF klasifikatorom i kNN klasifikatorom. Poređenja su rađenja i pre i posle upotrebe kriterijuma selekcije svojstava kako bi se redukovala dimenzionalnost podataka. Podaci sadrže merenja dve klase zemljišnog prekrivača. Kreirani su trening i testni skupovi podataka. Trening skupovi se sastoje od „čistih“ piksela, tj. onih koji se odnose na regiju za koju postoji samo jedna klasa. Primenjen je SVM model, pri čemu su trenirani linearno razdvojivi i linearno nerazdvojivi SVM-ovi, korišćenjem uzora izabranih iz trening seta. Razmatrana su dva SVM jezgra: linearno i Gausovo. SVM parametri (C i σ za Gausovo jezgro) su optimizovani analizom performansi nad trening podacima. Utvrđeno je da je SVM sa Gausovim jezgrom dao najbolje performanse u smislu tačnosti klasifikacije, sa i bez selekcije svojstava (Webb & Copsey, 2011).

(Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015) opisuju proces klasifikacije za detekciju infekcije na listovima šećerne repe. Proces uključuje: ekstrakciju svojstava iz opaženog reflektovanog spektra (vegetacionih indeksa, VI), učenje nelinearnog modela klasifikacije (nelinearni SVM) nad trening podacima i dalju primenu modela nad novim podacima za detekciju piksela koji predstavljaju zaraženi list, i na kraju vizuelizaciju predviđanja (Slika 5.4.3) (Behmann, Mahlein, Rumpf, Romer, & Plumer, 2015).



Slika 5.4.3: Grafički prikaz sva tri koraka procesa klasifikacije.

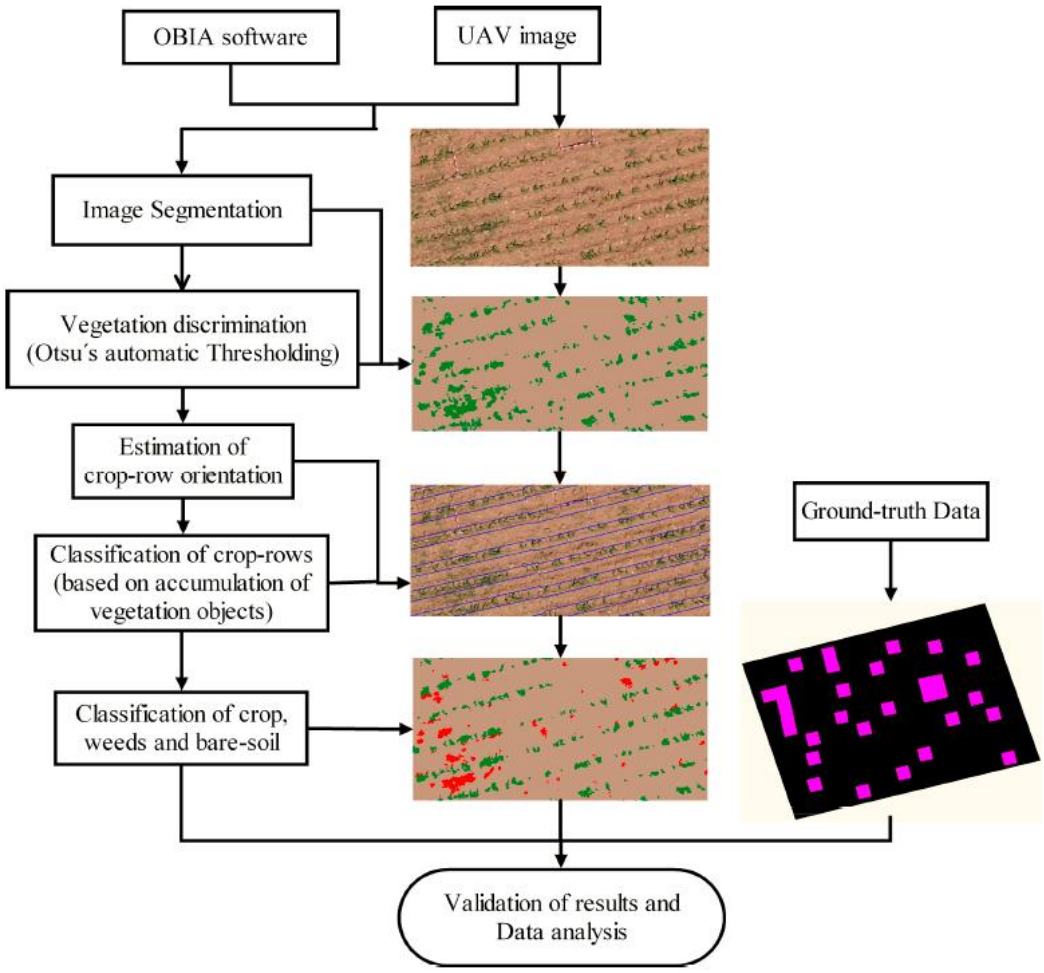
(Muhammed, 2005) se bave identifikacijom jedinstvenih potpisa za specifičan stres kod biljaka, uprkos promenama koje se odnose na normalan razvoj i rast useva. Studija koju sprovode autori se bavi karakterisanjem i procenom nivoa gljivičnih oboljenja kod pšenice. Oni koriste referentne skupove podataka koji se sastoje od hiperspektralnih vektorskih podataka refleksije i odgovarajućih nivoa bolesti procenjenih na polju. Nakon normalizacije hiperspektralnih vektora na nove hiperspektralne podatke, koristi se klasifikator najbližih suseda (kNN) za klasifikaciju novih podataka u odnosu na referentne podatke. Odgovarajući potpisi stresa se onda računaju korišćenjem modela linearne transformacije (Muhammed, 2005).

(Nieuwenhuizen, Tang, Hofstee, Muller, & Henten, 2007) razmatraju razvoj i poređenje dva algoritma mašinske vizije (eng. machine vision) za detekciju korova u poljima šećerne repe. Slike su prikupljene kolor fotoaparatom sa platforme koja se kreće po zemlji. Obrada slike se sastoji od tri glavna koraka, predprocesiranja slike, klasifikacije piksela i klasifikacije objekata biljki. U drugom koraku je zeleni materijal biljke segmentiran u odnosu na pozadinu – zemljište. Nako eliminacije pozadine, preostali pikseli biljaka su transformisani pomoću EGRBI transformacione matrice, koja odvaja intenzitet informacija od informacije o boji i dalje omogućuje analizu samo na osnovu boja. Za klasifikaciju piksela korišćeni su dva metoda. Prvi metod je kombinacija *k*-means klasterovanja i Bajesovog klasifikatora. Pikseli biljaka su klasterovani pomoću *k*-means algoritma, sa 8 polaznih, nasumično izabranih centroida. Klasteri korova su identifikovani na slici i označeni ručno na trening slici. Odgovarajuće RGB vrednosti označenih klastera su predstavljale ulaz kao priorni podaci, predstavljajući klasu korova za to polje, za Bajesovu rutinu klasifikacije. Drugi metod je bio treniranje neuronske mreže za klasterovanje na osnovu Euklidskog rastojanja, na osnovu čega je formiran klasifikator (Nieuwenhuizen, Tang, Hofstee, Muller, & Henten, 2007).

(Jovanović, i drugi, 2015) predstavljaju različite metode analize satelitskih slika, sa ciljem identifikacije promena zemljinog prekrivača kroz određeni vremenski period. Predstavljene metode uključuju diferenciranje vegetacionih indeksa, nadgledanu klasifikaciju i klasifikaciju baziranu na objektima (eng. *Object-based classification*). Zaključeno je da nagledana klasifikacija daje najtačnije rezultate kod slika srednje prostorne rezolucije, a u studiji je korišćeno više različitih algoritama: paralelopipedski klasifikator, klasifikacija na osnovu prostornih karakteristika, najmanja udaljenost i maksimalna uslovna verovatnoća. Tehnika klasifikacije maksimalnom izglednošću je korišćena za sve spektralne opsege, svake od slika. *Analiza slika na osnovu objekata* (eng. *Object based image analysis*, OBIA) predstavlja jedan od novijih načina klasifikacije, koji je nastao iz potrebe da se proces obrade slike prilagodi ljudskom razumevanju slike, objekata i prostora (Jovanović, i drugi, 2015).

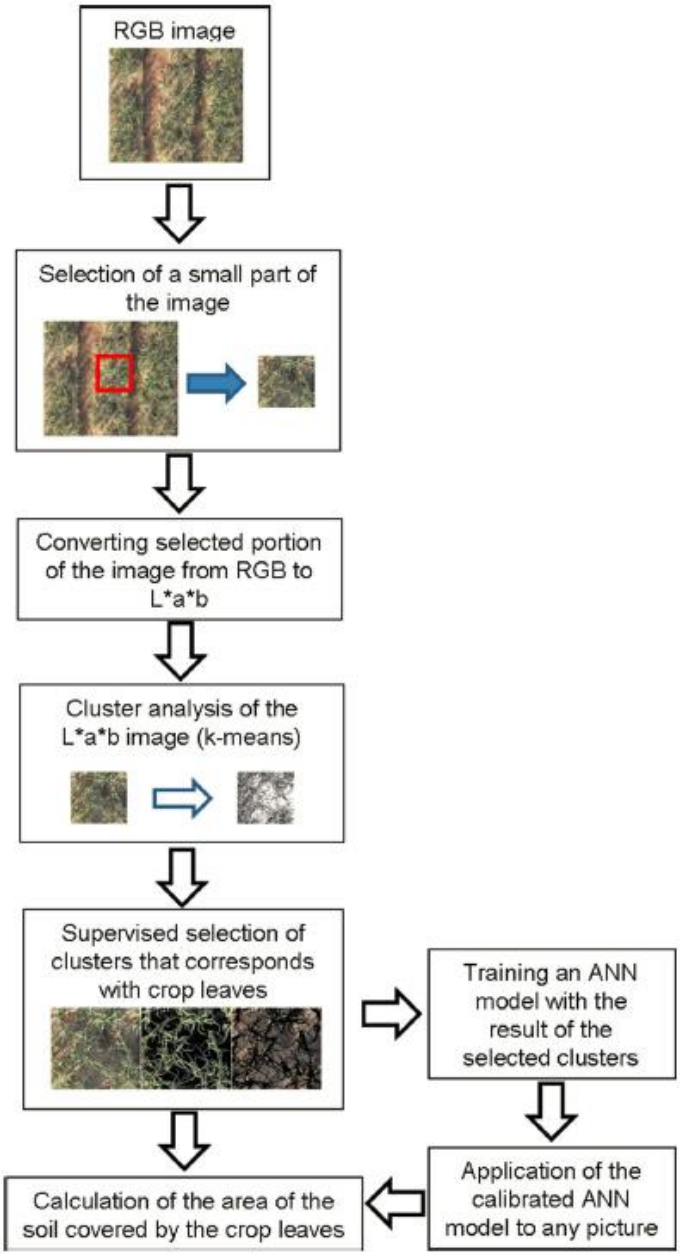
(Torres-Sánchez, López-Granados, Castro, & Peña-Barragán, 2013) opisuju tehničku specifikaciju i konfiguraciju bespilotne letelice (UAV) za prikupljanje daljinskih slika za lokalizovano upravljanje korovom u ranom delu sezone (eng. *early season site-specific weed management*, ESSWM). Takođe evaluiraju prostorne i spektralne karakteristike slika neophodnih za identifikaciju korova (u poljima suncokreta). Autori navode kako metode zasnovane samo na pikselima mogu biti neuspešne kod diskriminacije useva i korova u fazi sadnica za visine snimanja veće od 30 metara, zbog spektralne sličnosti između tih klasa vegetacije. Autori navode kako se spektralna ograničenja mogu rešiti implementiranjem naprednih algoritama poput OBIA metodologije, koja identifikuje prostorno i spektralno homogene jedinice (koje se nazivaju objektima) pomoću grupisanja susednih piksela putem procedure segmentacije. Nakon toga, višestruka svojstva lokalizacije, teksture, bliskosti i hijerarhijskih veza se koriste da drastično povećaju tačnost klasifikacije slike. Npr. kod useva u ranoj fazi, relativna pozicija biljke u redu useva može biti ključno svojstvo identifikacije korova, pre nego spektralna informacija (svaka biljke koja se ne nalazi u redu useva bi se mogla smatrati korovom). Prema tome, autori predlažu strategiju za robustnu klasifikaciju slika iz prikupljenih dronovima, koja uključuje dva koraka: a) diskriminaciju vegetacije (korova i useva) od zemljišta pomoću spektralnih informacija, i b) diskriminaciju korova od redova useva korišćenjem OBIA metodologije (Torres-Sánchez, López-Granados, Castro, & Peña-Barragán, 2013).

(Peña, Torres-Sánchez, Serrano-Pérez, Castro, & López-Granados, 2015), slično kao i u gornjoj studiji, razvijaju OBIA (analiza slika na osnovu objekata) proceduru za mapiranje korova. Procedura bazira na algoritmu mapiranja korova, istraženom na poljima kukuruza u ranom delu sezone, i prilagođena je specijalnim karakteristikama useva suncokreta. Čitav proces je automatizovan i sastavljen je od niza rutima, koje su: (a) segmentacija polja u podoblasti, (b) segmentacija podoblasti u objekte, (c) diskriminacija objekata vegetacije, (d) klasifikacija redova useva, (e) diskriminacija korova i useva, i (f) procena nivoa zastupljenosti korova. Šematski prikaz dat je na Slici 5.4.4 (Peña, Torres-Sánchez, Serrano-Pérez, Castro, & López-Granados, 2015).



Slika 5.4.4: Šematski prikaz OBIA procedure primenjene za klasifikaciju useva u redovima i detekciju korova.

(Ballesteros, Ortega, Hernández, & Moreno, 2014) sprovode studiju prikupljanja slika pomoću konvencionalne RGB kamere korišćenjem bespilotne letelice (UAV, tj. drona) i njihove obrade za dobijanje georeferenciranih orto-slika, sa ciljem karakterisanja glavnih parametara rasta biljaka neophodnih za menadžment navodnjavanja useva u polu-sušnim uslovima. Oni prvo opisuju proces prikupljanja slika i procedure obrade, a zatim primenjuju predloženu metodologiju na studiju slučaja. Korišćeni metod je šematski prikazan na slici 5.4.5 (Ballesteros, Ortega, Hernández, & Moreno, 2014).



Slika 5.4.5: Dijagram korišćenog algoritma (metoda); ANN je veštačka neuronska mreža.

(Song, i drugi, 2009) se bave studijom čiji je cilj razvoj i poređenje različitih metoda označavanja menadžment zona (MZ) na poljima pšenice. Oni koriste *fuzzy k-means* algoritam klasterovanja za definisanje menadžment zona, zajedno sa fazi indeksom performansi (FPI) i modifikovanom entropijom particionisanja (MPE) za određivanje optimalnog broja klastera. Statističke analize, uključujući Kappa koeficijent, su sprovođenje za sistematsko poređenje metoda za generisanje MZ-ova (Song, i drugi, 2009).

(Hung, Xu, & Sukkarieh, 2014) predlažu alternativni pristup zasnovan na učenju korišćenjem učenja svojstava za minimizaciju potrebnih manualnih napora. Sistem je primenjen za klasifikaciju invanzivnih vrsta korova. Oni primenjuju učenje svojstava za generisanje baze filtera slike koji omogućuje ekstrakciju svojstava koji diskriminišu korov od interesa u odnosu na ostale pozadinske objekte. Ta svojstva se objedinjuju (udružuju) da bi se sumirala statistika slike i formirao ulaz u linearni klasifikator koji klasifikuje deluje slike kao korov ili pozadina (Hung, Xu, & Sukkarieh, 2014).

(Franke & Menz, 2007) koriste tri daljinski uzorkovane slike visoke rezolucije za sprovođenje prostorno-vremensku analizu dinamike širenja bolesti kod pšenice, na eksperimentalnim poljima. Oni klasifikuju podatke podatke u oblasti koje predstavljaju različite nivoe ozbiljnosti oboljenja pomoću stabla odlučivanja, korišćenjem rezultata uparenog filtriranja mešanog podešavanja (MTMF) i vegetativnog indeksa normalizovane razlike (NDVI). Rezultati klasifikacije su poređeni sa podacima sa zemlje (ground truth) (Franke & Menz, 2007).

Ovde završamo pregled studija iz literature; a neki dodatni primeri se mogu naći u (Rajan & Maas, 2009; Römer, i drugi, 2011; Stein, i drugi, 1998) a odličan pregled različitih tehnika i metodologija dat je još u (Lee, Alchanatis, Yang, Hirafuji, & Moshou, 2010).

6 STATISTIČKI PRISTUP DEFINISANJU ZONE OSETLJIVOSTI

Statističke metode obezbeđuju donošenje odluka u sadašnjosti na osnovu ponašanja sistema u prethodnim trenutcima. Korišćenje probabilističkog pristupa predstavlja jedan od najegzaktnijih načina za uspešno predviđanje (Radojičić, 2001). Kod određivanja kategorija merenih pojava važno je odrediti granice tih pojava, odnosno odrediti one oblasti u kojima možemo sa sigurnošću pretpostaviti da entiteti pripadaju datoj kategoriji. U slučajevima kada nismo u mogućnosti da utvrdimo stanje (klasu) entiteta, potrebno je definisati *oblast* ili *zonu osetljivosti* koja predstavlja granične vrednosti kategorija (Radojičić, 2001). Ako je proizvod merenja neke pojave njen intenzitet na osnovu koje se vrši određivanje kategorije (klasifikacija), na osnovu višedimenzionalnog uzorka moguće je definisati odgovarajuću meru za to, koja se primenom statističkih metoda može testirati. Osnov za uspešno rešavanje takvog problema je sam proces merenja. Generalno, ispitivanje (istraživanje) predstavlja metodologiju objektivnog ili subjektivnog utvrđivanja osobina ispitivanog objekta, odnosno svojstva ispitivane pojave. U skladu s tim, ispitivanje se zasniva na procesu merenja odgovarajućih veličina koje određuju objekat, odnosno posmatranu pojavu (Radojičić, 2001).

Merenje kao proces ne predstavlja lak zadatak, posebno u realnim uslovima kada se suočavamo sa raznim uticajima koji ometaju kvalitetno merenje. Proces kreiranja kvalitetne mere za bilo koju vrstu pojave i učenje iz dobijenih rezultata, s ciljem implementiranja dobijene mere kao mere intenziteta (nivoa, kategorije) posmatrane pojave, predstavlja osnovni i suštinski cilj samog procesa (Radojičić, 2001). Statistički pristup merenju bazira se na samoj prirodi statistike. Statistika kao nauka bavi se istraživanjima masovnih pojava. Proces statističkog merenja obuhvata metode snimanja (prikupljanja) i obrade posmatranih pojava. Kao rezultat njihove primene dobijamo brojčane informacije o posmatranim pojavama. Za uspešno i efikasno merenje potrebno je iz populacije sa određenom karakteristikom izabrati odgovarajući uzorak, koji će moći efikasno da reprezentuje populaciju koja se posmatra. Na osnovu tako izabranog uzorka vrši se postupak statističkog zaključivanja (Radojičić, 2001).

U našem slučaju, ispitivanje i merenje veličina se vrši pomoću daljinskog uzorkovanja, objekti ili posmatrane pojave su zemljišni prekrivač ili konkretnije, u primerima koje istražujemo – kod primena u preciznoj poljoprivredi, različiti usevi, korov, zemljište, dok merene veličine su npr. broj biljaka, nivo biljne mase, zdravstveno stanje useva, nivo zakorovljenosti i sl. Dodatno, za nas je od posebnog značaja merenje veličine, odnosno intenziteta pojava, koje rezultira određivanjem kategorija ili klasifikacijom (npr. prepoznavanje biljaka u odnosu na zemljište, prepoznavanje obolelih od zdravih biljaka, i sl.). Drugim rečima, intezitet (merenih) pojava u nastavku

možemo posmatrati kao klase, odnosno dodeljene kategorija, entiteta. Time proces i cilj merenja možemo nazvati klasifikacijom.

Iz statističkog ugla, problem klasifikacije nastaje kada istraživač napravi potreban broj merenja i želi da klasifikuje individue (entitete) u jednu od nekoliko kategorija na bazi ovih mera. Istraživač ne može da identifikuje individue sa kategorijama direktno već mora da koristi dobijene rezultate merenja. U mnogim slučajevima postoji konačan broj kategorija ili populacija od kojih su individue potekle i svaka populacija je kategorizovana verovatnoćom raspodele tih mera. Stoga se individue podrazumevaju kao slučajne opservacije ovih populacija (Radojičić, 2001).

Osnovno pitanje je: iz koje populacije potiče navedena individua (entitet) sa određenim merama? Problem klasifikacije može se posmatrati kao problem statističke funkcije odluke. Imamo određen broj hipoteza, a svaka hipoteza je definisana raspodelom opservacija. Mi moramo da prihvatimo jednu od ovih hipoteza i odbacimo ostale. Ukoliko su dve populacije poznate mi imamo elementaran problem testiranja jedne hipoteze specifične raspodele u odnosu na drugu. U nekim okolnostima kategorije su definisane unapred u smislu da je verovatnoća raspodele u potpunosti poznata. Tada prema funkciji gustine, odnosno verovatnoći pripadnosti entiteta kategorijama, možemo ih dodeliti jednoj od kategorija. U ostalim slučajevima forma svake raspodele može biti poznata, ali parametri raspodele su ocenjeni kao primeri iz te populacije. Problem nastaje kada su verovatnoće pripadanja entiteta jednoj ili drugoj kategoriji dosta male i ne zadovoljavaju nivo poverenja koji želimo da zadržimo. Definisanjem metodologije određivanja zone osetljivosti rešavamo problem na način da identifikujemo elemente koji se nalaze u graničnoj oblasti, između dve kategorije, a koju nazivamo *zona osetljivosti* (Radojičić, 2001).

Dakle, kao i u opštem slučaju, kod daljinskog uzrokovanja se postavlja pitanje metoda koji se koristi za određivanje gustina verovatnoća $p(x|\omega)$ (problem procene gustina, npr. pomoću histograma), kako bi se potom primenila Bajesova klasifikacija. Za postizanje veće tačnosti, naročito je od značaja pomenuta zona osetljivosti i način kako da što tačnije klasifikujemo entitete koji u nju upadaju. Ta pitanja i metodi su predmet istraživanja ove disertacije, gde dajemo pretpostavke za metod a zatim ih pokušamo potvrditi (ili odbaciti) kao korisne za postizanje veće tačnosti. To je upravo ono što je definisano polaznim hipotezama ove disertacije, datim u poglavlju 1.1.

6.1 Bajesova procedura i određivanje regiona klasifikacije

Vratimo se prvo na Bajesovu proceduru klasifikacije i metod sličan metodu *Sečenja gustine* (eng. *Density Slicing*) koji smo ranije opisali. Posmatrajmo načine definisanja minimalnih gubitaka kod pogrešne klasifikacije u dva slučaja.

6.1.1 Slučaj kada su priorne verovatnoće poznate

U prvom slučaju pretpostavljamo da su nam priorne verovatnoće kategorija poznate za slučaj dve populacije. Neka verovatnoća opservacije koja dolazi iz populacije P_1 bude q_1, a verovatnoća opservacije koja dolazi iz populacije P_2 bude q_2. Verovatnoća mogućih opcija populacije P_1 su definisane funkcijama raspodele. Tretirajmo slučaj kada raspodela ima određenu gustinu, iako slučaj diskretne verovatnoće dozvoljava isti pristup (Radojičić, 2001).

Neka gustina populacije P_1 bude $p_1(x)$, a gustina populacije P_2 bude $p_2(x)$. Ako imamo region R_1 klasifikacije P_1 verovatnoća tačne klasifikacije opservacije koja u stvari potiče iz populacije P_1 je

$$P(1|1,R)=\int_{R_1} p_1(x)dx$$

gde je $dx=dx_1,\ldots,dx_p$, a verovatnoća pogrešne klasifikacije opservacije koja dolazi iz populacije P_1 je

$$P(2|1,R)=\int_{R_2} p_1(x)dx$$

Verovatnoća tačne klasifikacije za opservaciju koja dolazi iz populacije P_2 je

$$P(2|2,R)=\int_{R_2} p_2(x)dx$$

dok je verovatnoća pogrešne klasifikacije za opservaciju koja dolazi iz populacije P_2 je

$$P(1|2,R)=\int_{R_1} p_2(x)dx$$

Kako je verovatnoća izvlačenja opservacija iz populacije P_1 jednaka q_1, onda je verovatnoća tačne klasifikacije $q_1P(1|1, R)$. Ovo je verovatnoća događaja u gornjem levom uglu tabele 4.3.1. Analogno ovome su i ostala tri slučaja, odnosno ostala tri događaja tabele 4.3.1.

Sada se pitamo koji je prosek očekivanih gubitaka pogrešne klasifikacije? To je suma proizvoda uzroka svake pogrešne klasifikacije pomnožena sa verovatnoćom njihovog pojavljivanja i ona je

$$C(2|1)*P(2|1, R)*q_1 + C(1|2)*P(1|2, R)*q_2. \qquad (6.1.1)$$

To je prosek gubitaka koji želimo da minimiziramo tj. hoćemo da podelimo naš prostor na regione R_1 i R_2 tako da očekivani gubitak bude što je moguće manji. Kao što je ranije rečeno, procedura koja minimizira taj gubitak za dato q_1 i q_2 se zove *Bajesova procedura*. Razmotrimo sada problem izbora takvih regiona R_1 i R_2 da minimiziraju gornji izraz, s tim da imamo poznate prioritetne kategorije (Radojičić, 2001).

Uslovna verovatnoća koja dolazi iz populacije P_1 date opservacije x je

$$\frac{q_1 p_1(x)}{q_1 p_1(x)+q_2 p_2(x)}$$

Ako pretpostavimo da je C(1|2)=C(2|1)=1 tada je očekivani gubitak

$$q_1 \int_{R_2} p_1(x)dx + q_2 \int_{R_1} p_2(x)dx \qquad (6.1.2)$$

Ovo je ujedno i verovatnoća pogrešne klasifikacije, koju je potrebno minimizirati. Za datu tačku (opservaciju) x minimiziramo verovatnoću pogrešne klasifikacije dodeljivanjem populacije koja ima višu uslovnu verovatnoću. Ako je

$$\frac{q_1 p_1(x)}{q_1 p_1(x)+q_2 p_2(x)} \geq \frac{q_2 p_2(x)}{q_1 p_1(x)+q_2 p_2(x)}$$

tada biramo populaciju P_1, u suprotnom biramo populaciju P_2. Time smo minimizirali verovatnoću pogrešne klasifikacije u svakoj tački, odnosno minimizirali smo je na čitavom prostoru. Odatle potiče pravilo

$$\begin{aligned} R_1 &: q_1 p_1(x) \geq q_2 p_2(x) \\ R_2 &: q_1 p_1(x) < q_2 p_2(x) \end{aligned} \qquad (6.1.3)$$

Ako je $q_1p_1(x)=q_2p_2(x)$ tačka može biti klasifikovana ili kao da je iz P_1 ili iz P_2 (u ovom slučaju smo je smestili u region R_1), a ako je $q_1p_1(x)+q_2p_2(x)=0$ za datu tačku x, tačka može biti u bilo kom regionu R_1 ili R_2. Sada se primećuje da je matematički problem koji moramo rešiti: za date nenegativne konstante q_1 i q_2 i nenegativne funkcije $p_1(x)$ i $p_2(x)$, izabrati regione R_1 i R_2 tako da se minimizira izraz (6.1.2), a rešenje je izraz (6.1.3) (Radojičić, 2001).

Ukoliko želimo da minimiziramo (6.1.1) tada možemo pisati

$$[C(2|1)q_1]\int_{R_2} p_1(x)dx + [C(1|2)q_2]\int_{R_1} p_2(x)dx$$

a biramo regione R_1 i R_2 prema

$$\begin{aligned} R_1 &: [C(2|1)q_1]p_1(x) \geq [C(1|2)q_2]p_2(x) \\ R_2 &: [C(2|1)q_1]p_1(x) < [C(1|2)q_2]p_2(x) \end{aligned}$$

s tim da $[C(2|1)q_1]$ i $[C(1|2)q_2]$ su nenegativne konstante. Drugačiji način zapisa je

$$R_1 : \frac{p_1(x)}{p_2(x)} \geq \frac{C(1|2)q_2}{C(2|1)q_1}$$
$$R_2 : \frac{p_1(x)}{p_2(x)} < \frac{C(1|2)q_2}{C(2|1)q_1} \quad . \qquad (6.1.4)$$

Dakle, ukoliko su q_1 i q_2 prioritetne verovatnoće opservacija iz populacije P_1 sa gustinom $p_1(x)$ i populacije P_2 sa gustinom $p_2(x)$ respektivno, i ukoliko je "cena" ili „trošak“ pogrešne klasifikacije opservacije iz populacije P_1 kao da je iz populacije P_2 C(2|1), a opservacije koja iz populacije P_2 kao da je iz populacije P_1 C(1|2), tada regioni klasifikacije R_1 i R_2 definisani pod (6.1.4) minimiziraju očekivani gubitak (Radojičić, 2001).

6.1.2 Slučaj kada su priorne verovatnoće nepoznate

Drugi slučaj koji ćemo razmatrati je slučaj kada nema prioritetnih verovatnoća kategorija. U ovom slučaju očekivani gubitak ukoliko opservacija potiče iz P_1 je

$$C(2|1)\bullet P(2|1, R) = r(1, R)$$

a za opservaciju koja potiče iz P_2 je

$$C(1|2)\bullet P(1|2, R) = r(2, R).$$

Generalno, ne znamo da li je neka opservacija iz populacije P_1 ili P_2. Procedura R je skoro toliko dobra kao procedura R* ukoliko

$$r(1, R)<=r(1, R^*) \text{ i } r(2, R)<=r(2, R^*);$$

R je bolje od R* ukoliko je bar jedna od ovih nejednakosti stroga (bez znaka jednakosti). Obično ne postoji procedura koja je bolja od svih ostalih, ili bar toliko dobra kao i sve ostale procedure. Procedura R je usvojiva ukoliko ne postoji ni jedna procedura bolja od R, pa ćemo se zanimati za celokupnu klasu usvojivih procedura. Pod određenim uslovima ova klasa je isto što i klasa *Bajesovih procedura*.

Klasa procedure je kompletna ukoliko za svaku proceduru van klase postoji jedna klasa koja je bolja. Klasa se naziva esencijalno kompletnom ukoliko za svaku proceduru van klase u klasi postoji bar jedna procedura koja je toliko dobra. Pod određenim uslovima može se pokazati da su dve klase minimalno kompletne tj. mogu se podrazumevati procedure jednakim ukoliko se jedino razlikuju na skupu nula verovatnoće. Princip koji obično vodi ka jedinstvenim procedurama je *min-max* princip. Procedura je min-max ako je maksimum očekivanih gubitaka minimalan. Sa tradocionalne tačke gledišta ovo se može podrazumevati kao optimum procedure (Radojičić, 2001).

U mnogim slučajevima klasifikacije statističari ne mogu da dodele priorne verovatnoće populacijama. U tim slučajevima tražimo klase za usvojive procedure, odnosno prvo treba pokazati da su Bajesove procedure usvojive. Neka je Bajesova procedura R=(R_1, R_2) za dato q_1 i q_2. Pitanje je da li postoji procedura R*=(R_1*, R_2*) takva da je P(1|2, R*)<=P(1|2, R) i P(2|1, R*)<=P(2|1, R) sa barem jednom strogom nejednakošću? Kako je R Bajesova procedura onda je

$$q_1 P(2|1,R) + q_2 P(1|2,R) \le q_1 P(2|1,R^*) + q_2 P(1|2,R^*)$$

a ta nejednakost može biti napisana kao

$$q_1\left[P(2|1,R) - P(2|1,R^*)\right] \le q_2\left[P(1|2,R^*) - P(1|2,R)\right]$$

Može se pokazati da ako je P{p_2(x)=0|P_1}=0 i P{p_1(x)=0|P_2}=0 tada je svaka Bajesova procedura usvojiva (Radojičić, 2001). Takođe, može se dokazati razmatranje pod kojim podrazumevamo da je svaka usvojiva procedura Bajesova procedura – ukoliko važi:

$$P\left\{\frac{p_1(x)}{p_2(x)} = k|P_i\right\} = 0; \quad i = 1,2; \quad 0 \le k \le \infty$$

tada je svaka usvojiva procedura Bajesova procedura i tada je klasa Bajesovih procedura minimalno kompletna (Radojičić, 2001). Neka važi P(i | j,q_1) = P(i | j, R), gde je R Bajesova procedura koja odgovara q_1. P(i | j, q_1) je neprekidna funkcija po q_1. P(2|1, q_1) varira od 1 do 0, kada q_1 ide od 0 do 1. P(1|2, q_1) varira od 0 do 1, tako da postoji vrednost q_1, kažemo q_1*, takvo da važi P(2|1, q_1*)=P(1|2, q_1*) i to je *min-max* rešenje. U slučaju da postoji neka druga procedura R* takva da je max{P(2|1, R*), P(1|2, R*)}< =P(2|1, q_1*)=P(1|2, q_1*) ovo će biti kontradiktorno sa činjenicom da je svako Bajesovo rešenje usvojivo (Radojičić, 2001).

6.1.3 Generalizacija problema klasifikacije

Posmatraćemo problem klasifikacije u jednoj od nekoliko populacija. Razmatranje možemo proširiti na slučaj kad imamo više od dve populacije. Neka P_1,...,P_m bude m populacija sa funkcijama gustine $p_1(x)$,..., $p_m(x)$, respektivno, a regioni se dele na m delova R_1,..., R_m. Ukoliko opservacija pada u region R_i, onda dolazi iz populacije P_i. „Trošak“ pogrešne klasifikacije opservacije, koja je iz populacije P_i gledamo kao da dolazi iz populacije P_j, obeležimo sa C(j|i). Tada je verovatnoća pogrešne klasifikacije

$$P(j|i,R) = \int_{R_j} p_i(x)dx$$

Pretpostavimo da imamo priorne verovatnoće svih populacija q_1,...,q_m. Tada je očekivani gubitak

$$\sum_{i=1}^{m} q_i \left\{ \sum_{j=1}^{m} C(j \mid i) P(j \mid i, R) \right\}$$

Potrebno je izabrati regione $R_1,...,R_m$ tako da gornji izraz bude minimalan (Radojičić, 2001). Kada imamo priorne verovatnoće za populacije, možemo da definišemo uslovnu verovatnoću opservacije koja potiče iz populacije davanjem vrednosti komponentama vektora X. Uslovna kategorija opservacije koja dolazi iz populacije P_i je

$$\frac{q_i p_i(x)}{\sum_{k=1}^{m} q_k p_k(x)}$$

Ukoliko klasifikujemo opservaciju koja dolazi iz populacije P_j tada očekujemo gubitak

$$\sum_{i=1}^{m} \frac{q_i p_i(x)}{\sum_{k=1}^{m} q_k p_k(x)} C(j \mid i)$$

Minimizirani očekivani gubitak na ovoj tački ukoliko izaberemo *j* tako da minimiziramo taj izraz, odnosno

$$\sum_{i=1}^{m} q_i p_i(x) C(j \mid i)$$

Za svako *i* treba izabrati ono *j* koje daje minimum. Ova procedura dodeljuje tačku *x* jednom od regiona R_j. Posmatrajući ovu proceduru za svako *x* definišemo regione $R_1,...,R_m$. Procedura klasifikovanja tada ima zadatak da klasifikuje opservaciju u populaciju P_j ukoliko one padaju u region R_j (Radojičić, 2001).

Ukoliko je q_i prioritetna verovatnoća dobijanja opservacija iz populacije P_i sa gustinom $p_i(x)$, $i=1,...,m$, tada u regionima klasifikacije $R_1,...,R_m$ koji minimiziraju očekivane vrednosti, dodeljujemo *x* regionu R_k ako

$$\sum_{i=1}^{m} q_i p_i(x) C(k|i) < \sum_{i=1}^{m} q_i p_i(x) C(j|i)$$

za $j,k=1,...,m$ i $j<>k$. Ukoliko je verovatnoća jednakosti izmedju desne i leve strane jednaka nuli za svako *k* i *j* u okviru P_i (za svako *i*) tada je procedura minimiziranja jedinstveno prihvatljiva.

Kako se ovaj metod ponaša kada je $C(j|i)=1$ za svako *i* i *j* ($i<>j$)? Tada dobijamo

$$q_j p_j(x) < q_k p_k(x)$$

za j,k=1,...,m i j<>k. U ovom slučaju tačka x je u regionu R_k, ukoliko je k indeks za koje je $q_i p_i(x)$ maksimum i tada je P_k najverovatnija populacija (Radojičić, 2001).

Ako nemamo priorne verovatnoće onda ne možemo definisati bezuslovne očekivane gubitke za procedure klasifikacije. Uslovni očekivani gubitak ukoliko je opservacija iz P_j je

$$\sum_{j=1}^{m} C(j \mid i) P(j \mid i, R) = r(i, R)$$

Procedura R je isto toliko dobra kao i R* ako vazi $r(i, R) <= r(i, R^*)$, i=1,...,m. Procedura R je bolja ukoliko je bar jedna od nejednakosti stroga, a ista je usvojiva ukoliko nijedna od procedura R* nije bolja. Klasa procedura je kompletna ukoliko za svaku proceduru R izvan klase postoji procedura R* u klasi koja je bolja (Radojičić, 2001).

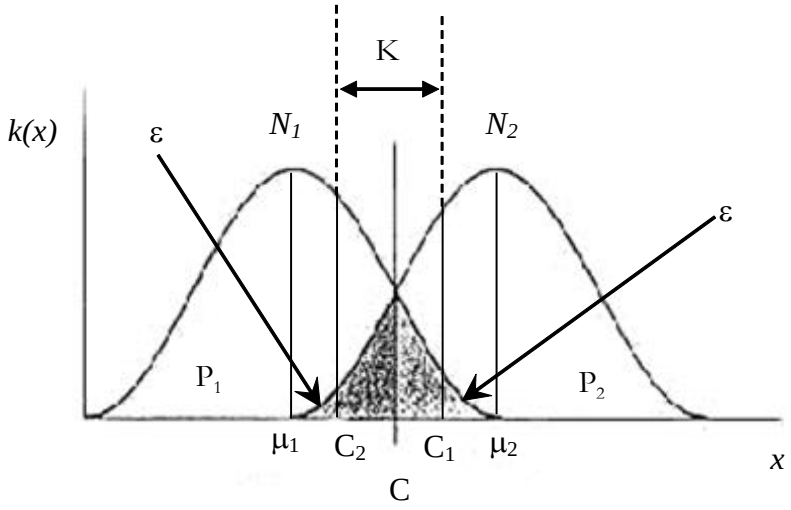
6.2 Određivanje zone osetljivosti i stepena preklapanja

Definišimo sada zonu osetljivosti i prikažimo način računanja za opšte uslove i probleme klasifikacije.

6.2.1 Definicija zone osetljivosti za populacije sa normalnom raspodelom

Neka su date dve populacije koje imaju normalnu raspodelu $N_i(\mu_i, \sigma_i^2)$, (i=1,2) gde su μ_i očekivane vrednosti, a σ_i varijanse prve, odnosno druge populacije. Neka je ε unapred zadata vrednost na osnovu koje ćemo odrediti da li opservaciju svrstavamo u populaciju P_1 ili P_2 ili će ona pripadati oblasti osetljivosti koja nas u ovom trenutku interesuje, pa je želimo odrediti.

Na slici 6.2.1 sve tačke koje se nalaze desno od tačke C_1 kategorizovaćemo u populaciju P_2, sve tačke koje se nalaze levo od tačke C_2 kategorizovaćemo u populaciju P_1, dok će tačke iz intervala (C_1, C_2) pripadati takozvanoj zoni osetljivosti, a sve u zavisnosti od zadate vrednosti ε (Radojičić, 2001).



Slika 6.2.1: Zona osetljivosti dve populacije

Prema onome što je prikazano na slici imamo

$$\int_{-\infty}^{C_1} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx + \int_{C_1}^{+\infty} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx = 1$$

gde je vrednost integrala

$$\int_{C_1}^{+\infty} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx$$

upravo ona koja je data sa ε, odnosno važi

$$\int_{C_1}^{+\infty} \frac{1}{\sigma_1\sqrt{2\pi}} e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx \le \varepsilon$$

Da bi odredili granične tačke zone osetljivosti koristićemo gore navedeni izraz i Gausovu krivu, na osnovu koje je definisana funkcija normalne raspodele. Zato imamo, posle uvođenja smena u integralu, sledeće

$$y = \frac{x-\mu_1}{\sigma_1} \quad \text{i} \quad dy = \frac{1}{\sigma_1} dx$$

$$\frac{1}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{C_1} e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{C_1-\mu_1}{\sigma_1}} e^{-\frac{y^2}{2}} dy = 1-\varepsilon$$

Iz vrednosti funkcije raspodele

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx$$

u našem slučaju imamo

$$\Phi\left(\frac{C_1-\mu_1}{\sigma_1}\right)=1-\varepsilon$$

pa je

$$C_1=z\sigma_1+\mu_1 \text{ gde je } z=\Phi^{-1}(1-\varepsilon)$$

Analogno prethodnom postupku odredićemo i drugu graničnu tačku oblasti osetljivosti tj. tačku C_2 kao

$$C_2=\Phi^{-1}(1-\varepsilon')\sigma_2+\mu_2$$

gde je vrednost $\varepsilon'=\varepsilon$ (Radojičić, 2001).

Dakle određenu tačku mi ćemo svrstati u jedan od tri regiona:

- $K_1=(-\infty,C_2)$ pripada populaciji P_1
- $K_2=(C_{1,}+\infty)$ pripada populaciji P_2
- $K=(C_2,C_1)$ pripada zoni osetljivosti čije su granične tačke date izrazima iz prethodnog postupka.

Time je zona osetljivosti je konačno određena sa

$$\mathrm{K}=\left(\Phi^{-1}(1-\varepsilon)\sigma_2+\mu_2,\Phi^{-1}(1-\varepsilon)\sigma_1+\mu_1\right)$$

U slučaju kada je $C_2<C_1$ tada tu oblast nazivamo *Pozitivna zona osetljivosti*, dok u slučaju kada je $C_2>C_1$ tada tu oblast nazivamo *Negativna zona osetljivosti*. U okviru pozitivne zone osetljivosti regioni K_1 i K_2 za zadato ε ne prepokrivaju oblast osetljivosti K, pa je s te strane ona pozitivnog karaktera, dok u slučaju negativne oblasti osetljivosti regioni K_1 i K_2 za zadato ε prepokrivaju oblast osetljivosti K pa stoga ima negativan karakter tj. povećava mogućnost greške (Radojičić, 2001).

6.2.2 Određivanje stepena preklapanja

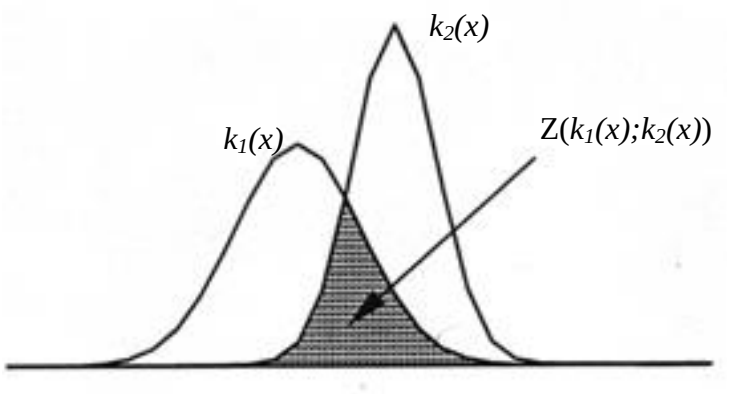
Sledeće pitanje koje se postavlja jeste kako da kvantitativno ocenjujemo poklapanje raspodela. Osnovna ideja u (Radojičić, 2001) je da se odredi udeo preseka površina u ukupnoj površini na datom intervalu (Slika 6.2.1). Označimo ukupnu površinu sa T i presek sa P i izračunajmo ih na sledeći način, ukoliko pretpostavimo:

$$\mathrm{T}_{C_2}^{C_1}(k_1,k_2)=\sum_{x=C_2}^{C_1}\max\left[k_1(x),k_2(x)\right] \text{ za } k_1(x)\geq 0 \text{ i } k_2(x)\geq 0$$

$$\mathrm{P}_{C_2}^{C_1}(k_1,k_2)=\sum_{x=C_2}^{C_1}\min\left[k_1(x),k_2(x)\right] \text{ za } k_1(x)\geq 0 \text{ i } k_2(x)\geq 0$$

Ukoliko znamo ova dva podatka, a njih lako dobijamo ako pređemo sa gornjih suma na odgovarajuće integrale, možemo izračunati i udeo površine P u ukupnoj površini T. Ovako dobijeni podatak nazivamo *ocena poklapanja funkcija* $k_1(x)$ i $k_2(x)$i označićemo je sa Z (Slika 6.2.2):

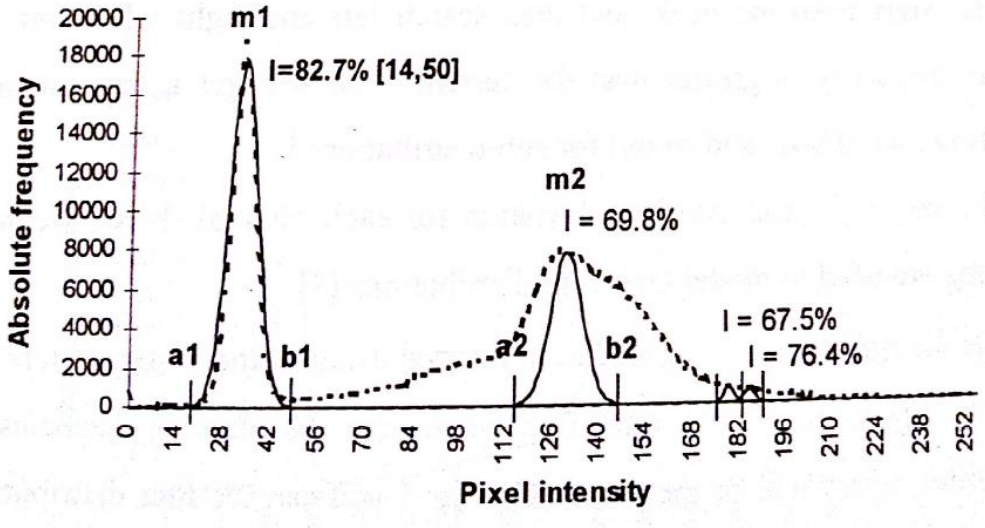
$$Z_a^b(k_1,k_2)=\frac{P_a^b(k_1,k_2)}{T_a^b(k_1,k_2)}$$



Slika 6.2.2: Određivanje stepena preklapanja zone osetljivosti

Ovaj rezultat se može kretati u intervalu od 0 do 1. Nula znači da poklapanja nema, dok jedan znači da je poklapanje potpuno, odnosno da se radi o identičnim funkcijama. Pokazalo se da sva poklapanja sa $Z \approx 0.9$ ili više znače da su raspodele jako slične (Radojičić, 2001).

Jedan primer upotrebe stepena preklapanja se može naći u (Radojičić, Vukmirović, & Glišin, 1997), gde se koristi za identifikaciju onih piksela koji predstavljaju šum na slici, koja se inače koristi za detekciju zvezda na slici noćnog neba. Pretpostavljajući da šum dolazi sa Gausovom raspodelom, autori predlažu metodološki pristup za eliminaciju šuma koji koristi stepen preklapanja za njegovu identifikaciju: prvo se apsolutne frekvencije (pojavljivanja intenziteta piksela) grupišu u intervale intenziteta piksela (kako bi se izbegli mali i loše definisani šiljci), i pronalaze se lokalni minimumi oko svakog šiljka. Zatim se računa srednja vrednost i standardna devijacija svakog šiljka i na osnovu toga se računa Gausova raspodela. Potom se takva raspodela modela testira na preklapanje za uzorkom pomoću gornje formule za T, P i Z. Za diskretne raspodele Z (ili I) je „identitet“, P je presek, i T je ukupna površina. Slika 6.2.3 prikazuje rezultate iz primera, postoje četiri intervala sa četiri odgovarajuća modela. Zatim se nalazi interval sa najvećim procentom preklapanja (na osnovu testiranja, procenjeno je da preklapanje od 80% znači da je pronađen šum na slici) i taj interval se smatra šumom (jer odgovara u najvećoj meri Gausovoj raspodeli) i svi pikseli koji mu pripadaju se zamenjuju pozadinskom bojom, čime se šum eliminiše (Radojičić, Vukmirović, & Glišin, 1997).



Slika 6.2.3: Krive Gausove raspodele dodeljene kandidatima šuma, uključujući procenat preklapanja modela i uzorka.

6.2.3 Klasifikacija u jednu od više populacija sa normlanom raspodelom

Analogno prethodnom postupku kada smo imali primer klasifikacije u jednu od dve populacije sa normalnom raspodelom, možemo razmatranje proširiti na slučaj za više populacija. Dakle, pokušaćemo da rezultate i zaključke koji su važili za slučaj od dve populacije generalizujemo za opšti slučaj od *m* populacija.

Smatraćemo poznatim očekivane vrednosti i varijanse za svih *m* populacija, odnosno, raspodele za svih *m* populacija $\mathrm{N}_i\left(\mu_i,\sigma_i^2\right)$. Konkretno, sada ćemo vršiti klasifikaciju u jedan od 2*m*-1 regiona, gde ćemo sa sigurnošću moći da tvrdimo da se tačka nalazi u tačno jednoj od *m* populacija, ili će pak ona pripadati jednoj od *m*-1 zona osetljivosti, koliko ih ima za slučaj *m* populacija, a na način na koji smo je definisali u prethodnom poglavlju. Sama zona osetljivosti, odnosno njene granične tačke, kojima je ona određena zavisiće od parametra ε koji ima isto značenje kao i u prethodnom postupku za slučaj od dve populacije (Radojičić, 2001).

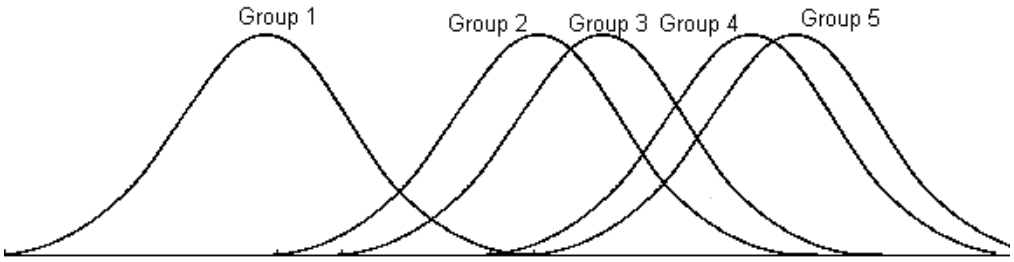
Granične tačke za prvu zonu osetljivosti su iste kao i za slučaj dve populacije, a njihove vrednosti za neku *i*-tu zonu osetljivosti su:

- $C_d(i)=z\sigma_i+\mu_i$ - desna granična tačka *i*-te zone osetljivost
- $C_d(i)=z\sigma_{i+1}+\mu_i$ - leva granična tačka *i*-te zone osetljivosti

gde je $z=\Phi^{-1}(1-\varepsilon)$, a Φ je funkcija raspodele oblika

$$\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{x^2}{2}}dx\,.$$

Stepen preklapanja populacija i i i+1(i=1, m-1) određen je na isti način kao i u slučaju dve populacije. Dakle, i u ovom opštem slučaju, stepen preklapanja je dat kao odnos preseka površina i-te i i+1-ve populacije i ukupne površine ove dve populacije, a sve ovo na datom intervalu (Slika 6.2.4) (Radojičić, 2001).



Slika 6.2.4: Klasifikacija u jednu od više populacija.

6.3 Određivanje i primena zone osetljivosti u daljinskom uzorkovanju

Na osnovu prethodno iznetog teoretskog osnova, u ovom poglavlju se predlaže i opisuje statistički metod određivanja i primene zone osetljivosti u daljinskom uzorkovanju. Metod možemo nazvati „Dvofazna klasifikacija daljinski uzorkovanih podataka primenom zone osetljivosti“.

Zadatak i problem koji želimo rešiti je razdvajanje (klasifikacija) podataka (piksela) u dve klase – usev i pozadinu (namerno se koristi termin „pozadina“ umesto „zemljište“, jer „pozadina“ može sadržati i ostale objekte koji nisu zemljište, tj. razdvajaćemo piksele na usev i „ostalo“). Time uprošćavamo model na klasifikaciju u dve kategorije, pri čemu ćemo poistovetiti spektralne klase i klase informacija (dve klase u oba slučaja, sa 1:1 preslikavanjem), ali se model lako proširuje na klasifikaciju u više klasa itd. Metod sadrži niže opisane korake i elemente kojim se definiše.

6.3.1 Ulazni podaci

Polazni podaci su dati u vidu daljinski uzorkovane slike – poljoprivredne površine gde želimo razlikovati usev od pozadine. Možemo pretpostaviti opšti slučaj da se slika sastoji od 3 kanala (RGB) ali se model lako uopštava i u slučaju većeg (ili manjeg) broja kanala i svetlosnih opsega sadržanih u podacima slike. Podaci se mogu predstaviti u vidu matrice koja ima 2+broj kanala dimenzija (prva i druga dimenzija određuju x i y poziciju svakog piksela, a piksel je definisan sa d vrednosti, zavisno od dubine slike – broja kanala). U primerima koji su korišćeni, svaka od d vrednosti je definisana sa 1 bajtom, bez predznaka, te može uzimati 256 (0-255) različitih vrednosti.

Iako su ulazni podaci trokanalni (RGB), tokom istraživanja i opisa metodologije podaci su prvo konvertovani u HSV ili HSL ekvivalent, a potom korišćena samo dva kanala (HS, odnosno *hue* i *saturation*, ranije pomenuti), iz prostog praktičnog razloga mogućnosti grafičke reprezentacije rezultata-histograma: podaci svedeni u dve dimenzije, sa frekvencijom-histogramom kao trećom, se mogu grafički prikazati 3D dijagramom, dok bi bio nemoguć prikaz ukupno 4 dimenzije. Time se spektralna rezolucija primenjenog metoda svela sa, u slučaju tro-kanalne slike, 256^3 (približno 16.8 miliona) na HS rezoluciju od 180x256 (približno 46 hiljada). Time je umanjena tačnost klasifikacije koju je moguće postići, međutim predloženi model funkcioniše identično i u slučaju sa 3 (i više) dimenzija-kanala ulaznih ili trening podataka (jedino što podatke od značaja ne možemo pregledno grafički prikazati).

6.3.2 Predprocesiranje

Predprocesiranje sprovodimo kako bismo otklonili šum u ulaznim podacima, odnosno kako bismo smanjili nivo (nepotrebnih) detalja, odnosno naglasili grupe podataka od značaja. Ovo postižemo ranije pomenutim *low-pass* filtriranjem, naročito

tehnikom zamućivanja (eng. *bluring*). Jedna od najpopularnijih tehnika takvog filtriranja (zamućivanja) je Gausovo zamućivanje (eng. *Gaussian blur*) čija se primena ovde i predlaže. Gausovo zamućivanje koristi Gausovu funkciju, za normalnu raspodelu, za transofrmaciju vrednosti svakog piksela:

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

U praksi, pri procesiranju slika, kalkulacija se postiže primenom jezgra ili konvolucione matrice koja se primenjuje za računanje novih vrednosti piksela. Aproksimacija Gausovog zamućivanja, za jezgro veličine 5x5 piksela, se postiže primenom sledećeg jezgra:

$$\frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$$

U istraživanju i testiranju tokom izrade ove disertacije se pokazalo da se dobri rezultati postižu Gausovim zamućivanjem jezgra veličine 5x5 i 15x15, te se isto može predložiti kao sastavni deo i korak predložene metodologije koji pospešuje postignutu tačnost. Ovde treba biti pažljiv jer zamućivanje može dovesti do gubitka informacije u slučaju veoma malih objekata-fenomena na slici, pa se kod tih slučajeva preporučuje upotreba jezgra manjih dimenzija (ili čak nastavak obrade bez zamućivanja). Pored toga, sprovedeni su testovi korišćenjem algoritma Otklanjanja šuma nelokalnim sredinama (eng. *Non-local Means Denoising*), ali se Gausovo zamućivanje pokazalo kao dovoljno za predmetnu svrhu.

6.3.3 Trening podaci (uzorci)

U ovom koraku je potrebno definisati trening set (podatke), odnosno odabrati reprezentativne piksele dveju klasa koje želimo identifikovati (usev i pozadina). Potrebno je da analitičar identifikuje reprezentativne segmente tih oblasti na originalnoj slici i da zabeleži delove tih oblasti u vidu maske, odnosno dve maske, jedna koja označava trening podatke useva, a druga koja označava trening podatke pozadine. Maska je jednokanalna slika istih prostornih dimenzija, širine i visine, kao originalna slika, gde su belom bojom (vrednost 255) označeni pikseli koji predstavljaju oblast od interesa, a crnom bojom (vrednost 0) označeni maskirani, odnosno pikseli koji nisu od značaja. Tokom izrade ove disertacije, razvijeno je softversko rešenje (opisano Prilogu A) koje uključuje i mogućnost brze selekcije poligona na originalnoj fotografiji i njihovo čuvanje u datoj formi maske.

Nastavak predloženog metoda se zasniva na činjenici da izabrane oblasti (posebno usev, posebno pozadina) su reprezentativni uzorci tih klasa na ostatku slike, odnosno da na osnovu takvog spektralnog potpisa, sadržanog u trening podacima, se mogu identifikovati svi pikseli datih klasa koji su dovoljno slični zadatom spektralnom potpisu (boji, odnosno bojama u svi kanalima). Činjenicu da određene klase podataka se mogu identifikovati spektralnim potpisom svojih uzoraka koriste i drugi algoritmi iz literature, poput *Projekcije histograma unazad* (eng. *Histogram backprojection*) (Swain & Ballard, 1992) i *GrabCut* (Rother, Kolmogorov, & Blake, 2004).

6.3.4 Određivanje gustina raspodela i stepena preklapanja

Sad je potrebno metodom histograma odrediti gustine raspodela klasa na osnovu zadatih trening podataka. Za frekvenciju pojavljivanja piksela određene boje u uzorku podataka ćemo smatrati da direktno određuje verovatnoću pojavljivanja takvog piksela u datoj klasi na čitavoj populaciji, odnosno verovatnoću da piksel sa određenom bojom pripada datoj klasi:

$$\hat{p}(x) = \frac{n_j}{\sum_j^N n_j dx}$$

U opštem slučaju skup trening podatak je dovoljno velik i daje dovoljno glatku gustinu, i ne sadrži nagle promene, odnosno vrhove i „rupe“ na histogramu. Međutim, u slučaju kada to nije tako, odnosno kada je jedan od skupova trening podataka dovoljno mali, potrebno je izgladiti ga računanjem gustine metodom jezgra:

$$\hat{p}(x) = \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

gde $K(x)$ data funkcija jezgra. U ovom istraživanju je korišćena Normalna (Gausova) funkcija jezgra čija je analitička forma:

$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

a parametar h (*paramtar širenja* ili *izglađivanja*, nazivan i *protok*, eng. *bandwidth*) je tokom istraživanja uzimao vrednosti 0,25, 0,5, i 0,75.

U slučaju, kako je navedeno, kada tretiramo podatke sa njihova dva kanala (HS), gustine su predstavljene sa dve matrice dimenzija 180x256 (dimenzija vrednosti *Hue* x dimenzija vrednosti *Saturation*), koje možemo nazvati H_{usev} i $H_{pozadina}$, čije su vrednosti na lokaciji (i, j) frekvencije pojavljivanja piksela boje H=i i S=j.

Sledeći uvedeni korak u predloženi metod je normalizacija, odnosno skaliranje, jednog od dva histograma na „veličinu“ drugog. Ovo uvodimo kako bismo eliminisali uticaj veličine uzorka, tj. kako analitičar ne bi morao voditi računa o tome da odredi

uzorke različitih klasa koji su iste ili približno iste veličine, već se može fokusirati na to da obuhvate reprezentativne piksele. Ako je:

$$S_{usev} = \sum_{i=1}^{180} \sum_{j=1}^{256} H_{usev_{i,j}}$$

i analogno $S_{pozadina}$ za pozadinu, onda će za $S_{pozadina} > S_{usev}$ biti:

$$H_{usev-norm_{i,j}} = H_{usev_{i,j}} \cdot \frac{S_{pozadina}}{S_{usev}}$$

i analogno u obrnutom slučaju (jer nije potrebno da skaliramo oba histograma, već je dovoljno da to uradimo na jednom, koji sadrži ukupno manji broj piksela).

Sada po analogiji sa ranije definisanim (u poglavlju 6.2.2) možemo odrediti stepen preklapanja tako dobijenih gustina raspodela:

$$P = \sum_{i=1}^{180} \sum_{j=1}^{256} \min[H_{usev-norm_{i,j}}, H_{pozadina-norm_{i,j}}]$$

$$T = \sum_{i=1}^{180} \sum_{j=1}^{256} \max[H_{usev-norm_{i,j}}, H_{pozadina-norm_{i,j}}]$$

$$Z = \frac{P}{T}$$

čime smo dobili stepen preklapanja Z kao procenat preklapanja spektralnih potpisa uzoraka za usev i pozadinu. Ovo istraživanje je pokazalo da se stepen preklapanja <10% može smatrati kao prihvatljivo razdvojenim uzorkom, dok je poželjno da bude <5%, ili, naravno, što manji.

6.3.5 Određivanje posteriornih verovatnoća Diskriminacionom analizom

U sledećem koraku je potrebno na osnovu trening podataka koji su dati u formi frekvencija u predhodno normalizovanim histogramima odrediti liniju ili ravan podele pomoću diskriminacione analize, odosno posteriorne verovatnoće pripadanja svakog entiteta (svih piksela na našoj slici) jednoj od klasa.

Ovde linija ili ravan podele može biti određena Linearnom diskriminacionom analizom (eng. *Linear Discriminant Analysis*, LDA) ili Kvadratnom diskriminacionom analizom (eng. *Quadratic Discriminant Analysis*, QDA), u kom slučaju podelu definiše kvadratna ravan a ne linearna. Tokom istraživanja se pokazalo da postoje slučajevi u kojima kvadratna ravan vrši mnogo bolju podelu na klase (u slučaju nekonveksnih gustina raspodele pojedinih klasa), te je preporučljivo podatke klasifikovati pomoću QDA, odnosmo kvadratnim klasifikatorom.

Kao što smo ranije videli, i LDA i QDA se izvode iz jednostavnih probabilističkih modela koji modeluju uslovnu, posteriornu, verovatnoću podataka $P(X|y=k)$ za svaku klasu k. Predviđanje se dobija korišćenjem Bajesovog pravila:

$$P(y = k|X) = \frac{P(X|y = k)P(y = k)}{P(X)} = \frac{P(X|y = k)P(y = k)}{\sum_l P(X|y = l) \cdot P(y = l)}$$

i biramo onu klasu k koja maksimizira ovu uslovnu verovatnoću. Konkretnije, kod LDA i QDA, $P(X|y)$ se modeluje kao multivarijantna Gausova raspodela sa gustinom:

$$p(X|y = k) = \frac{1}{(2\pi)^n|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(X - \mu_k)^t \sum_k^{-1} (X - \mu_k)\right)$$

Da bismo koristili ovaj model kao klasifikator, potrebno je prvo da procenimo iz trening podataka prioritetne verovatnoće klasa $P(y=k)$, na osnovu odnosa broja instanci klasa k. Međutim, zbog izjednačavanja-normalizacije histograma mi ovde pretpostavljamo i dobijamo da klase imaju identične prioritetne verovatnoće. Zatim, potrebne su srednje vrednosti klasa μ_k (empirijske srednje vrednosti klasa) i na kraju matrice kovarijanse (pomoću empirijskih matrica kovarijansi uzorka ili regularizovanog estimatora). Na kraju, kod LDA se pretpostavlja da Gausove raspodele za svaku klasu dele iste matrice kovarijanse, što rezultira linearnim ravnima odlučivanja, dok kod QDA nema pretpostavki o matricama kovarijanse Gausovih raspodela, što rezultira kvadratnim ravnima odlučivanja.

6.3.6 Određivanje zone osetljivosti

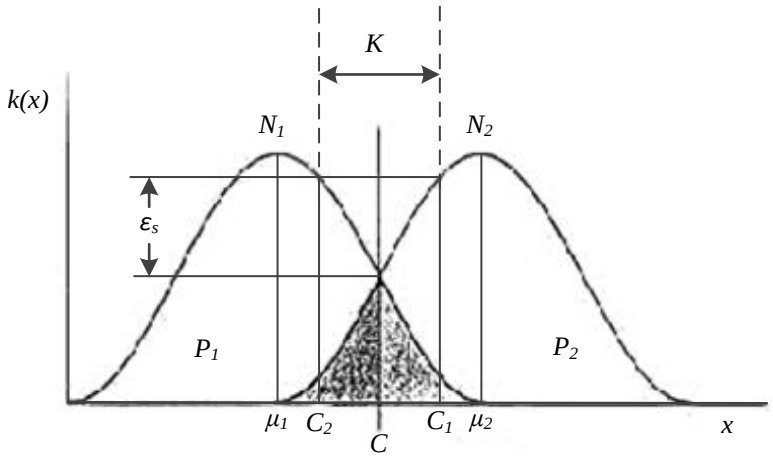
Sada je potrebno da odredimo regione odlučivanja i zonu osetljivosti. Kao što smo rekli u prethodnom poglavlju, samtramo priorne verovatnoće jednakim, a zbog uprošćenja modela možemo smatrati i jednakim „troškove“ pogrešne klasifikacije. Drugim rečima za ranije date regione odlučivanja:

$$R_1 : [C(2|1)q_1]p_1(x) \geq [C(1|2)q_2]p_2(x)$$
$$R_2 : [C(2|1)q_1]p_1(x) < [C(1|2)q_2]p_2(x),$$

mi rešavamo slučaj kada je C(2|1) = C(1||2) (trošak pogrešne klasifikacije je identičan) i kada je $q_1 = q_2 = 0.5$ (priorne verovatnoće su nepoznate, odnosno jednake). Za situacije kada bi ove vrednosti bile različite, odnosno poznate, model se veoma lako uopštava dodavanjem ovih vrednosti. U prethodnim poglavljima dali smo zonu osetljivosti kao:

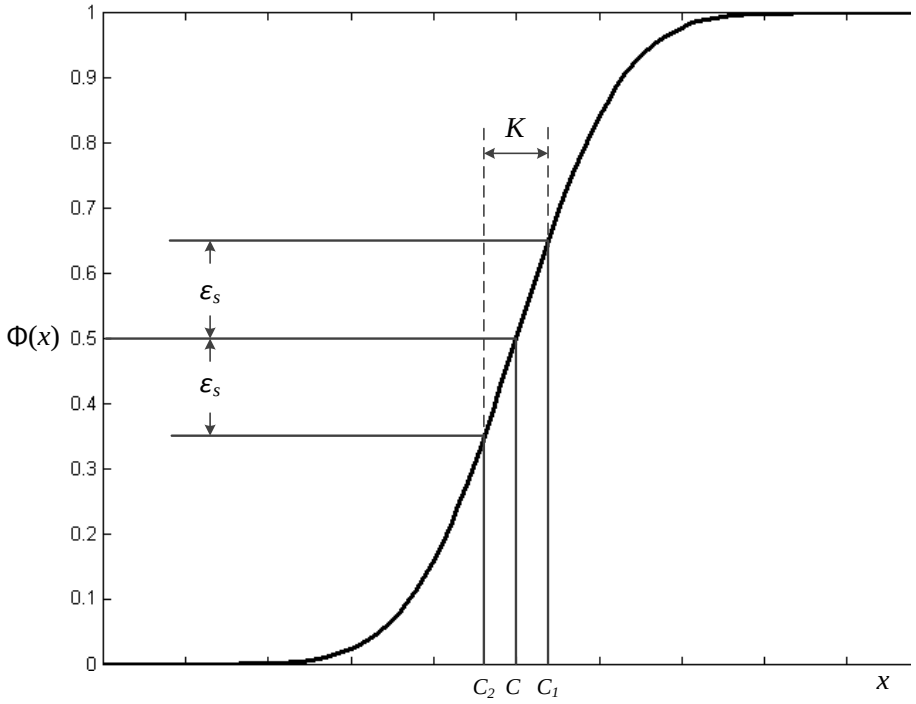
$$\mathrm{K} = \left(\Phi^{-1}(1-\varepsilon)\sigma_2 + \mu_2, \Phi^{-1}(1-\varepsilon)\sigma_1 + \mu_1\right)$$

međutim u ovom metodu ovodimo nešto modifikovan parametar ε, koji možemo nazvati ε_s (s od eng. *sensitivity* – osetljivost) i koji, za razliku od ranijeg parametra, predstavlja vrednost kao na Slici 6.3.1.



Slika 6.3.1: Zona osetljivosti dve populacije određena parametrom ε_s

Pošto su posteriorne verovatnoće izračunate pomoću QDA dobijene u vidu kumulativne Gausove raspodele, gde $\Phi(x)<0,5$ znači pripadnost jednoj a $\Phi(x)>0,5$ znači pripadnost drugoj klasi, pogodnije je zonu osetljivosti prikazati kao na Slici 6.3.2.



Slika 6.3.1: Zona osetljivosti dve populacije određena parametrom ε_s i kumulativnom Gausovom raspodelom

Te sada zonu osetljivosti $K=(C_2, C_1)$ određujemo pomoću:

$$K = (\Phi^{-1}(0,5 - \varepsilon_s),\ \Phi^{-1}(0,5 + \varepsilon_s))$$

gde je važno napomenuti da će u našem i opštem slučaju zona *K* biti višedimenzionalna oblast, npr. 2-dimenzionalna u HS prostoru, 3-dimenzionalna u RGB, itd.

6.3.7 Klasifikacija podataka iz zone osetljivosti

U ovom koraku predlaže se primena zone osetljivosti za povećanje tačnosti klasifikacije. Entitete (piksele) koji upadnu u zonu osetljivosti treba tretirati na poseban način kako bismo obezbedili da se klasifikuju što tačnije. Drugim rečima, sve piksele $x < C_2$ i $x > C_1$ jednostavno klasifikujemo u date klase, veoma brzom metodom diskriminacione analize, dok za piksele $x \in [C_2, C_1]$ (što bi trebalo da je značajno manji broj piksela od ukupnog broja ukoliko je parametar ε_s dobro postavljen) možemo primeniti drugu metodu klasifikacije. Konkretno, u sprovedenom istraživanju za klasifikaciju piksela iz zone osetljivosti korišćena je kNN klasifikacija, i to za RGB, tj. trokanalnu dubinu. Prednost kNN metode je što tretira svaki piksel posebno, tj. ne svrstava ih u ravni odlučivanja i time može postići veliku tačnost pri klasifikaciji. Nedostatak je što je metod veoma spor, upravo iz razloga što posebno tretira svaki piksel. Prema tome, za ovaj manji broj piksela iz zone osetljivosti možemo primeniti sporu ali tačnu metodu klasifikacije i time uspešno povećati ukupnu tačnost klasifikacije pri neznatnom povećanju ukupno utrošenih resursa.

Druge mogućnosti bi bile da iterativno razmatramo zonu osetljivosti sa dodatnim trening setom ili da gustinu raspodele u zoni osetljivosti računamo pomoću više pristupa, ali smo svakako u čitavom postupku lokalizovali problem na manji broj podataka čime se smanjuje kako potreban ručni rad analitičara, tako i računska kompleksnost. Ovde se ostavlja prostor za dalju razradu predloženog metoda.

6.3.8 Merenje tačnosti klasifikacije

Za merenje tačnosti postignute klasifikacije, kako bismo mogli meriti i oceniti postignute rezultate, prvenstveno koristimo mere ukupna tačnost i Kappa statistika:

$$OA = \frac{\sum_{i=1}^{r} x_{ii}}{N}$$

$$\widehat{K} = \frac{N \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} (x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{r} (x_{i+} * x_{+i})}.$$

Ostaje još pitanje u odnosu na šta merimo postignutu tačnost? U sprovedenom istraživanju merena je tačnost u odnosu na klasifikaciju postignutu kNN metodom za čitavu sliku, kao veoma tačne ali veoma spore metode, te u odnosu na date uzorke

(trening podatke), koji se mogu smatrati potvrđenim „*ground truth*“ podacima. Da bi se potvrdila uspešnost predložene metode, po istom principu je poređena sa klasifikacijom postignutom pomoću kNN metode nad HS dubinom, SVM klasifikacijom, kao i rezultatom QDA klasifikacije bez upotrebe zone osetljivosti. Rezultati postignuti nad jednim brojem primera su prikazani u sledećem poglavlju.

7 STUDIJA PRIMERA

U ovom poglavlju su opisane tri studije slučaja za koje je sproveden predloženi metod kako bi se potvrdila njegova uspešnost. Naravno, tokom istraživanja metod je testiran na mnogo većem broju slučajeva, ali zbog sažetosti ovde su prikazana samo tri. Što se tiče parametra osetljivosti, ε_s, testirane su različite vrednosti ali zbog sažetosti ovde prikazujemo rezultate za vrednosti 0,25, 0,35 i 0,40.

7.1 Primer 1 – Ogledna polja uljane repice

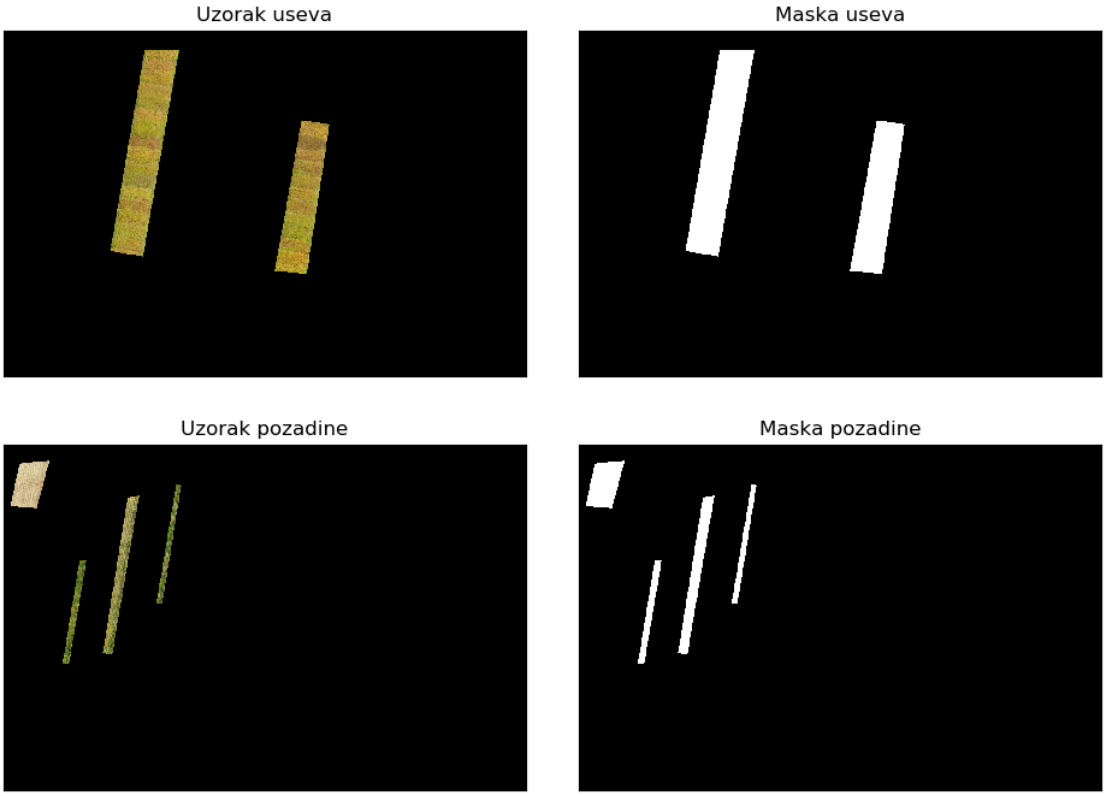
7.1.1 Ulazni podaci i predprocesiranje

Slika korišćena za testiranje je Uljana repica (sa lokacije u Kanadi), pri čemu se ovde radi o oglednom polju koje sadrži veliki broj različitih mikroplotova (malih oglednih lokacija na kojima se seje različiti hibrid, različito tretira, itd.) (Slika 7.1.1). Sprovedeno je predprocesiranje Gausovim zamućivanjem, pri čemu se pokazalo da veličina jezgra 15 daje bolje rezultate od 5, tj. bolje se pokazao veći nivo zamućivanja.

Slika 7.1.1: Originalna slika oglednog polja uljane repice

7.1.2 Trening podaci (uzorci)

Korišćenjem razvijenog softverskog rešenja određeni su trening podaci – oblasti na slici koje predstavljaju uzorak useva i uzorak pozadine (Slika 7.1.2). Tabela 7.1.1 prikazuje detalje veličine slike i uzoraka. Kao što se vidi iz tabele, ovaj test je izvršen sa relativno velikim brojem trening piksela, odnosno velikim zadatim uzorcima za usev i pozadinu.



Slika 7.1.2: Određeni trening podaci useva (gore) i pozadine (dole), sa odgovarajućim maskama

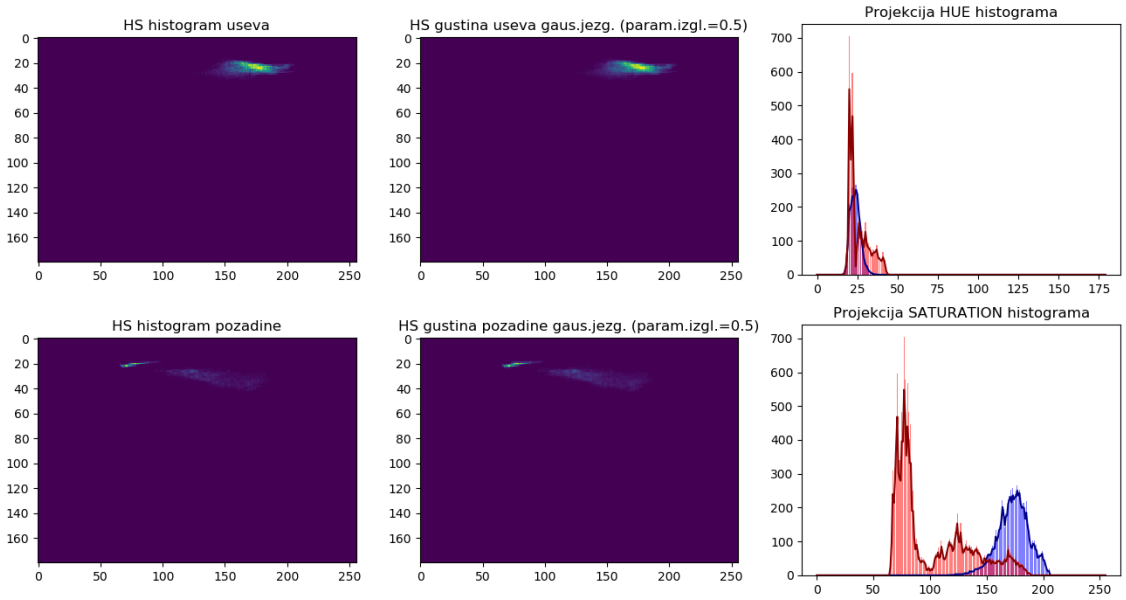
Tabela 7.1.1:Detalji veličine slike i uzoraka (trening podataka)

Visina slike	Širina slike	Ukupno piksela	Pikseli za trening-usev	Pikseli za trening-pozadina
696	1.055	734.280	46.889 (6,39%)	17,228 (2,35%)

7.1.3 Gustine raspodela i stepen preklapanja

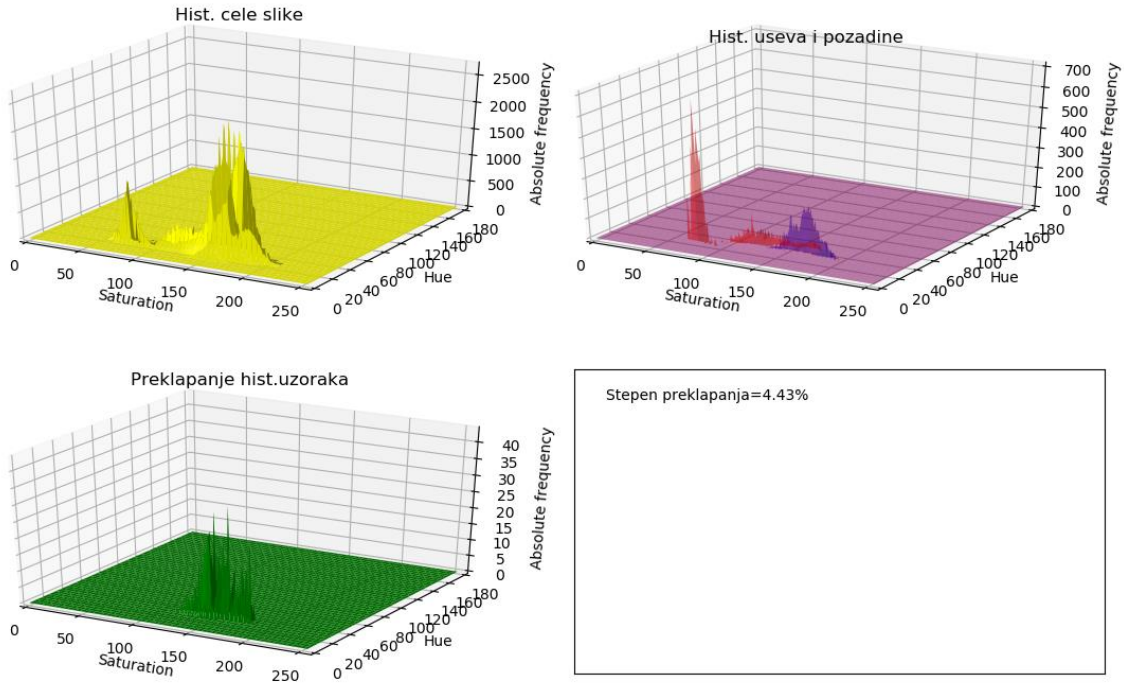
Preama opisanom metodu histogrami trening podataka su normalizovani (ujednačeni). Dobijeni histogrami za dvodimenzionalne podatke (*Hue* i *Saturation*, H i S) su prikazani na Slici 7.1.3, gde su prvo u formi „*heat* mape“ prikazani histogrami

uzoraka, zatim gustine izglađene Gausovim jezgrom (sa parametrom izglađivanja 0,5), i na kraju su prikazane projekcije tih histograma na H i S ravan (crvenom i plavom bojom), gde su svetlijim „bar“ linijama prikazane vrednosti histograma a tamnijom bojom linija izglađene gustine. Na projekcijama histograma se može primetiti da H (odnosno boja) ne razdvaja toliko dobro dva uzorka, koliko to čini S (zasićenost bojom).



Slika 7.1.3: HS histogrami uzorka, izglađene gustine uzorka i projekcije na H i S ravan

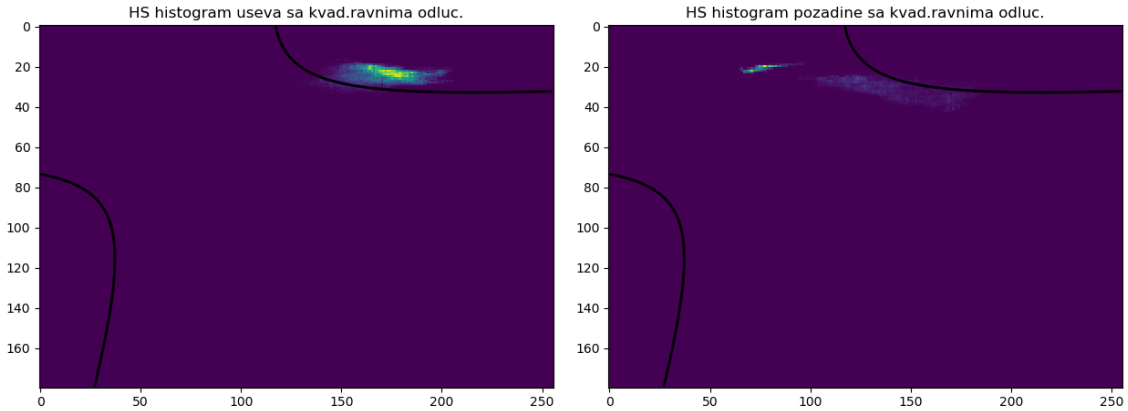
Na slici 7.1.4 je 3D dijagramom prvo prikazan histogram čitave slike, zatim su prikazani histogrami uzorka useva i pozadine (na istom dijagramu), zatim dijagram oblasti preklapanja ta dva histograma. U ovom primeru, Stepen preklapanja iznosi 4,43% (ukupno 3.975 piksela) što daje dovoljno dobro razdvajanje.



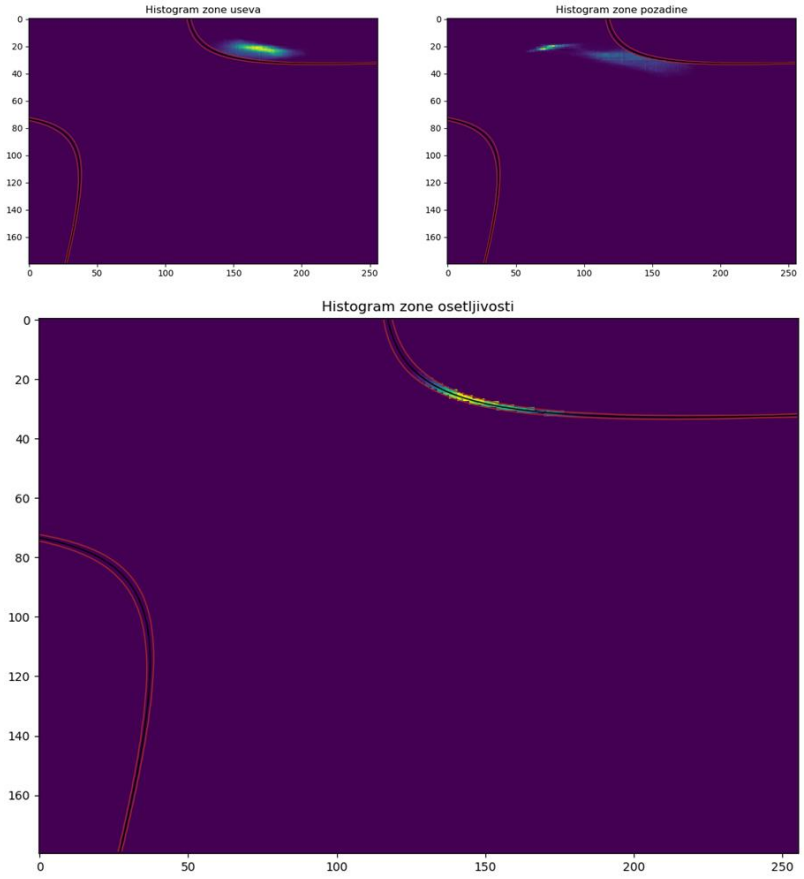
Slika 7.1.4: Prikaz u 3D histograma slike, useva i pozadine i oblasti preklapanja

7.1.4 Diskriminaciona analiza i zona osetljivosti

Primenom Kvadratne diskriminacione analize, na osnovu uzoraka useva i pozadine određene su kvadratne ravni odlučivanja prikazane na Slici 7.1.5 i za ε_s = 0,25 određene su ravni odlučivanja i zona osetljivosti, sa samo onim pikselima koji im pripadaju, kako je dato na slici 7.1.6.

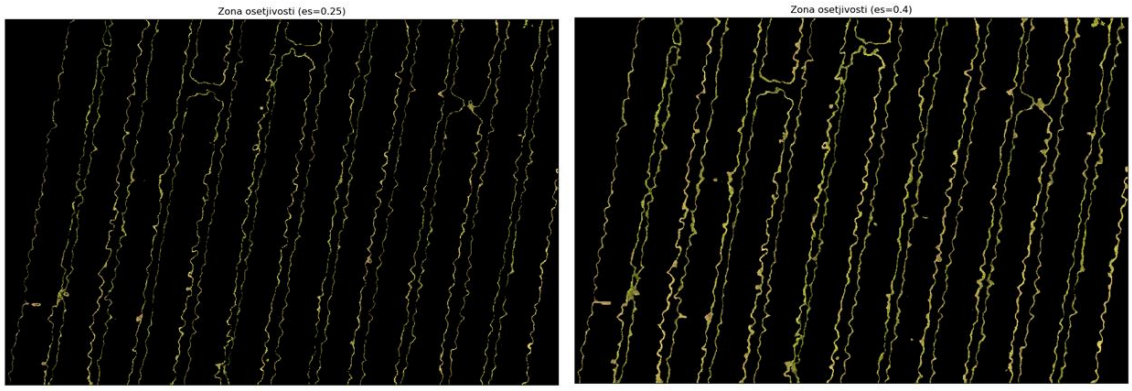


Slika 7.1.5: Ravni odlučivanja dobijene QDA



Slika 7.1.6: Klasifikacija HS vrednosti piksela i zona osetljivosti (ε_s = 0,25)

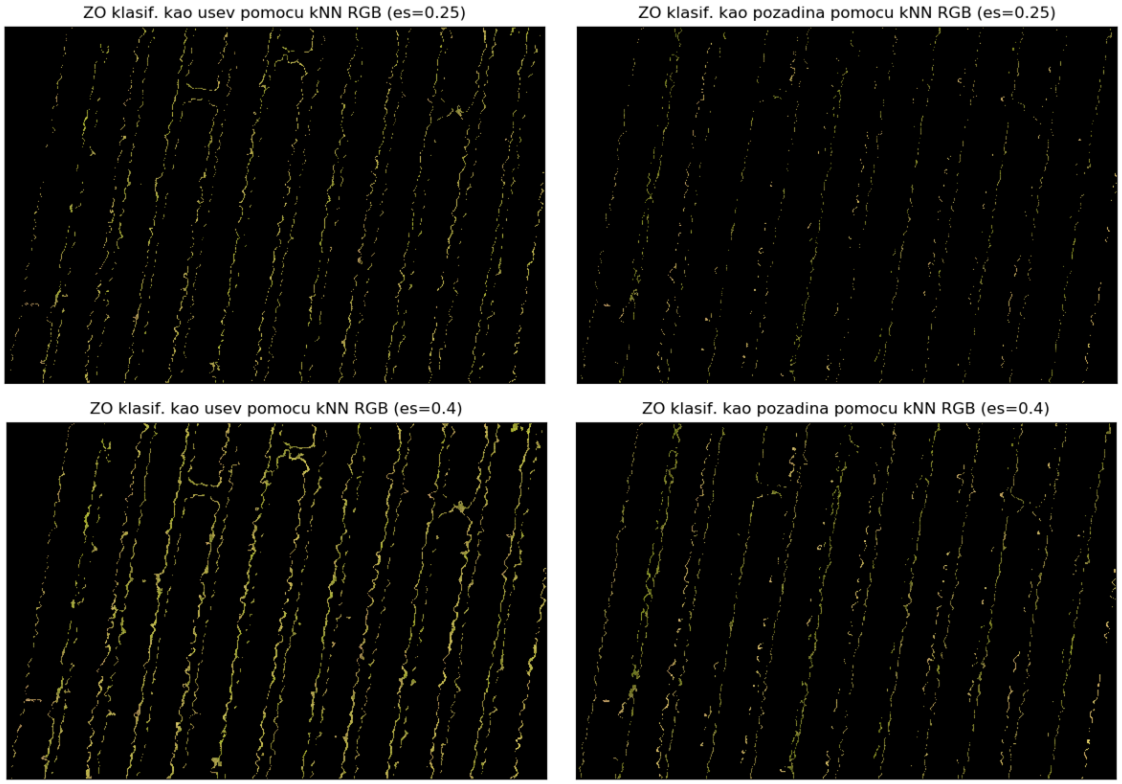
Na osnovu tako određenih ravni odlučivanja i zone osetljivosti na osnovu histograma uzoraka useva i pozadine, pikseli su klsaifikovani kao usev i kao pozadina, a pikseli slike koji upadaju u zonu osetljivosti su prikazani na slici 7.1.7.



Slika 7.1.7: Pikseli iz zone osetljivosti za ε_s = 0,25 i ε_s = 0,40

7.1.5 Klasifikacija podataka iz zone osetljivosti

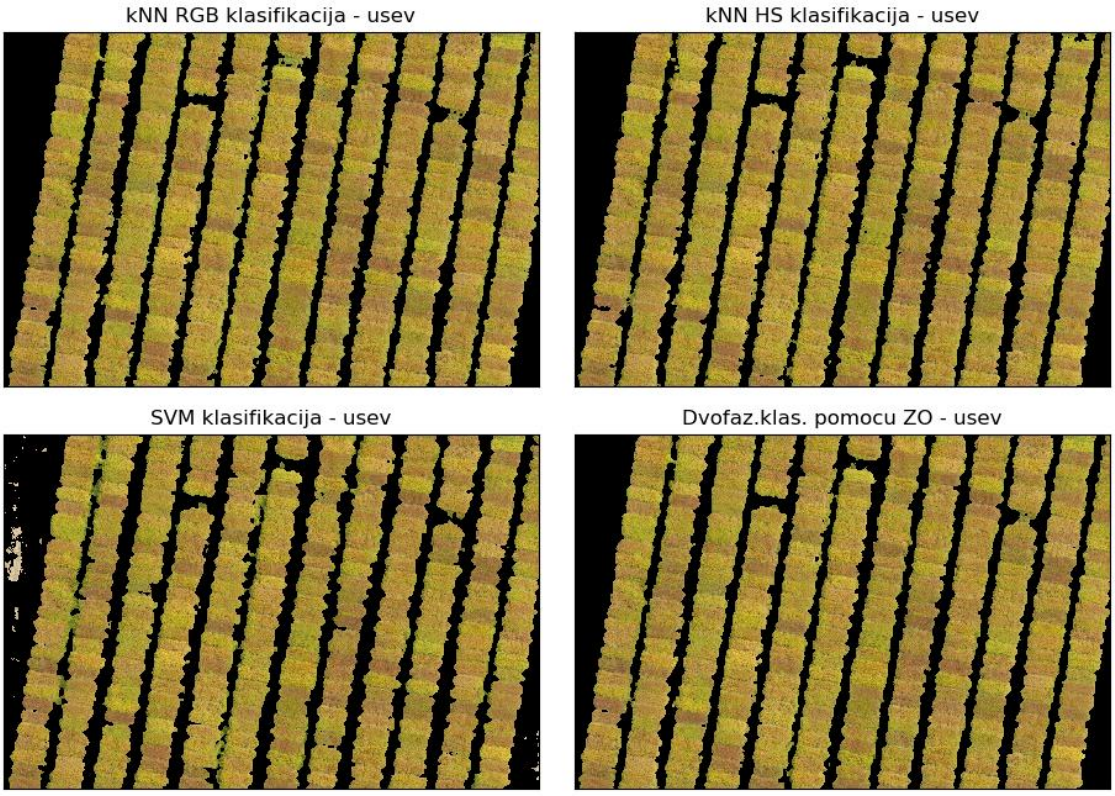
Sada tako određene piksele zone osetljivosti klasifikujemo dodatnom metodom klasifikacije, odnosno u našem slučaju kNN metodom po sva tri kalana. Na Slici 7.1.8 su prikazani posebno pikseli sa Slike 7.1.7 klasifikovani na taj način kao usev i kao pozadina, za oba primera (za ε_s = 0,25 i ε_s = 0,40).



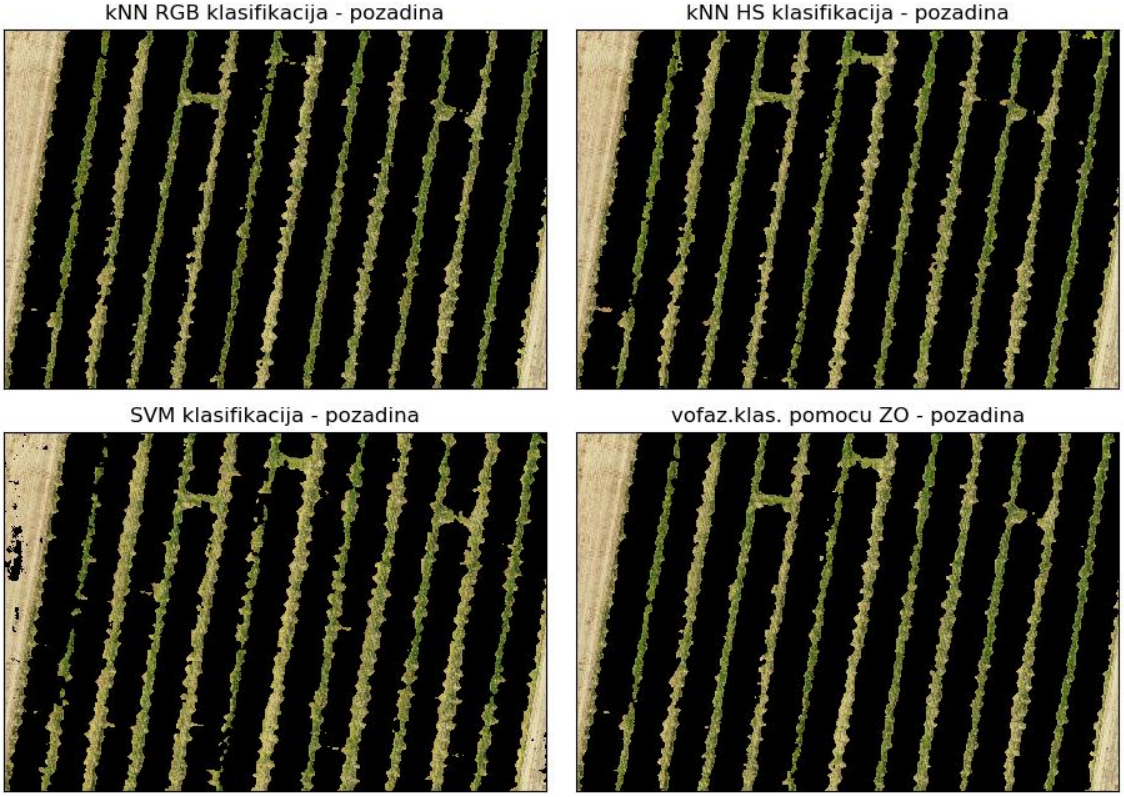
Slika 7.1.8: Pikseli iz zone osetljivosti klasifikovani kao usev i kao pozadina

7.1.6 Merenje tačnosti klasifikacije

Kao što je opisano u 6.3.8, sada je prvo slika klasifikovana na osnovu polaznih trening podataka metodama kNN sa RGB dubinom (sva 3 kanala), kNN sa HS dubinom (sa 2 kanala), metodom SVM i metodom QDA bez zone osetljivosti. Slika 7.1.9 daje uporedni prikaz postignutog izdvajanja klase useva a Slika 7.1.10 prikaz izdvajanja klase pozadine po svim metodama klasifikacije (izuzev QDA bez zone osetljivosti, zbog sažetosti prikaza, ali ona je svakako sastavni deo predloženog metoda sa zonom osetljivosti).



Slika 7.1.9: Uporedni prikaz klasifikacije (useva) pomoću četiri različita metoda



Slika 7.1.10: Uporedni prikaz klasifikacije (pozadina) pomoću četiri različita metoda

Sada je ostalo da prikažemo mere uspešnosti različitih primenjenih metoda klasifikacije. Rečeno je već da je utrošak resursa, pre svega računarskih, od velikog značaja za šta je jasna motivacija kada se uzme u obzir da se metod predlaže za klasifikaciju daljinski uzorkovanih podataka-slika koje su po pravilu veoma velikih dimenzija, odnosno sadrže velike količine podataka (piksela). U primerima koji se ovde navode su korišćene slike malih dimenzija, tako da iako su vremena izvršavanja algoritama za klasifikaciju mala, pa se mogu učiniti neznačajnim, ta vremena značajno rastu sa porastom veličine slike i veoma su važna za uspešnost primene metode klasifikacije. Svi algoritmi su, zbog uporedivosti, izvršavani na istoj računarskoj konfiguraciji pri istom opterećenju sistema ostalim programima. Ta konfiguracija je, informativno, Intel Core i5-3320M CPU 2.60GHz, 8.00 GB RAM, sa Windows 7 Professional 64-bit operativnim sistemom.

Tabela 7.1.2 prikazuje uporedna vremena potrebna za izvršenje klasifikacije ostalim metodama, Tabela 7.1.3, zajedno sa brojem piksela identifikovanim u zoni osetljivost i njihovim delom klasifikovanim kao usev, uporedna vremena klasifikacije pomoću QDA i klasifikacije piksela iz zone osetljivosti kNN metodom (u RGB opsegu), a Tabela 7.1.4 uporedni prikaz Ukupne tačnosti (OA) i Kappa statistike u odnosu na sliku klasifikovanu u celosti sa kNN (RBG) metodom, kao i Ukupnu tačnost postignutu u odnosu na zadati trening set podataka (uzorke useva i pozadine).

Tabela 7.1.2: Uporedna vremena klasifikacije referentnim metodama

Metod klasifikacije	Vreme trajanja (minuti:sekunde)
kNN RGB	04:34
kNN HS	00:42
SVM	00:08
QDA (bez zone osetljivosti)	00:00

Tabela 7.1.3: Vreme i broj piksela klasifikacije sa zonom osetljivosti

Parametar ε_s	Veličina zone osetljivosti	Vreme trajanja (minuti:sekunde)	Pikseli iz ZO klasifikovani kao usev
0,25	27.799 (3,79%)	00:15	18.419 (66,26%)
0,35	44.751 (6,09%)	00:15	29.411 (65,72%)
0,40	57.242 (7,80%)	00:16	36.989 (64,62%)

Tabela 7.1.4: Izmerena tačnosti klasifikacije

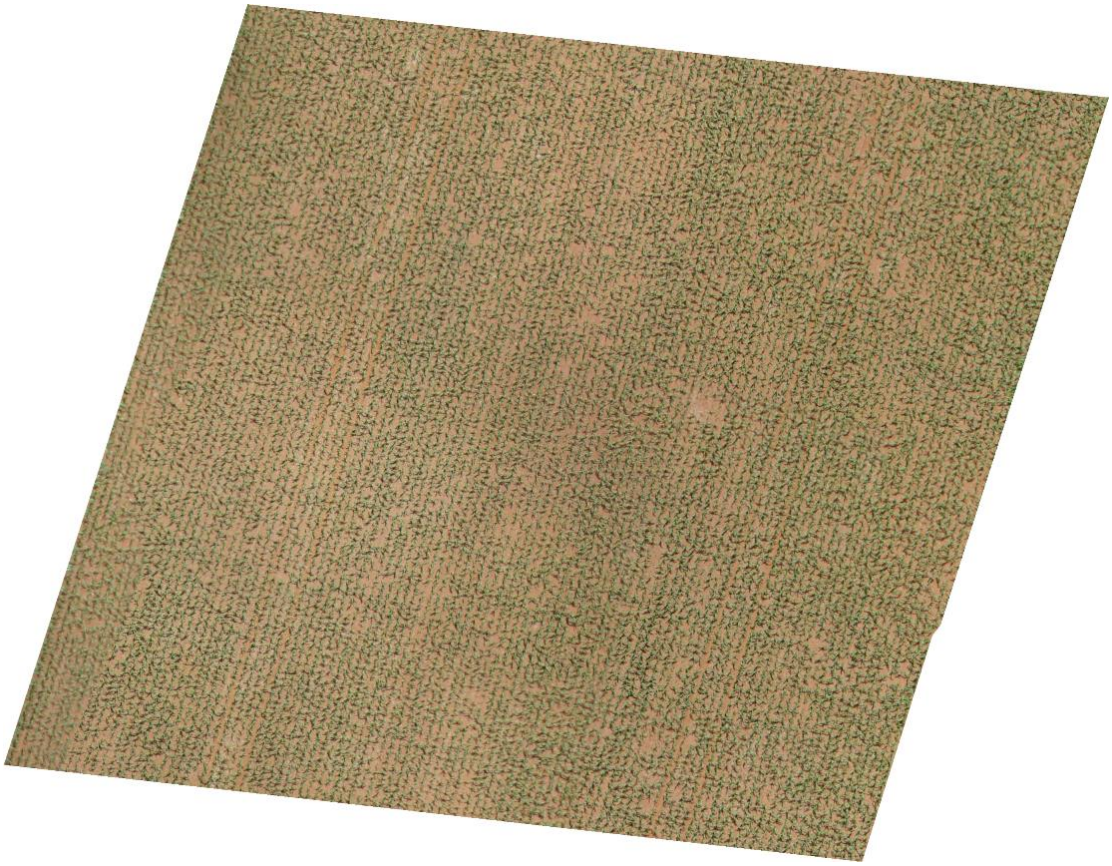
Primenjeni metod	Ukupna tačnost (ref. kNN RGB)	Kappa statistika (ref. kNN RGB)	Ukupna tačnost (ref. trening uzorak)
kNN RGB	100,00%	100,00%	96,53%
kNN HS	96,18%	90,25%	95,09%
SVM	91,70%	79,70%	91,91%
QDA	95,83%	89,49%	95,13%
Sa ZO (ε_s = 0,25)	97,40%	93,39%	95,73%
Sa ZO (ε_s = 0,35)	98,03%	94,98%	96,06%
Sa ZO (ε_s = 0,40)	98,45%	96,04%	96,20%

Na osnovu prikazanih podataka može se zaključiti da predloženi metod Dvofazne klasifikacije korišćenjem zone osetljivosti postiže najveću tačnost u poređenju sa ostalim metodama, uz efikasno vreme izvršavanja.

7.2 Primer 2 – Kukurz u fazi V5

7.2.1 Ulazni podaci i predprocesiranje

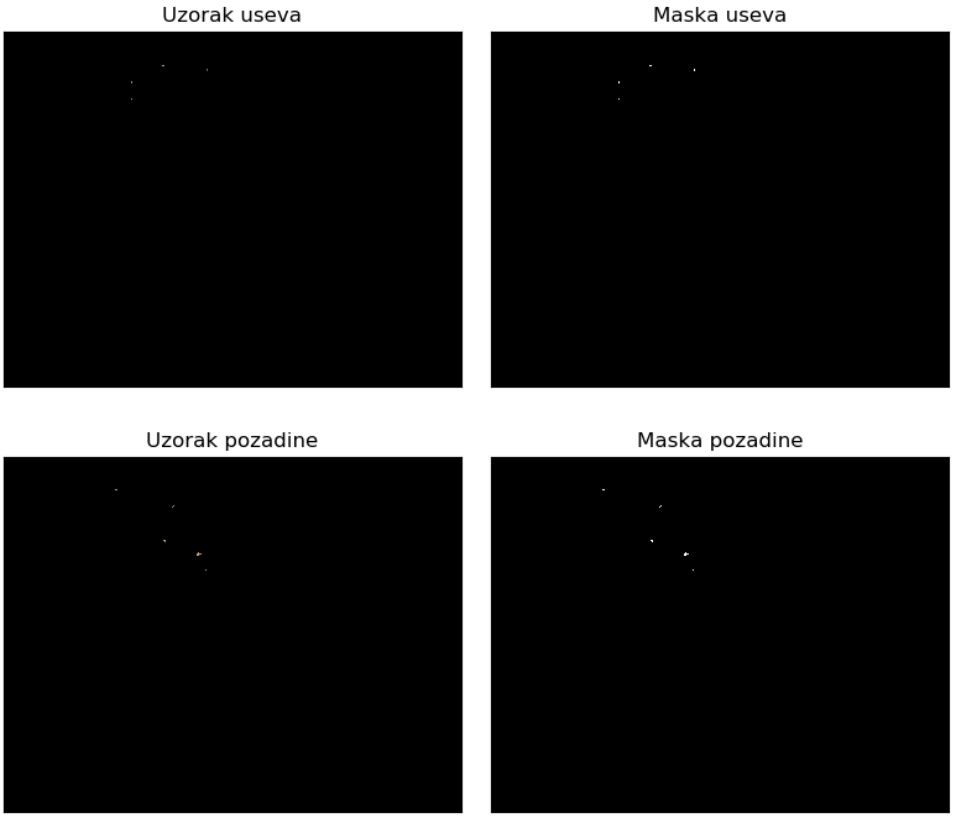
Slika korišćena za testiranje je Kukurzuz (sa lokacije u Srbiji), u fazi V5 (Slika 7.2.1). Sprovedeno je predprocesiranje Gausovim zamućivanjem, pri čemu se pokazalo da veličina jezgra 5 daje bolje rezultate od 15, budući da su biljke kukuruza prilično male na slici te je ipak potrebno zadržati veću oštrinu slike. U ovom primeru želimo testirati metod u slučaju rasutih biljaka, pre nego zgusnutih kao u prethodnom primeru, kao i sa veoma malim skupom trening podataka.



Slika 7.2.1: Originalna slika kukuruza u fazi V5

7.2.2 Trening podaci (uzorci)

I ovde su, korišćenjem razvijenog softverskog rešenja, određeni trening podaci – oblasti na slici koje predstavljaju uzorak useva i uzorak pozadine (Slika 7.2.2). Tabela 7.2.1 prikazuje detalje veličine slike i uzoraka. U ovom slučaju testirana je identifikacija „rasutih“ useva i to sa veoma malim uzorkom – trening setom.



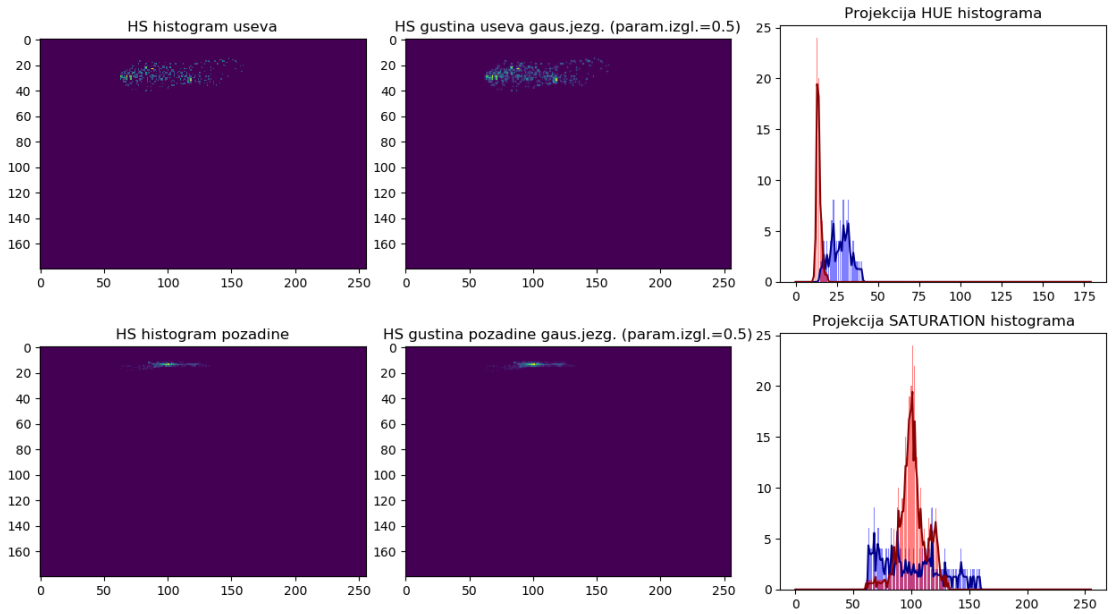
Slika 7.2.2: Određeni trening podaci useva (gore) i pozadine (dole), sa odgovarajućim maskama

Tabela 7.2.1:Detalji veličine slike i uzoraka (trening podataka)

Visina slike	Širina slike	Ukupno piksela	Pikseli za trening-usev	Pikseli za trening-pozadina
2.149	2.772	5.957.028	414 (0,01%)	839 (0,01%)

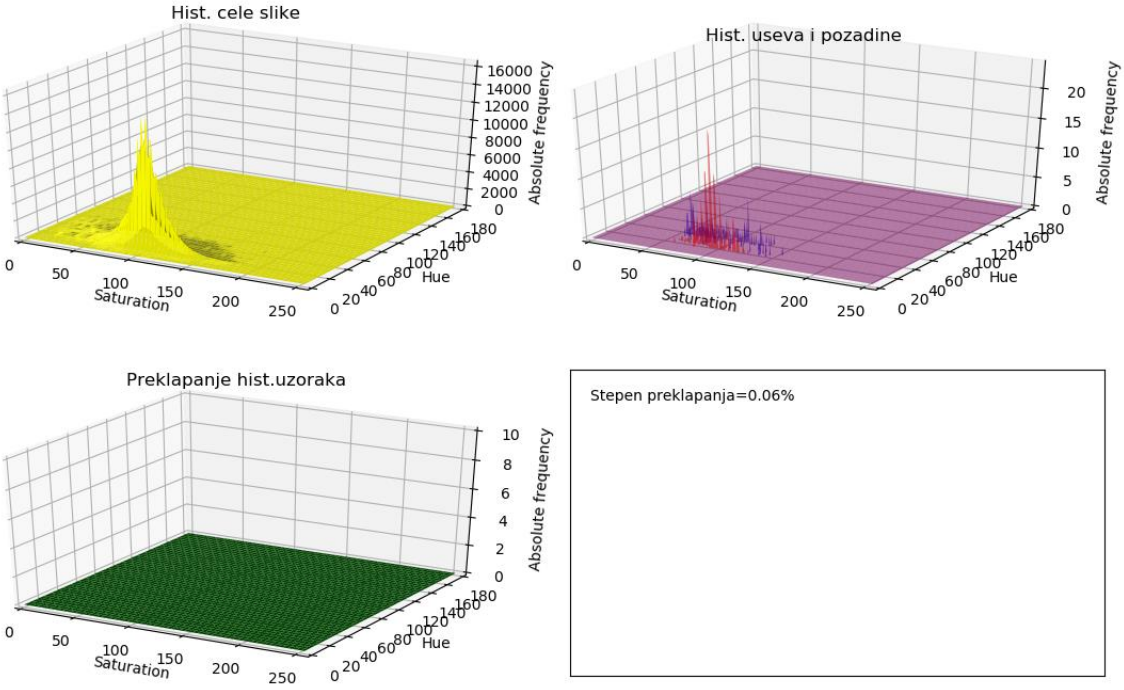
7.2.3 Gustine raspodela i stepen preklapanja

Dobijeni ujednačeni histogrami za dvodimenzionalne podatke (*Hue* i *Saturation*, H i S) su prikazani na Slici 7.2.3, gde su takođe prvo u formi „*heat* mape“ prikazani histogrami uzoraka, zatim gustine izglađene Gausovim jezgrom (sa parametrom izglađivanja 0,5), i na kraju su prikazane projekcije tih histograma na H i S ravan (crvenom i plavom bojom). Boja (Hue) bolje doprinosi razdvajanju nego Saturation.



Slika 7.2.3: HS histogrami uzorka, izglađene gustine uzorka i projekcije na H i S ravan

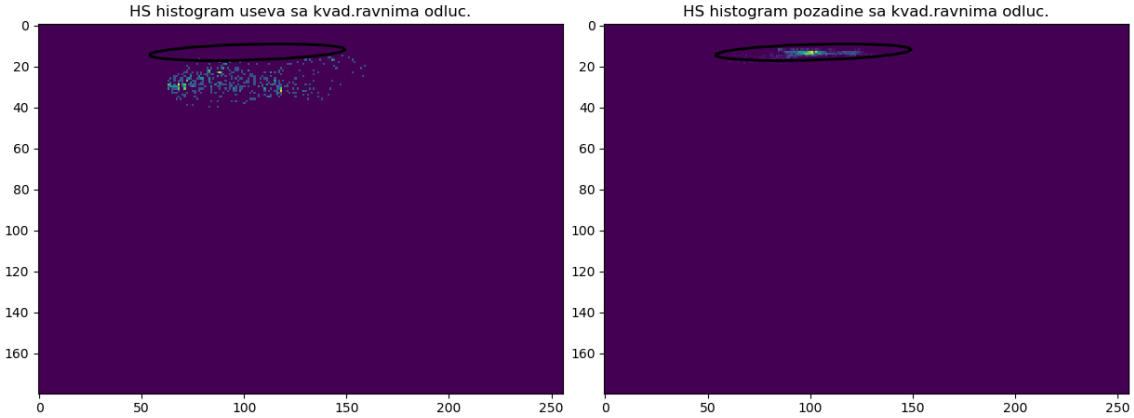
Na slici 7.2.4 je 3D dijagramom je ponovo prvo prikazan histogram čitave slike, zatim su prikazani histogrami uzorka useva i pozadine, dijagram oblasti preklapanja ta dva histograma. U ovom primeru, Stepen preklapanja iznosi 0,06% (samo jedan jedini piskel) što je veoma dobro razdvajanje.



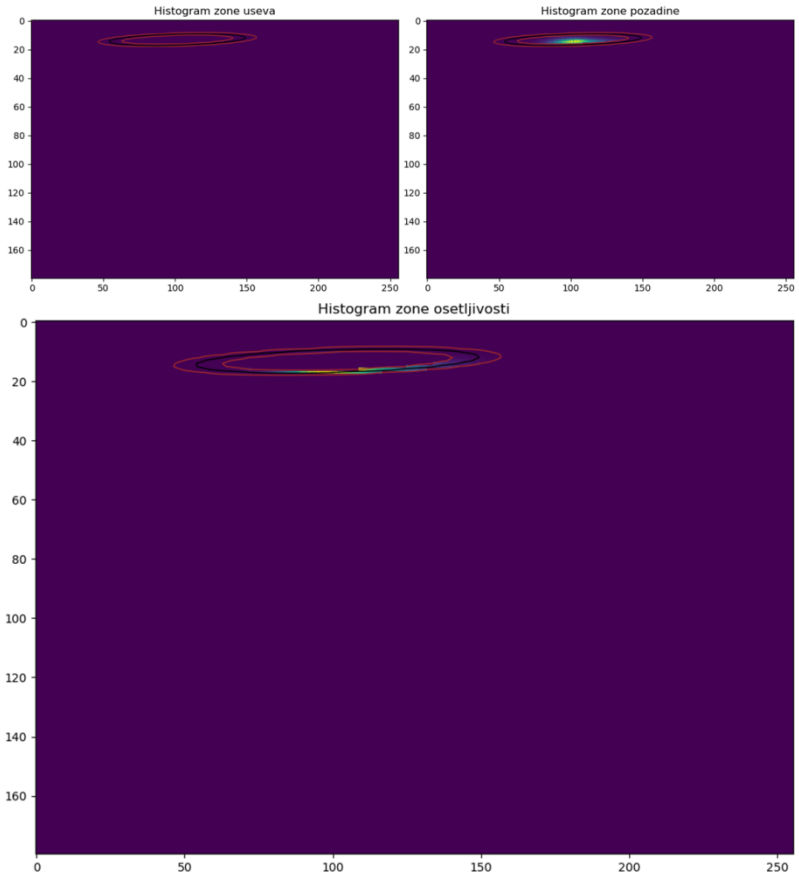
Slika 7.2.4: Prikaz u 3D histograma slike, useva i pozadine i oblasti preklapanja

7.2.4 Diskriminaciona analiza i zona osetljivosti

Ponovo primenom Kvadratne diskriminacione analize, na osnovu uzoraka useva i pozadine određene su kvadratne ravni odlučivanja (Slika 7.2.5). Za ε_s = 0,25 određene su ravni odlučivanja i zona osetljivosti, sa samo onim pikselima koji im pripadaju, kako je dato na slici 7.2.6.

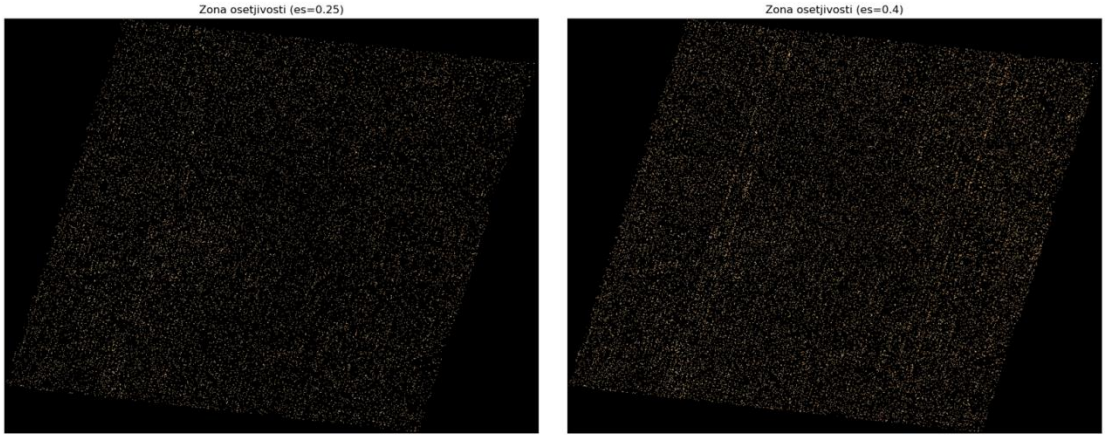


Slika 7.2.5: Ravni odlučivanja dobijene QDA



Slika 7.2.6: Klasifikacija HS vrednosti piksela i zona osetljivosti (ε_s = 0,25)

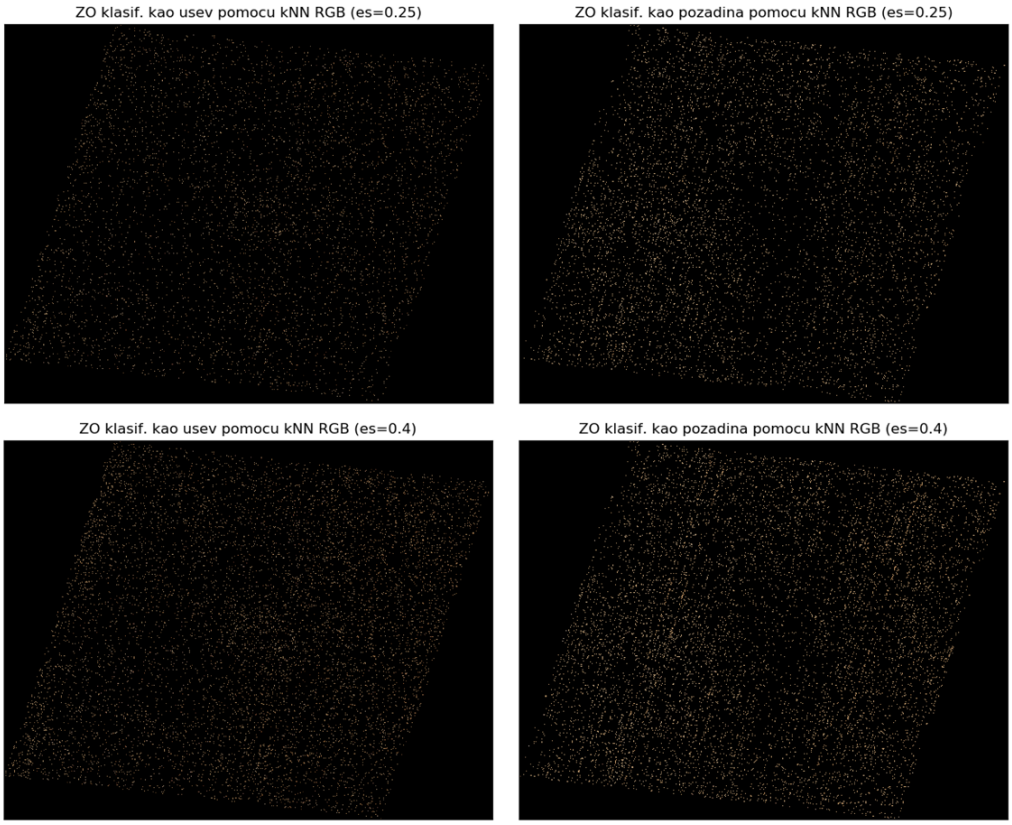
Pikseli slike koji upadaju u zonu osetljivosti su prikazani na slici 7.2.7, dok su ostali na osnovu ravni odlučivanja klasifikovani u odgovarajuće klase.



Slika 7.2.7: Pikseli iz zone osetljivosti za ε_s = 0,25 i ε_s = 0,40

7.2.5 Klasifikacija podataka iz zone osetljivosti

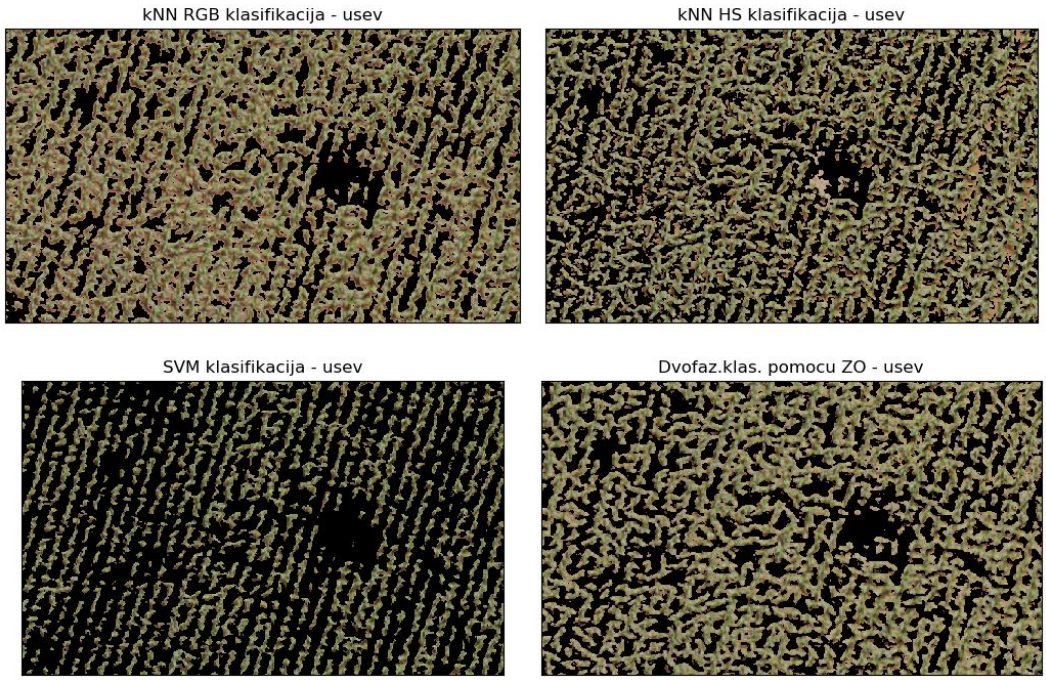
Tako određene piksele zone osetljivosti klasifikujemo dodatno, metodom kNN po sva tri kalana. Na Slici 7.2.8 su pikseli klasifikovani kao usev i kao pozadina, za oba primera (za ε_s = 0,25 i ε_s = 0,40).



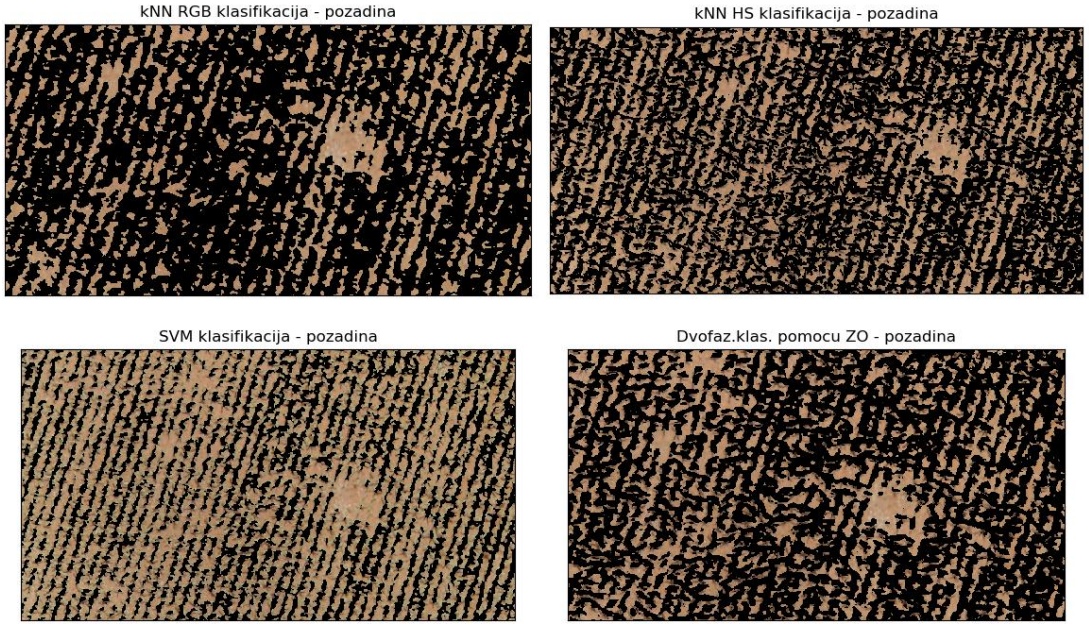
Slika 7.2.8: Pikseli iz zone osetljivosti klasifikovani kao usev i kao pozadina

7.2.6 Merenje tačnosti klasifikacije

Zatim je takođe slika prvo klasifikovana na osnovu polaznih trening podataka metodama kNN sa RGB dubinom (sva 3 kanala), kNN sa HS dubinom (sa 2 kanala), metodom SVM i metodom QDA bez zone osetljivosti. Slika 7.2.9 daje uporedni prikaz postignutog izdvajanja klase useva a Slika 7.2.10 prikaz izdvajanja klase pozadine po svim metodama klasifikacije. I jedan i drugi prikaz su uveličani, kako bi se rezultati klasifikacija mogla primetiti.



Slika 7.2.9: Uporedni (uveličani) prikaz klasifikacije (useva) pomoću četiri različita metoda



Slika 7.2.10: Uporedni (uveličani) prikaz klasifikacije (pozadina) pomoću četiri različita metoda

Tabela 7.2.2 prikazuje uporedna vremena potrebna za izvršenje klasifikacije ostalim metodama, Tabela 7.2.3, zajedno sa brojem piksela identifikovanim u zoni osetljivost i njihovim delom klasifikovanim kao usev, uporedna vremena klasifikacije pomoću QDA i klasifikacije piksela iz zone osetljivosti kNN metodom (u RGB opsegu), a Tabela 7.2.4 uporedni prikaz Ukupne tačnosti (OA) i Kappa statistike u odnosu na sliku klasifikovanu u celosti sa kNN (RBG) metodom, kao i Ukupnu tačnost postignutu u odnosu na zadati trening set podataka (uzorke useva i pozadine).

Tabela 7.2.2: Uporedna vremena klasifikacije referentnim metodama

Metod klasifikacije	Vreme trajanja (minuti:sekunde)
kNN RGB	07:03
kNN HS	04:40
SVM	00:28
QDA (bez zone osetljivosti)	00:01

Tabela 7.2.3: Vreme i broj piksela klasifikacije sa zonom osetljivosti

Parametar ε_S	Veličina zone osetljivosti	Vreme trajanja (minuti:sekunde)	Pikseli iz ZO klasifikovani kao usev
0,25	283.789 (4,76%)	01:40	127.974 (45,09%)
0,35	359.789 (6,04%)	02:18	178.600 (49,64%)
0,40	425.929 (7,15%)	02:55	217.492 (51,06%)

Tabela 7.2.4: Izmerena tačnosti klasifikacije

Primenjeni metod	Ukupna tačnost (ref. kNN RGB)	Kappa statistika (ref. kNN RGB)	Ukupna tačnost (ref. trening uzorak)
kNN RGB	100.00%	100.00%	97.04%
kNN HS	56.28%	14.93%	91.29%
SVM	74.71%	47.57%	81.16%
QDA	54.77%	12.10%	99.34%
Sa ZO (ε_s = 0,25)	56.80%	16.11%	99.46%
Sa ZO (ε_s = 0,35)	57.46%	17.41%	99.52%
Sa ZO (ε_s = 0,40)	57.97%	18.44%	99.53%

Iako u ovom slučaju SVM klasifikacija daje veću tačnost u odnosu na kNN RGB, kada gledamo postignutu tačnost u odnosu na uzorak kojeg možemo smatrati „ground-truth“ podacima, predloženi metod daje značajno bolje rezultate. Takođe, sagledavanjem uporednog prikaza izdvojenih useva i pozadine različitim metodama, i posmatrajući izdvojeni usev SVM klasifikacijom se čini da je klasifikacija dobro izvršena jer se ne primećuje mnogo pozadine. Međutim kada pogledamo pozadinu izdvojenu SVM metodom, možemo primetiti da se uviđa veliki broj piksela koji zapravo predstavljaju usev, što ukazuje na to da je usev suviše „restriktivno“ izdvojen. Prema tome, lako se uočava da su rezultati postignuti korišćenjem zone osetljivosti pouzdaniji, što se i vidi i iz tačnosti u odnosu na uzorke (trening podatke).

7.3 Primer 3 – Plantaža manga

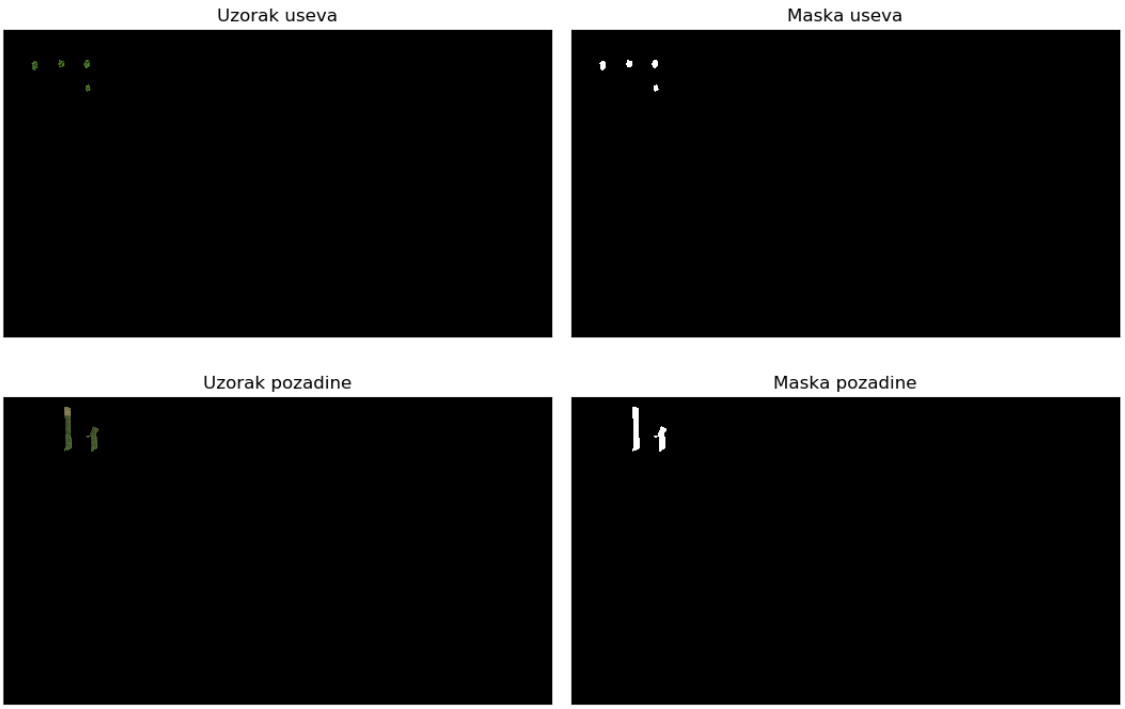
7.3.1 Ulazni podaci i predprocesiranje

Slika korišćena za ovaj test je plantaža voća mango (sa lokacije iz Portorika) (Slika 7.3.1). Sprovedeno je predprocesiranje Gausovim zamućivanjem, pri čemu se pokazalo da veličina jezgra 15 daje dobre rezultate, te je dalje korišćeno u ovom primeru. Ovde želimo testirati metod u primeru slike veće veličine, kao i u slučaju kada je očigledno da se boja useva i pozadine ne razlikuj značajno (što je vidljivo sa slike).

Slika 7.3.1: Originalna slika plantaže manga

7.3.2 Trening podaci (uzorci)

Trening podaci – oblasti na slici koje predstavljaju uzorak useva i uzorak pozadine su dati na Slici 7.3.2. Tabela 7.3.1 prikazuje detalje veličine slike i uzoraka. U ovom primeru su jasno vidljivi (okom posmatrača-analitičara) objekti koji predstavljaju pojedinačne biljke.



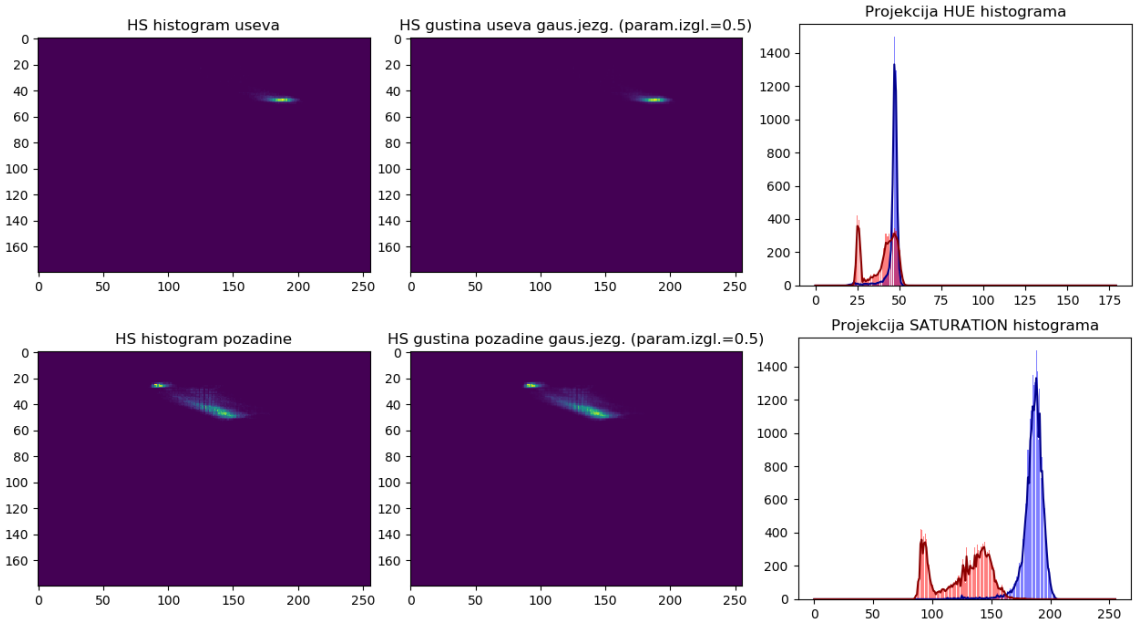
Slika 7.3.2: Određeni trening podaci useva (gore) i pozadine (dole), sa odgovarajućim maskama

Tabela 7.3.1:Detalji veličine slike i uzoraka (trening podataka)

Visina slike	Širina slike	Ukupno piksela	Pikseli za trening-usev	Pikseli za trening-pozadina
3.815	6.823	26.029.745	23.531 (0,09%)	66.469 (0,26%)

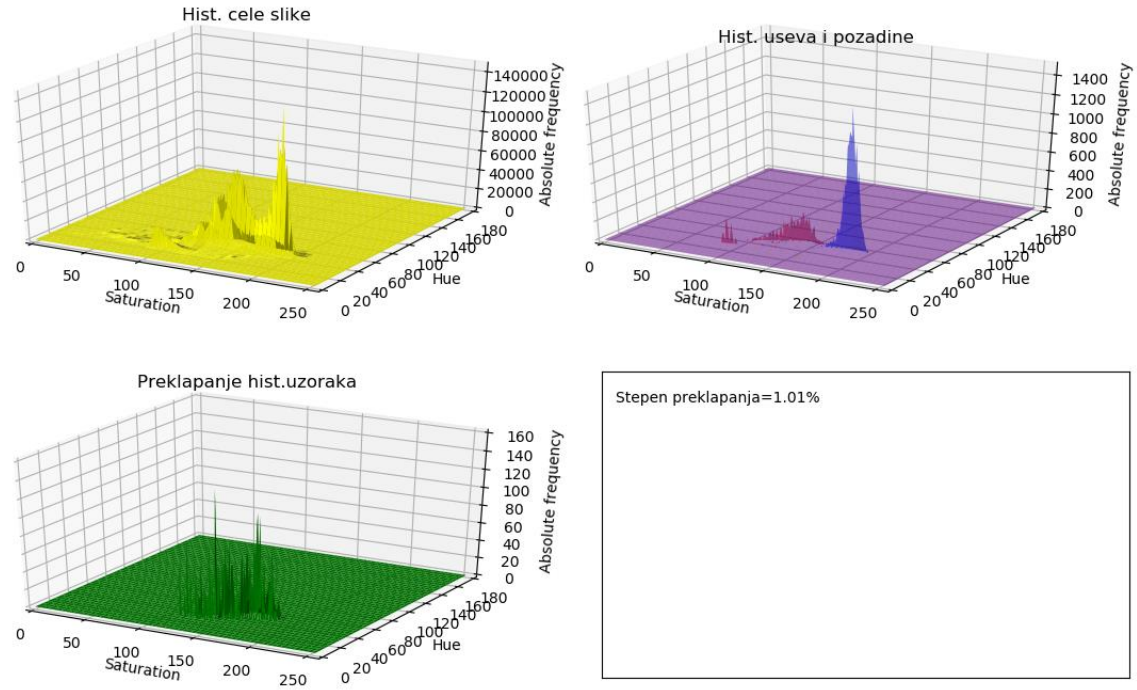
7.3.3 Gustine raspodela i stepen preklapanja

Dobijeni ujednačeni histogrami za dvodimenzionalne podatke (*Hue* i *Saturation*, H i S) su prikazani na Slici 7.3.3, gde su takođe prvo u formi „*heat* mape“ prikazani histogrami uzoraka, zatim gustine izglađene Gausovim jezgrom (sa parametrom izglađivanja 0,5), i na kraju su prikazane projekcije tih histograma na H i S ravan (crvenom i plavom bojom). Ovde se jasno vidi iz projekcija histograma na H i S ravan, da, kako je rečeno u uvodnom pasusu ovog poglavlja, boja (H) ne doprinosi toliko razdvajanju koliko to čini Saturation (S).



Slika 7.3.3: HS histogrami uzorka, izglađene gustine uzorka i projekcije na H i S ravan

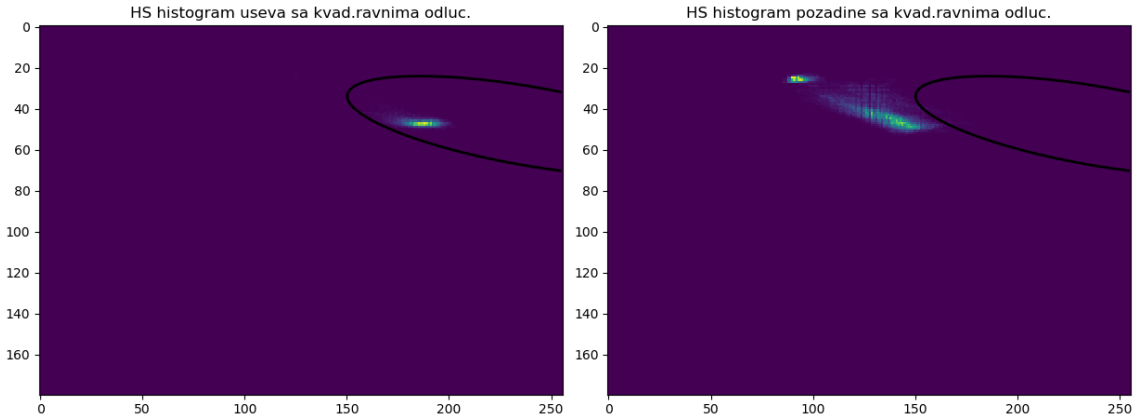
Na slici 7.3.4 je 3D dijagramom je ponovo prvo prikazan histogram čitave slike, zatim su prikazani histogrami uzorka useva i pozadine, dijagram oblasti preklapanja ta dva histograma. U ovom primeru, Stepen preklapanja iznosi 1,01% (1.331 piskel) što je veoma dobro razdvajanje.



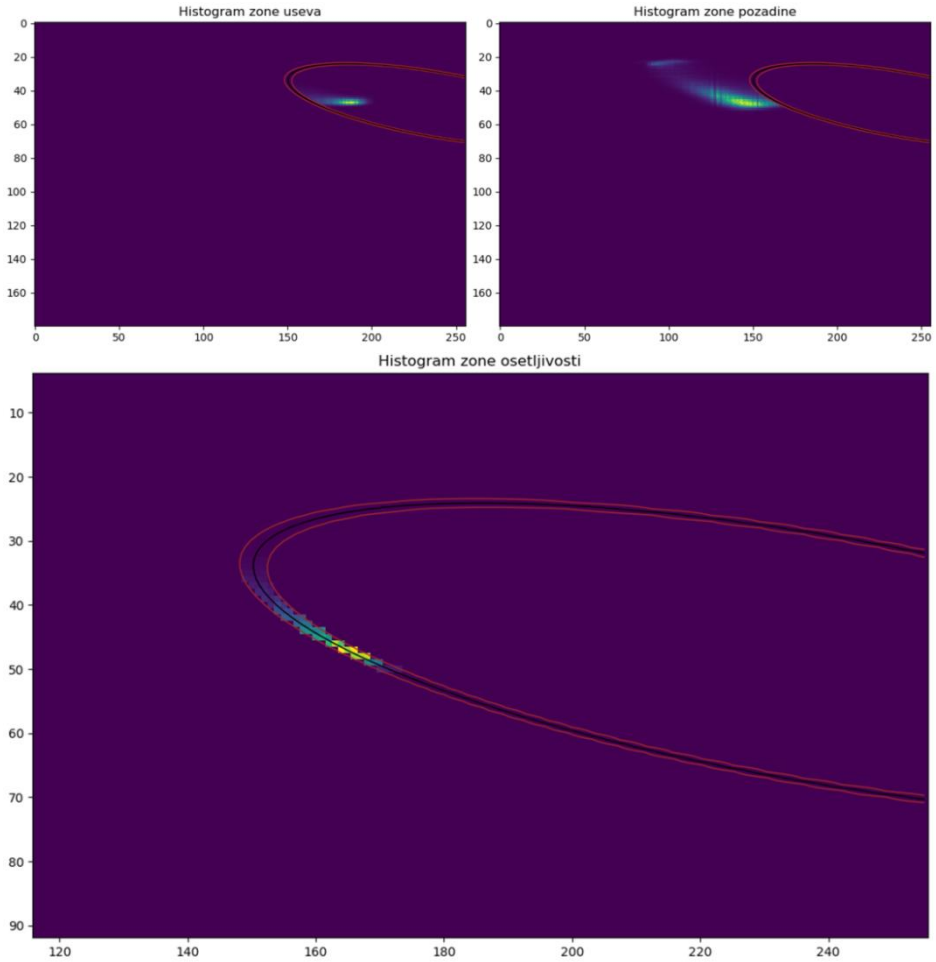
Slika 7.3.4: Prikaz u 3D histograma slike, useva i pozadine i oblasti preklapanja

7.3.4 Diskriminaciona analiza i zona osetljivosti

Pomoću QDA, na osnovu uzoraka useva i pozadine određene su kvadratne ravni odlučivanja (Slika 7.3.5). Za $\varepsilon_s = 0{,}25$ određene su ravni odlučivanja i zona osetljivosti, sa samo onim pikselima koji im pripadaju, kako je dato na slici 7.3.6.

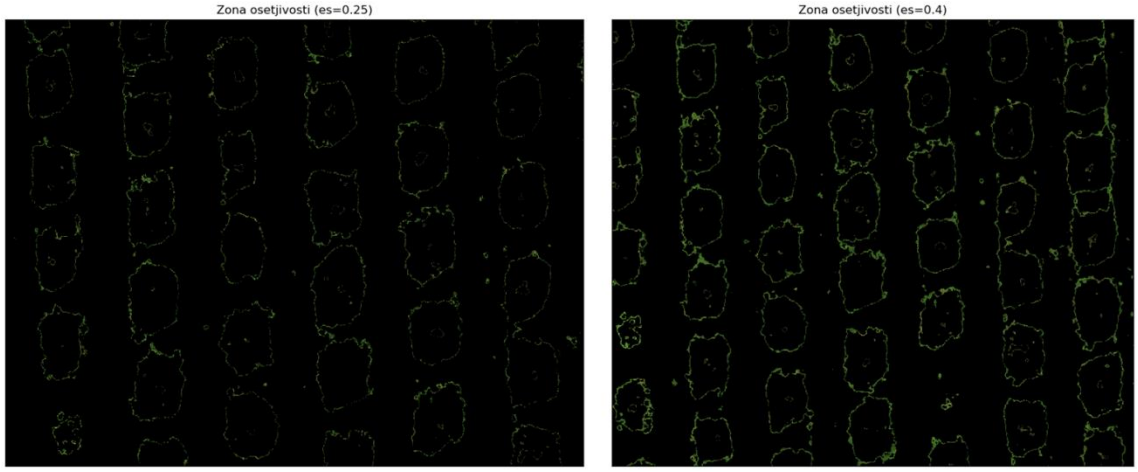


Slika 7.3.5: Ravni odlučivanja dobijene QDA



Slika 7.3.6: Klasifikacija HS vrednosti piksela i zona osetljivosti ($\varepsilon_s = 0{,}25$)

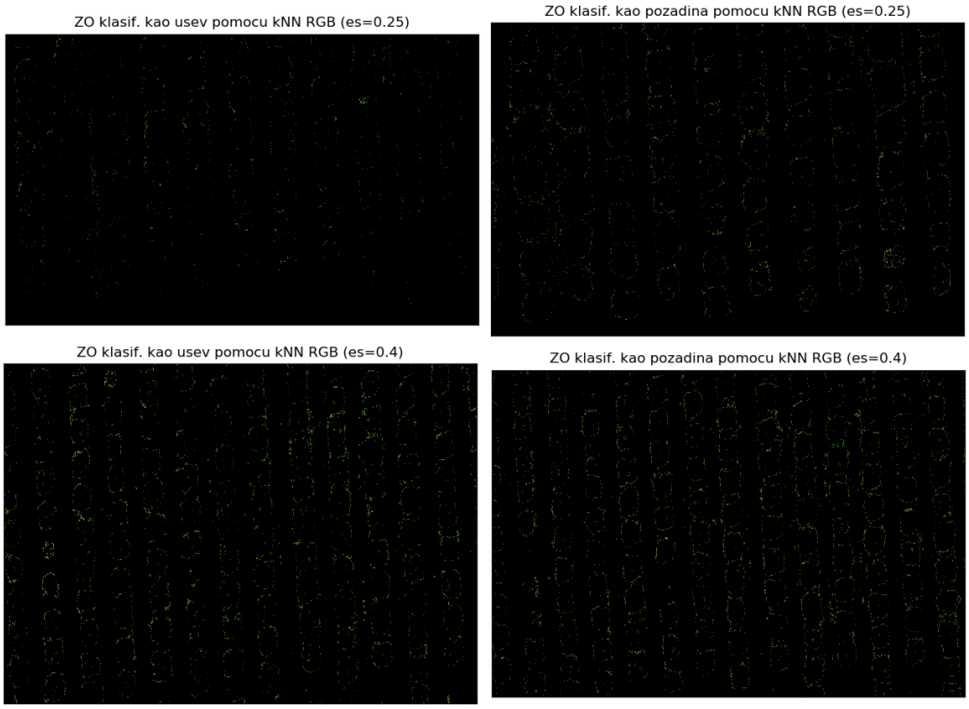
Pikseli slike koji upadaju u zonu osetljivosti su prikazani na slici 7.3.7, dok su ostali na osnovu ravni odlučivanja klasifikovani u odgovarajuće klase.



Slika 7.3.7: Pikseli iz zone osetljivosti za ε_s = 0,25 i ε_s = 0,40

7.3.5 Klasifikacija podataka iz zone osetljivosti

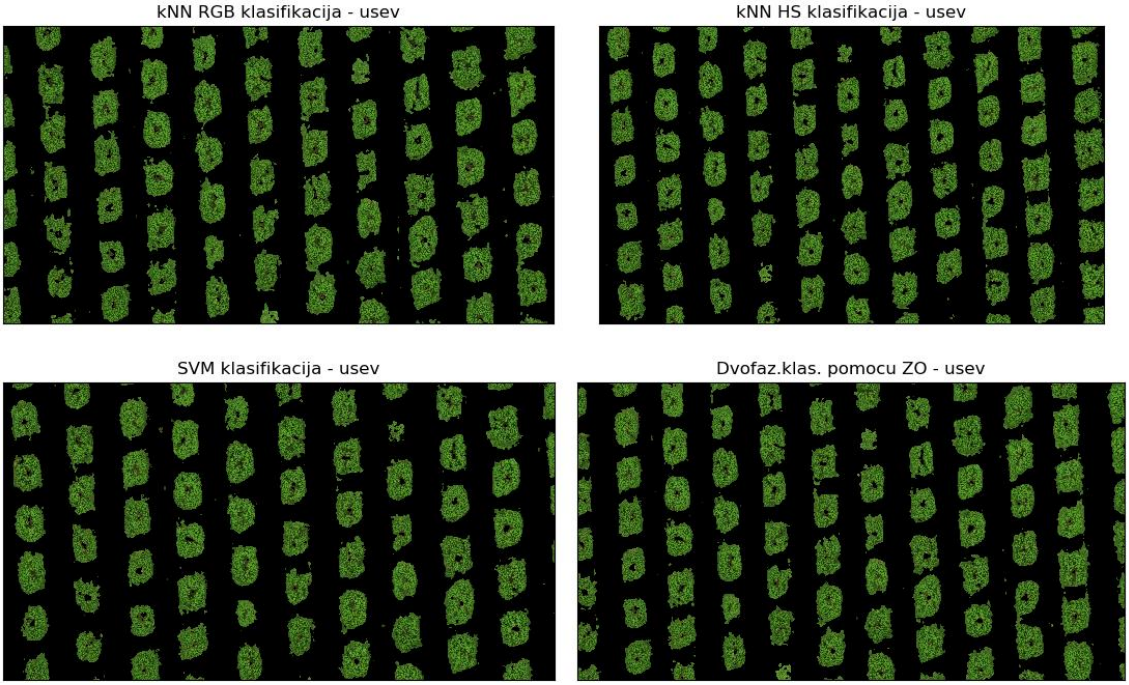
Tako određene piksele zone osetljivosti klasifikujemo dodatno, metodom kNN po sva tri kalana. Na Slici 7.3.8 su pikseli klasifikovani kao usev i kao pozadina, za oba primera (za ε_s = 0,25 i ε_s = 0,40).



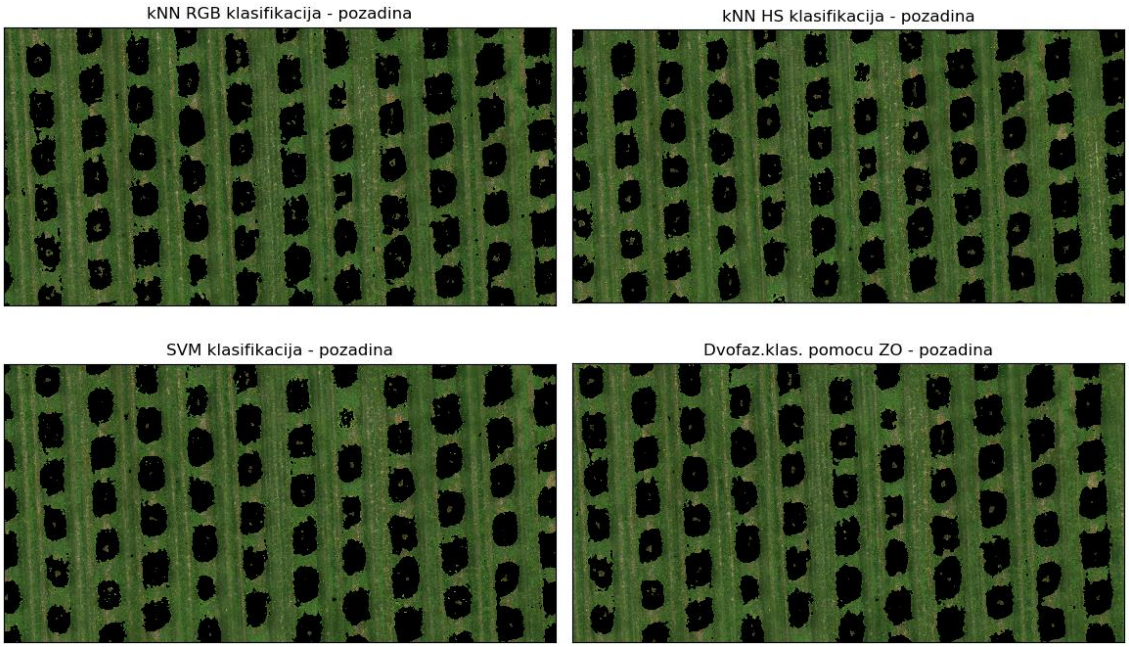
Slika 7.3.8: Pikseli iz zone osetljivosti klasifikovani kao usev i kao pozadina

7.3.6 Merenje tačnosti klasifikacije

Slika je opet prvo klasifikovana na osnovu polaznih trening podataka metodama kNN sa RGB dubinom (sva 3 kanala), kNN sa HS dubinom (sa 2 kanala), metodom SVM i metodom QDA bez zone osetljivosti. Slika 7.3.9 daje uporedni prikaz postignutog izdvajanja klase useva a Slika 7.3.10 prikaz izdvajanja klase pozadine po svim metodama klasifikacije (izuzev QDA bez zone osetljivosti, zbog sažetosti prikaza, ali ona je svakako sastavni deo predloženog metoda sa zonom osetljivosti).



Slika 7.3.9: Uporedni prikaz klasifikacije (useva) pomoću četiri različita metoda



Slika 7.3.10: Uporedni prikaz klasifikacije (pozadina) pomoću četiri različita metoda

Tabela 7.3.2 prikazuje uporedna vremena potrebna za izvršenje klasifikacije ostalim metodama, Tabela 7.3.3, zajedno sa brojem piksela identifikovanim u zoni osetljivost i njihovim delom klasifikovanim kao usev, uporedna vremena klasifikacije pomoću QDA i klasifikacije piksela iz zone osetljivosti kNN metodom (u RGB opsegu), a Tabela 7.3.4 uporedni prikaz Ukupne tačnosti (OA) i Kappa statistike u odnosu na sliku klasifikovanu u celosti sa kNN (RBG) metodom, kao i Ukupnu tačnost postignutu u odnosu na zadati trening set podataka (uzorke useva i pozadine).

Tabela 7.3.2: Uporedna vremena klasifikacije referentnim metodama

Metod klasifikacije	Vreme trajanja (minuti:sekunde)
kNN RGB	44:53
kNN HS	06:24
SVM	04:26
QDA (bez zone osetljivosti)	00:15

Tabela 7.3.3: Vreme i broj piksela klasifikacije sa zonom osetljivosti

Parametar ε_S	Veličina zone osetljivosti	Vreme trajanja (minuti:sekunde)	Pikseli iz ZO klasifikovani kao usev
0,25	545.366 (2,10%)	01:33	208.939 (38.31%)
0,35	907.547 (3.49%)	02:54	353.076 (38.90%)
0,40	1.217.861 (4.68%)	03:16	488.797 (40.14%)

Tabela 7.3.4: Izmerena tačnosti klasifikacije

Primenjeni metod	Ukupna tačnost (ref. kNN RGB)	Kappa statistika (ref. kNN RGB)	Ukupna tačnost (ref. trening uzorak)
kNN RGB	100.00%	100.00%	99.13%
kNN HS	97.46%	94.10%	97.83%
SVM	96.44%	91.81%	97.35%
QDA	97.04%	93.23%	98.17%
Sa ZO (ε_s = 0,25)	97.97%	95.33%	98.40%
Sa ZO (ε_s = 0,35)	98.35%	96.21%	98.45%
Sa ZO (ε_s = 0,40)	98.54%	96.65%	98.52%

U ovom primeru se iz podataka o tačnosti klasifikacije vidi da predložena metoda postiže veću tačnost u klasifikaciji od ostalih (kNN za HS, SVM i QDA), dok je potrebno vreme za njeno izvršenje kraće od ostalih (izuzev QDA bez zone osetljivosti), a naročito u odnosu na kNN (RGB) koja je značajno sporija. Time se može izneti zaključak da predložena metoda donosi poboljšanje i veću efikasnost klasifikacije.

8 ZAKLJUČAK

Razvojem naučne discipline daljinskog uzorkovanja i napretkom hardverskih platformi, veoma važan činilac u njihovoj primeni svakako postaje dalji napredak algoritama za procesiranje i analizu daljinski uzorkovanih podataka, što je upravo oblast kojom se bavi ova doktorska disertacija i opisano istraživanje. Daljinsko uzorkovanje je naučna oblast i tehnika za prikupljanje informacija o objektu (najčešće Zemljinoj površini) bez dolaženja u kontakt sa njim (dakle na daljinu, npr. iz letelice, satelita, itd.). Sprovodi se uzorkovanjem (očitavanjem) putem beleženja reflektovane ili emitovane energije (elektromagnetnog zračenja) objekta, procesiranjem, analiziranjem i primenom informacija. Daljinsko uzorkovanje često je upareno sa disciplinama obrade slika i geografskog informacionog sistema (GIS) za široku oblast geospatijalne nauke i tehnologije. U ovoj doktorskoj disertaciji je akcenat stavljen na oblast primene daljinskog uzorkovanja koja se naziva precizna poljoprivreda, a s obzirom na sve veću rastuću popularnost tih platformi, pažnju smo usmerili na prikupljanje daljinski uzorkovanih slika pomoću bespilotnih letelica (odnosno popularnih dronova).

Deo procesa koji se u ovoj doktorskoj disertaciji naročito opisuje je ekstrakcija smislenih informacija iz slike putem njene interpretacije i analize. Ona uključuje identifikaciju i/ili merenje različitih ciljnih objekata na slici kako bi se dobile korisne informacije o njima. Dakle, automatsko procesiranje i analiza digitalnih slika se sprovode za automatsku identifikaciju ciljnih objekata i dobijanje informacija, načelno bez ili sa veoma malo ručnih intervencija analitičara. Ipak, procesiranje i analiza se u praksi retko sprovode u potpunosti bez ljudskih intervencija i nadzora, ali se uvek teži procesu na što većem nivou automatizacije, odnosno sa što manjim brojem ručnih intervencija i unosa analitičara.

Većina standardnih funkcija obrade i analize slika su kategorizovane u: predprocesiranje, poboljšanja, transformacije i klasifikacija i analiza. U ovoj doktorskoj disertaciji smo se posebno bavili klasifikacijom daljinski uzorkovanih podataka – podataka predstavljenih u vidu slike. Ona se obično izvodi na višestrukim skupovima podataka a cilj je dodeljivanje svakog piksela slike određenoj klasi (npr. voda, vrsta šume, kukuruz, pšenica), na osnovu statističkih karakteristika intenziteta i obojenosti piksela (ili čak na osnovu prostorne povezanosti i okruženja piksela). Predstavljene su najpopularnije procedure klasifikacije u daljinskom uzorkovanju, koje se obično grupišu u dve grupe na osnovu korišćenih metoda:

- Klasifikacija bez nadzora ili nenadgledana klasifikacija, gde su detaljnije opisani *k*-means klasterovanje i Spektralno klasterovanje;
- Klasifikacija s nadzorom ili nadgledana klasifikacija, gde su detaljnije opisane metode koje baziraju na Bajesovoj teoriji odlučivanja (klasifikaciji) i Diskriminacionoj analizi, poput Sečenja gustine, Klasifikacie maksimalnom izglednošću, kao i Klasifikacija podržavajućim vektorima (SVM).

Spomenut je i niz drugih statističkih metoda koje se koriste kod daljinskog uzorkovanja, pored tehnika klasifikacije, kao što su regresija, kanonočka korelaciona analiza, Bajesove mreže uslovnih verovatnoća, itd.

Multivarijaciona analiza je prikazana kao skup statističkih metoda koje analiziraju višedimenziona merenja skupa objekata koji ispitujemo. Ona omogućuje analizu složenih nizova podataka, kada postoji mnogo nezavisnih i zavisnih promenljivih koje su u korelaciji, kako bi se obezbedile što sveobuhvatnije statističke analize. Pored ovog deskriptivnog zadatka metode multivarijacione analize se koriste u procesu zaključivanja i odlučivanja tako što se ocenjuje, na primer, stepen međuzavisnosti promenljivih i/ili testira njihova statistička značajnost. U disertaciji posebno su prikazane sledeće tehnike:

- Analiza grupisanja (klaster analiza) kao metoda koja predstavlja nenadgledanu tehniku klasifikacije. Kod klasifikacije u problemima daljinskog uzorkovanja klasterovanje se koristi kao tehnika koja pruža značajno veći nivo automatizacije procesa ali i postiže lošije rezultate. Zato je veći fokus stavljen na tehnike nadgledane klasifikacije.
- Diskriminaciona analiza je prikazana kao metoda koja se bavi problemom razdvajanja grupa i alokacijom opservacija u ranije definisane grupe. Ona ima dva osnovna cilja: Prvi, da utvrdi da li postoji statistički značajna razlika u sredinama dve ili više grupa, a zatim da odredi koja od promenljivih daje najveći doprinos utvrđenoj razlici. Ovaj cilj analize nazivamo diskriminacija ili razdvajanje među grupama. Drugi cilj odnosi se na utvrđivanje postupka za klasifikaciju opservacija na osnovu vrednosti nekoliko promenljivih u dve ili više razdvojenih, unapred definisanih grupa. Ovaj cilj analize nazivamo klasifikacija ili alokacija opservacija.

U osnovi razrađenih metoda klasifikacije kojima smo se bavili u ovoj doktorskoj disertaciji svakako je Bajesova teorija odlučivanja kao fundamentalan statistički pristup problemu klasifikacije. Pristup bazira na kvantifikaciji kompromisa između različitih odluka klasifikacije pomoću verovatnoće i cene („troška“) ili napora koji se javljaju tokom odlučivanja. Ona problem odlučivanja predstavlja u probabilističkom kontekstu. Iz Bajesove teorije odlučivanja je zatim izvedena čitava oblast statističkog prepoznavanja obrazaca, koja je prikazana opet u kontekstu klasifikacije podataka kojim se bavimo.

Kako je rečeno, problem klasifikacije može se posmatrati kao problem statističke funkcije odluke. Imamo određen broj hipoteza, a svaka hipoteza je definisana raspodelom opservacija. Mi moramo da prihvatimo jednu od hipoteza i odbacimo ostale. Ukoliko su dve populacije poznate mi imamo elementaran problem testiranja jedne hipoteze specifične raspodele u odnosu na drugu. U nekim okolnostima kategorije su definisane unapred u smislu da je verovatnoća raspodele u potpunosti poznata. U

ostalim slučajevima forma svake raspodele može biti poznata, ali parametri raspodele su ocenjeni kao primeri iz te populacije. Drugim rečima u konstrukciji procedure klasifikacije potrebno je minimizirati verovatnoću pogrešne klasifikacije, ili konkretnije, poželjno je minimizirati rezultate loših efekata pogrešne klasifikacije.

Kada je izlaz daljinskog uzorkovanja klasifikovana mapa, od suštinskog je značaja u ovoj disertaciji bilo proceniti njenu tačnost. Bez obzira na statistički metod klasifikacije piksela u ovoj doktorskoj disertaciji naglašena je važnost da klasifikaciju treba razmatrati isključivo uz određeni nivo tačnosti klasifikovanja koji je potrebno meriti (tačnost klasifikacije). U tom smislu je naročito bilo važno tokom istraživanja posvetiti posebnu pažnju tačnosti klasifikovanja, pa otuda je jasno i proističe zašto je bilo od posebnog značaja izučavanje zone osetljivosti kod klasifikovanja, koja se odnosi na upravo na onaj opseg vrednosti u kom se objekat ne može sa zadovoljavajućom sigurnošću klasifikovati u jednu od populacija, ili u našm slučaju piksel klasifikovati u jednu od spektralnih ili klasa informacija. Prema tome, cilj ovog istraživanja, u užem smislu, je bio razvoj metodologije za definisanje takve zone osetljivosti koja će pospešiti ukupnu tačnost klasifikovanja.

Kao primer naveli smo prvo načine definisanja minimalnih gubitaka u slučaju klasifikacije u dve populacije. To npr. može biti klasifikacija piksela u klasu „zemljište“ i u klasu „biljka“ (ili „usev“ i „korov“ i sl.), odnosno kako je dato u većem delu ove disertacije klasifikacija na „usev“ i „pozadinu“, tj. „ostalo“. Kada imamo prioritetne verovatnoće kategorije za dve populacije možemo označiti verovatnoću opservacije koja dolazi iz populacije P_1 sa q_1, a verovatnoću opservacije koja dolazi iz populacije P_2 sa q_2. Verovatnoća mogućih opcija populacije P_1 su definisane funkcijama raspodele. Onda je očekivanih gubitaka pogrešne klasifikacije suma proizvoda uzroka svake pogrešne klasifikacije pomnožena sa verovatnoćom njihovog pojavljivanja i ona je

$$C(2|1)*P(2|1, R)*q_1 + C(1|2)*P(1|2, R)*q_2$$

To je prosek gubitaka koji želimo da minimiziramo tj. hoćemo da podelimo naš prostor na regione R_1 i R_2 tako da očekivani gubitak bude što je moguće manji. Procedura koja minimizira taj gubitak za dato q_1 i q_2 se zove *Bajesova procedura*. Pokazali smo da ako pretpostavimo da je C(1|2)=C(2|1)=1 tada smo minimizirali verovatnoću pogrešne klasifikacije za regione:

$$R_1 : q_1 p_1(x) \geq q_2 p_2(x)$$
$$R_2 : q_1 p_1(x) < q_2 p_2(x)$$

odnosno biramo regione R_1 i R_2 prema

$$R_1 : [C(2|1)q_1]p_1(x) \geq [C(1|2)q_2]p_2(x)$$
$$R_2 : [C(2|1)q_1]p_1(x) < [C(1|2)q_2]p_2(x)$$

U opštem slučaju od m populacija ukoliko je q_i prioritetna verovatnoća dobijanja opservacija iz populacije P_i sa gustinom $p_i(x)$, i=1,...,m, tada u regionima klasifikacije $R_1,...,R_m$ koji minimiziraju očekivane vrednosti, dodeljujemo x regionu R_k ako

$$\sum_{i=1}^{m} q_i p_i(x) C(k|i) < \sum_{i=1}^{m} q_i p_i(x) C(j|i)$$

za j,k=1,...,m i j<>k a kada je C(j|i)=1 za svako i i j (i<>j)? Tada dobijamo

$$q_j p_j(x) < q_k p_k(x)$$

za j,k=1,...,m i j<>k. U ovom slučaju tačka x je u regionu R_k, ukoliko je k indeks za koje je $q_i p_i(x)$ maksimum i tada je P_k najverovatnija populacija.

Zona osetljivosti, koja predstavlja granične vrednosti regiona je definisana u slučajevima kada nismo u mogućnosti da utvrdimo stanje entiteta. Dakle, osim određivanja kategorija pojava, bitno je odrediti i granice tih pojava, tj. odrediti one oblasti u kojima možemo sa sigurnošću pretpostaviti da entiteti pripadaju definisanoj kategoriji, i one kada to ne možemo pretpostaviti sa sigurnošću. Problem nastaje kada su verovatnoće pripadanja entiteta jednoj ili drugoj kategoriji dosta male i ne zadovoljavaju nivo poverenja koji želimo da zadržimo. Pokazali smo da je kada su date dve populacije koje imaju normalnu raspodelu $N_i(\mu_i, \sigma_i^2)$, (i=1,2) zona osetljivosti konačno određena sa

$$K = \left(\Phi^{-1}(1-\varepsilon)\sigma_2 + \mu_2, \Phi^{-1}(1-\varepsilon)\sigma_1 + \mu_1\right)$$

gde su μ_i očekivane vrednosti, σ_i varijanse populacija, ε unapred zadata vrednost (parametar) i Φ funkcija (normalne) raspodele. Definisanjem metodologije određivanja zone osetljivosti rešavamo problem na način da identifikujemo elemente koji se nalaze u graničnoj oblasti, između dve kategorije, a koju nazivamo zona osetljivosti. Kako je deo metodologije predložene u ovoj disertaciji klasifikacija podataka pomoću LDA ili QDA, i kako se podaci o procenjenim (predviđenim) verovatnoćama dobijeni na taj način mogu predstaviti kumulativnom funkcijom normalne raspodele, predloženo metod određivanja zone osetljivosti je dat u nešto prilagođenoj formi:

$$K = (\Phi^{-1}(0{,}5 - \varepsilon_s),\ \Phi^{-1}(0{,}5 + \varepsilon_s))$$

gde je važno napomenuti da će u opštem slučaju zona K biti višedimenzionalna oblast, npr. 2-dimenzionalna u HS prostoru, 3-dimenzionalna u RGB, itd.

Kod postojećih metoda klasifikacije daljinski uzorkovanih podataka se ističu problemi u vezi sa velikom količinom podataka i kompromisom između tačnosti i brzine (performansi) klasifikacije, koji se ogledaju u činjenici da se i na tako velikom skupu podataka (entiteta za klasifikaciju) svi podaci tretiraju istim nivoom značajnosti i sa istim nivoom detalja (bilo da je nivo detalja suviše veliki – što dovodi do previše računski zahtevnih operacija ili suviše mali – što dovodi do nedovoljne tačnosti

klasifikacije). Uzimajući to u obzir, upravo je i bio cilj ovog istraživanja nalaženje pogodne metodologije za određivanje zone osetljivosti kod daljinskog uzorkovanja, koja će potom na prihvatljiv način klasifikovati piksele u odgovarajćue klase. Ključno je bilo da predloženu novu metodu klasifikacije podataka karakterišu:

- manja zahtevnost po pitanju utrošenih resursa za obradu podataka, naročito vremena, računarskih resursa, ljudskih resursa i
- veća tačnost klasifikacije u odnosu na metode kojima su performanse primarni kriterijum.

Shodno tome razvijen je originalni statistički pristup određivanju i primeni zone osetljivosti koji je nazvan: Dvofazna klasifikacija daljinski uzorkovanih podataka primenom zone osetljivosti. Rezultati doktorske disertacije u vidu navedene predložene metode imaju praktičnu primenu za široki broj slučajeva iz domena klasifikacije digitalnih fotografija, ne samo u oblasti daljinski uzorkovanih podataka u domenu precizne poljoprivrede. Iako je metodologija prikazana na studijama slučajeva klasifikacije daljinski uzorkovanih podataka u dve klase (usev i pozadina), ona se lako može generalizovati na klasifikaciju u više klasa i za klasifikaciju bilo kojih fenomena sadržanih u podacima.

Dvofazna klasifikacija daljinski uzorkovanih podataka primenom zone osetljivosti se može praktično primeniti i na druge probleme od interesa, koje karakteriše višedimenziona priroda i velika količina podataka (entiteta) koje je potrebno klasifikovati. Istraživanje ove doktorske disertacije pokazalo je da predložena metodologija klasifikacije zaista optimizuje proces po oba kriterijuma: i performanse-brzina obrade i postignuta tačnost, čime unapređuje rasprostranjene metode klasifikacije koje su u upotrebi. Mogućnosti za takvo unapređenje metoda klasifikacije ogledaju se u tome da se predloženi statistički model zasniva na dinamički određenom nivou detalja kojim se tretiraju različite grupe entiteta (piksela u ovom slučaju) za klasifikaciju. Samo manji deo entiteta za koje postoji značajna verovatnoća pogrešne klasifikacije se, predloženim metodom, tretira računski zahtevnim operacijama kako bi se obezbedio kompromis između performansi i tačnosti klasifikacije. To upravo predstavlja i jednu od glavnih prednosti predloženog metoda.

Time, na kraju, na osnovu istraživanja i izvršenih testiranja, možemo konstatovati da je sprovedeno istraživanje opisano u ovoj doktorskoj disertaciji u značajnoj meri potvrdilo polazne hipoteze istraživanja, sa detaljima kako je dato ispod.

Osnovna hipoteza, da je moguće ustanoviti Bajesovu proceduru koja prilikom klasifikacije podataka daljinskog uzorkovanja minimizira mogućnost greške klasifikacije, kao i zonu osetljivosti klasifikacije koja maksimizira broj tačnih klasifikacija: hipoteza je potvrđena kroz definiciju i validaciju metoda Dvofazne klasifikacije daljinski uzorkovanih podataka primenom zone osetljivosti, koja uspešno minimizira gubitak usled pogrešnih klasifikacija, odnosno pomoću statističkog pristupa i definisanja zone osetljivosti maksimizira tačnost klasifikacije (uzimajući troškove

pogrešne klasifikacije konstantim, odnosno identičnim za različite klase). Potvrda ove osnovne hipoteze ujedno predstavlja i osnovu naučnog doprinosa ove doktorske disertacije.

Ostale hipoteze, koje su potvrđene i čija potvrda predstavlja osnovu stručnog doprinosa ove doktorske disertacije:

- Klasifikacija na osnovu elemenata datih u osnovnoj hipotezi može povećati tačnost klasifikacije: hipoteza je potvrđena pomoću izrađenog softverskog rešenja kojim je sproveden veći broj nezavisnih testova (od kojih su tri prikazana u disertaciji) nad različitim ulaznim podacima, koji su pokazali da predložena metoda u opštem slučaju klasifikuje podatke na većem nivou tačnosti od drugih, široko primenjenih metoda;
- Primena takve metode može povećati nivo automatizacije obrade podataka (obrada podataka uz manji broj zadatih ulaza od strane analitičara): hipoteza je potvrđena takođe serijom testova sprovedenih u izrađenom softverskom rešenju, koji su pokazali da predložena i implementirana metoda s jedne strane zahteva manje računarskih resursa (time i vremena za izršavanje) a s druge strane zahteva minimalnu količinu ulaza od strane analitičara, u vidu malog broja brzo određenih uzoraka – trening podataka za svaku od klasa, i ne zahteva naknadnu interakciju od strane analitičara nakon obezbeđivanja trening podataka;
- Predloženi model se može primeniti na različite primene daljinskog uzorkovanja: hipoteza je potvrđena takođe sprovedenim testovima za različite vrste ulaza. Iako su u disertaciji opisani slučajevi primene u svrhu precizne poljoprivrede, testovi su sprovedeni nad heterogenim oblicima ulaznih podataka (npr. ratarske kulture u fazi rasta kada su biljke manje ili više zgusnute, plantaže kod kojih se biljke ne dodiruju, i sl.). Iako nije prikazano u okviru ove disertacije, opisani metod se lako primenjuje i na ostale primene daljinskog uzorkovanja, ne samo u poljoprivredi.

Konačno, mogućnosti za dalji pravac istraživanja ogledaju se u mogućim unapređenjima i proširenjima predloženog metoda na način da se razrađuje postupak u smislu broja iterativnih koraka primene metoda u zavisnosti od ulaznih parametara i podataka (npr. identifikovanje zone osetljivosti unutar zone osetljivosti, na višem nivou detalja), kao i da se razrade mogućnosti kombinovanja različitih metoda klasifikacije nad različitim skupovima entiteta identifikovanih predloženim metodom (entieti koji upadaju i koji su izvan zone osetljivosti).

8.1 Doprinosi doktorske disertacije

Doprinos doktorske disertacije ogleda se u definisanju sistematskog pregleda proučavane oblasti kao i predloga originalnog modela za klasifikaciju daljinski

uzorkovanih podataka. Razvijeni model može da se koristi za klasifikaciju daljinski uzorkovanih podataka, kao sredstva za identifikaciju i evaluaciju pojava i fenomena koji se javljaju na zemljinoj površini ili generalno na bilo kom objektu proučavanom daljinskom detekcijom.

Predložena metodologija, nazvana Dvofazna klasifikacija daljinski uzorkovanih podataka primenom zone osetljivosti, potvrđena je kroz veći broj testova (od kojih su u ovoj doktorskoj disertaciji izložena tri) izvršenim nad stvarnim podacima, kojima se rešavaju realni problemi iz prakse (npr. brojanje biljaka, analiza zdravstvenog stanja biljaka, identifikacija biljaka zahvaćenih nekim oblikom stresa) i uporednom analizom rezultata koje postižu ostale metode klasifikacije.

Doktorskom disertacijom se predlaže i softversko rešenje za brzo određivanje trening seta podataka, računanje zone osetljivosti i klasifikaciju podataka tom metodom, te računanje tačnosti klasifikacije za evaluaciju postignutih rezultata, kojim se ubrzava proces istraživanja i postizanje krajnjih rezultata.

Može se zaključiti da su rezultati, proistekli iz istraživanja u doktorskoj disertaciji, pružili nekoliko različitih doprinosa, među kojima su:

- Celovit prikaz problematike daljinskog uzorkovanja i klasifikacije podataka (slika) i modela koji se primenjuju.
- Predlog originalnog statističkog modela klasifikacije daljinski uzorkovanih podataka koji se zasniva na određivanju zone osetljivosti.
- Unapređenje klasifikacije koje se ogleda u povećanju tačnosti i smanjenom broju resursa.
- Implementacija predloženog modela u konkretnim problemima daljinskog uzorkovanja (s fokusom na preciznu poljoprivredu) i verifikacija dobijenih rezultata kroz praktičnu primenu modela.
- Potvrda postavljenih hipoteza i predstavljanje rezultata dobijenih predloženim modelom klasifikacije daljinski uzorkovanih podataka.
- Multidisciplinarnosti teme istraživanja koja se zasniva na pomenutim statističkim metodama.
- Predlog softverskog rešenja za klasifikaciju podataka pomoću opisanog metoda.

9 LITERATURA

1. Aber, S. E., & Aber, J. W. (2017). Geographic Information Systems and Remote Sensing. In S. E. Aber, & J. W. Aber, *Map Librarianship - A Guide to Geoliteracy, Map and GIS Resources and Services* (pp. 71-85). Elsevier Ltd.

2. Anderberg, M. R. (1973). *Cluster Analysis for Applications.* London: Elsevier Inc.

3. Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour Detection and Hierarchical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(5), 898-916.

4. Ballesteros, R., Ortega, J. F., Hernández, D., & Moreno, M. A. (2014). Applications of georeferenced high-resolution images obtained with unmanned aerial vehicles. Part I: Description of image acquisition and processing. *Precision Agriculture, 15,* 579-592.

5. Bauer, S. D., Korč, F., & Forstner, W. (2011). The potential of automatic methods of classification to identify leaf diseases from multispectral images. *Precision Agriculture, 12,* 361-377.

6. Behmann, J., Mahlein, A.-K., Rumpf, T., Romer, C., & Plumer, L. (2015). A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture, 16,* 239–260.

7. Benediktsson, J., Swain, P., & Ersoy, O. (1990). Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing, 28*(4), 540-552.

8. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York, USA: Springer.

9. Bogosavljević, S. (1985). *Apriorne metode klasifikacije ekonomskih pojava, Doktorska disertacija.* Beograd.

10. Bovik, A. (2009). *The Essential Guide to Image Processing.* London: Academic Press, Elsevier Inc.

11. Bulajić, M. (2002). *Geodemografski model tržišnog prostora Srbije, Doktorska disertacija.* Beograd: Univerzitet u Beogradu, Fakultet organizacionih nauka.

12. Campbell, J., & Shin, M. (2011). *Essentials to Geographic Information Systems.* Irvington: Flat World Knowledge, Inc.

13. Canada Centre for Remote Sensing. (2007). *Fundamentals of Remote Sensing.* Ottawa: Canada Centre for Remote Sensing.

14. Candiago, S., Remondino, F., Giglio, M. D., Dubbini, M., & Gattelli, M. (2015). Evaluating Multispectral Images and Vegetation Indices for Precision Farming Applications from UAV Images. *Remote Sensing, 7*, 4026-4047.

15. Congalton, R. G. (1991). A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing of Environment, 37*, 35-46.

16. Deekshatulu, B., Subhadra, P., Sasikala, K., Padmapriya, K., Lakshmi, V., Karunakar, T., . . . Kamaraju, M. (1995). On some applications of statistical methods in remote sensing. *Indian Journal of Pure and Applied Mathematics, 26*(6), 579-597.

17. Dobrota, M. P. (2015). *Statistički pristup formiranju kompozitnih indikatora zasnovan na Ivanovićevom odstojanju. Doktorska disertacija.* Beograd: Univerzitet u Beogradu, Fakultet organizacionih nauka.

18. Dobrota, M., & Dobrota, M. (2016). ARWU Ranking Uncertainty and Sensitivity: What If the Award Factor Was Excluded? *Journal of the Association for Information Science and Technology, 67*(2), 480-482.

19. Dobrota, M., Delibašić, B., & Delias, P. (2016). A Skiing Trace Clustering Model for Injury Risk Assessment. *International Journal of Decision Support System Technology, 8*(1), 56-68.

20. Đoković, A. M. (2013). *Strukturna korelaciona analiza u interpretaciji vektorskih koeficijenata korelacije, Doktorska disertacija.* Beograd: Univerzitet u Beogradu, Fakultet organizacionih nauka.

21. Duda, R. O., Hart, P. E., & Stork, D. G. (1995). *Pattern Classification and Scene Analysis* (2nd edition ed.). Hoboken, NJ, USA: John Wiley & Sons Inc.

22. Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification* (2nd ed.). New York, USA: John Wiley & Sons, Inc.

23. Dutta, R., Stein, A., & Patel, N. (2008). Delineation of Diseased Tea Patches Using MXL and Texture Based Classification. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 37*, 1693–1700.

24. Evans, F. (1998). Statistical Methods in Remote Sensing. *Proceedings of the 3rd National Earth Resource Assessment workshop*, (pp. 1-26). Brisbane, Australia.

25. Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision, 59*(2), 167-181.

26. Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern Recognition, 41,* 176-190.

27. Fischer, M. M., & Nijkamp, P. (1992). Geographic information systems and spatial analysis. *The Annals of Regional Science, 26*(1), 3-17. doi:10.1007/BF01581477

28. Fisher, P. F., & Pathirana, S. (1990). The evaluation of fuzzy membership of land cover classes in the suburban zone. *Remote Sensing of Environment, 34*(2), 121-132.

29. Fisher, R. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics, 7*(2), 179-188.

30. Franke, J., & Menz, G. (2007). Multi-temporal wheat disease detection by multi-spectral remote sensing. *Precision Agriculture, 8,* 161-172.

31. Gallego, J., Craig, M., Michaelsen, J., Bossyns, B., & Fritz, S. (2009). *GEOSS Community of Practice Ag 0703a. Best practices for crop area estimation with Remote Sensing.* Joint Research Centre, Institute for the Protection and Security of the Citizen. Luxembourg: European Commision, Publications Office of the European Union.

32. Gong, P., & Howarth, P. (1990). An assessment of some factors influencing multispectral land-cover classification. *Photogrammetric Engineering and Remote Sensing, 56*(5), 597-603.

33. Gonzalez, R. C., & Woods, R. E. (2002). *Digital Image Processing.* Upper Saddle River, New Jersey, USA: Prentice Hall.

34. Gonzalez-Dugo, V., Zarco-Tejada, P., Nicola's, E., Nortes, P. A., Alarco'n, J. J., Intrigliolo, D. S., & Fereres, E. (2013). Using high resolution UAV thermal imagery to assess the variability in the water status of five fruit tree species within a commercial orchard. *Precision Agriculture, 14,* 660-678.

35. Hagen, L., & Kahng, A. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 11*(9), 1074-1085.

36. Hague, T., Tillett, N. D., & Wheeler, H. (2006). Automated crop and weed monitoring in widely spaced cereals. *Precision Agriculture, 7,* 21-32.

37. Hanuschak, G., & Delincé, J. (2004). Utilization of Remotely Sensed Data and Geographic Information Systems (GIS) for Agricultural Statistics in the United States and the European Union. *Proceedings of the 3rd World Conference on Agricultural and Environmental Statistical Application* (pp. 2-4). Cancun, Mexico: Food and Agriculture Organization of the United Nations.

38. Hočevar, M., Širok, B., Godeša, T., & Stopar, M. (2014). Flowering estimation in apple orchards by image analysis. *Precision Agriculture, 15*, 466-478.

39. Huisman, O., & de By, R. A. (2009). *Principles of Geographic Information Systems: An Introductory Textbook.* Enschede: The International Institute for Geo-Information Science and Earth Observation (ITC).

40. Hung, C., Xu, Z., & Sukkarieh, S. (2014). Feature Learning Based Approach for Weed Classification Using High Resolution Aerial Images from a Digital Camera Mounted on a UAV. *Remote Sensing, 6*, 12037-12054.

41. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*, 651-666.

42. Jannoura, R., Brinkmann, K., Uteau, D., Bruns, C., & Joergensen, R. G. (2015). Monitoring of crop biomass using true colour aerial photographs taken from a remote controlled hexacopter. *Biosystems Engineering, 129*, 341-351.

43. Jeremić, V. (2012). *Statistički model efikasnosti zasnovan na Ivanovićevom odstojanju. Doktorska disertacija.* Beograd: Univerzitet u Beogradu, Fakultet organizacionih nauka.

44. Jovanović, D., Govedarica, M., Sabo, F., Bugarinović, Ž., Novović, O., Beker, T., & Lauter, M. (2015). Land Cover change detection by using Remote Sensing – A Case Study of Zlatibor (Serbia). *Geographica Pannonica*, *19*(4), 162-173.

45. Jovanović, V., Đurđev, B., Srdić, Z., & Stankov, U. (2012). *Geografski informacioni sistemi.* Beograd: Univerzitet u Novom Sadu, Univerzitet Singidunum.

46. Kapetsky, J. M., & Aguilar-Manjarrez, J. (2007). *Geographic information systems, remote sensing and mapping for the development and management of marine aquaculture.* Rome: Food and Agriculture Organization of the United Nations.

47. Khorram, S., Wiele, C. F., Koch, F. H., Nelson, S. A., & Potts, M. D. (2016). *Principles of Applied Remote Sensing.* New York: Springer Science+Business Media LLC.

48. Klecka, W. R. (1980). *Discriminant Analysis.* Newbury Park, California: Sage Publications Inc.

49. Kovačić, Z. J. (1994). *Multivarijaciona Analiza.* Beograd, Srbija: Univerzitet u Beogradu, Ekonomski fakultet.

50. Krapivin, V. F., Varotsos, C. A., & Soldatov, V. Y. (2015). Remote-Sensing Technologies and Data Processing Algorithms. In *New Ecoinformatics Tools in Environmental Science* (pp. 119-219). Cham: Springer International Publishing.

51. Lachenbruch, P. A. (1975). *Discriminant Analysis.* New York: Hafner Press.

52. Lapaine, M., & Frančula, N. (2000/2001). Kartografija i daljinska istraživanja. *Bilten Znanstvenog vijeća za daljinska istraživanja i fotointerpretaciju, 15-16*, 145-154.

53. Larsolle, A., & Muhammed, H. H. (2007). Measuring crop status using multivariate analysis of hyperspectral field reflectance with application to disease severity and plant density. *Precision Agriculture, 8*, 37-47.

54. Lechner, A. M., Langford, W. T., Bekessy, S. A., & Jones, S. D. (2012). Are landscape ecologists addressing uncertainty in their remote sensing data? *Landscape Ecology, 27*(9), 1249-1261.

55. Lee, W., Alchanatis, V., Yang, C., Hirafuji, M., & Moshou, D. (2010). Sensing technologies for precision specialty crop production. *Computers and Electronics in Agriculture, 74*, 2-33.

56. Levin, N. (1999). *Fundamentals of Remote Sensing.* Tel Aviv, Israel: Remote Sensing Laboratory, Geography Department, Tel Aviv University.

57. Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographical Information Systems and Science.* Chichester: John Wiley & Sons.

58. Luxburg, U. v. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395-416.

59. Ma, Y., Wu, H., Wang, L., Huang, B., Ranja, R., Zomaya, A., & Jie, W. (2015). Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems, 51*, 47-60.

60. Manić, E., Gajović, V., & Popović, S. (2016). Geografski informacioni sistemi u poljoprivredi. *Ekonomske ideje i praksa*(21), 45-58.

61. Manson, S. M., Bonsal, D. B., Kernik, M., & Lambin, E. F. (2015). Geographic Information Systems and Remote Sensing. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., pp. 64-68). Elsevier Ltd.

62. McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition.* Hoboken, New Jersey: John Wiley & Sons, Inc.

63. Mok, P., Huang, H., Kwok, Y., & Au, J. (2012). A robust adaptive clustering analysis method for automatic identification of clusters. *Pattern Recognition, 45*, 3017-3033.

64. Morris, R. D., Kottas, A., Taddy, M., Furfaro, R., & Ganapol, B. D. (2008). A Statistical Framework for the Sensitivity Analysis of Radiative Transfer Models. *IEEE Transactions on Geoscience and Remote Sensing, 46*(12), 4062-4074.

65. Morrison, D. F. (1976). *Multivariate Statistical Methods.* New York: McGraw-Hill.

66. Muhammed, H. H. (2005). Hyperspectral Crop Reflectance Data for characterising and estimating Fungal Disease Severity in Wheat. *Biosystems Engineering, 91*(1), 9-20.

67. Mynemi, R. B., Hall, F. G., Sellers, P. J., & Marshak, A. L. (1995). The Interpretation of Spectral Vegetation Indexes. *IEEE Transactions on Geoscience and Remote Sensing, 33*(2), 481-486.

68. Nguyen, H. (2009). *Spatial Statistical Data Fusion for Remote Sensing Applications, Dissertation.* Los Angeles: University of California.

69. Nieuwenhuizen, A. T., Tang, L., Hofstee, J. W., Muller, J., & Henten, E. J. (2007). Colour based detection of volunteer potatoes as weeds in sugar beet fields using machine vision. *Precision Agriculture, 8*, 267-278.

70. Oommen, T., Misra, D., Twarakavi, N. K., Prakash, A., Sahoo, B., & Bandopadhyay, S. (2008). An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing. *Mathematical Geosciences, 40*, 409-424.

71. Patz, T., & Preusser, T. (2012). Fast Parameter Sensitivity Analysis of PDE-Based Image Processing Methods. *12th European Conference on Computer Vision* (pp. 140-153). Florence: Springer Berlin Heidelberg.

72. Peña, J. M., Torres-Sánchez, J., Serrano-Pérez, A., Castro, A. I., & López-Granados, F. (2015). Quantifying Efficacy and Limits of Unmanned Aerial

Vehicle (UAV) Technology for Weed Seedling Detection as Affected by Sensor Resolution. *Sensors, 15*, 5609-5626.

73. Peña-Barragan, J. M., López-Granados, F., Jurado-Exposito, M., & Garcia-Torres, L. (2010). Sunflower yield related to multi-temporal aerial photography, land elevation and weed infestation. *Precision Agriculture, 11*, 568-585.

74. Pérez-Ortiz, Pena, J., Gutiérrez, P., Torres-Sánchez, J., Hervás-Martínez, C., & López-Granados, F. (2015). A semi-supervised system for weed mapping in sunflower crops using unmanned aerial vehicles and a crop row detection method. *Applied Soft Computing, 37*, 533-544.

75. Prabhakar, T. V., Gintu, X., Geetha, P., & Soman, K. P. (2015). Spatial Preprocessing Based Multinomial Logistic Regression For Hyperspectral Image Classification. *Procedia Computer Science, 46*, 1817-1826.

76. Radojičić, Z. (2001). *Statističko merenje intenziteta pojava, Magistarski rad.* Beograd: Univerzitet u Beogradu, Fakultet organizacionih nauka.

77. Radojičić, Z. (2007). *Statistički model ocenjivanja na subjektivno procenjenim karakteristikama. Doktorska disertacija.* Beograd: Univerzitet u Beogradu, Fakultet organizacionih nauka.

78. Radojičić, Z., Vukmirović, D., & Glišin, I. (1997). Application of the Statistical Methods In CCD Stellar Photometry. *Proceedings of the 4th Balcan Conference on Operational Research. 2.* Thessaloniki, Greece: Hellenic Operational Research Society (HELORS).

79. Rajan, N., & Maas, S. J. (2009). Mapping crop ground cover using airborne multispectral digital imagery. *Precision Agriculture, 10*, 304-318.

80. Rao, C. R. (1964). *The Use and Interpretation of Principal Component Analysis in Applied Research.* Sankhya: Indian Statistical Institute.

81. Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* New York: John Wiley & Sons.

82. Rogerson, P. A. (2001). *Statistical Methods for Geography*. London: SAGE Publications Ltd.

83. Römer, C., Bürling, K., Hunsche, M., Rumpf, T., Noga, G., & Plümer, L. (2011). Robust fitting of fluorescence spectra for pre-symptomatic wheat leaf rust detection with Support Vector Machines. *Computers and Electronics in Agriculture, 79*, 180-188.

84. Rother, C., Kolmogorov, V., & Blake, A. (2004). "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG), 23*(3), 309-314.

85. Rumpf, T. (2012). *Finding spectral features for the early identification of biotic stress in plants, Dissertation.* Institut für Geodäsie und Geoinformation. Bonn: Landwirtschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

86. Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., & Plümera, L. (2010). Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture, 74,* 91-99.

87. Sellers, P. J. (1985). Canopy reflectance, photosynthesis and transpiration. *International Journal of Remote Sensing, 6*(8), 1335-1372.

88. Shi, J., & Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888-905.

89. Signal and Information Processing Laboratory. (2013). *SV7: K-means and Spectral Clustering.* Zurich: Swiss Federal Institute of Technology.

90. Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science, 21*(2), 110-124.

91. Song, X., Wang, J., Huang, W., Liu, L., Yan, G., & Pu, R. (2009). The delineation of agricultural management zones with high resolution remotely sensed data. *Precision Agriculture, 10,* 471-487.

92. Srivastava, P. K., Mukherjee, S., Gupta, M., & Islam, T. (2014). *Remote Sensing Applications in Environmental Research.* Springer International Publishing.

93. Stein, A., Bastiaanssen, W. G., Bruin, S. D., Cracknell, A. P., Curran, P. J., Fabbri, A. G., . . . Saldana, A. (1998). Integrating spatial statistics and remote sensing. *International Journal of Remote Sensing, 19*(9), 1793-1814.

94. Sutton, T., Dassau, O., & Sutton, M. (2009). *A Gentle Introduction to GIS.* East London: Spatial Planning and Information, Department of Land Affairs, Eastern Cape.

95. Swain, M. J., & Ballard, D. H. (1992). Indexing via Color Histograms. In S. A.K., & W. H. (Ed.), *Active Perception and Robot Vision. NATO ASI Series (Series F:*

Computer and Systems Sciences). 83, pp. 261-273. Berlin: Springer, Berlin, Heidelberg.

96. Tempfli, K., Kerle, N., Huurneman, G. C., & Janssen, L. L. (2009). *Principles of Remote Sensing: An introductory textbook.* Enschede: The International Institute for Geo-Information Science and Earth Observation (ITC).

97. Tomlinson, R. F. (1968). *A Geographic Information System for Regional Planning.* Ottawa: Department of Forestry and Rural Development, Government of Canada.

98. Torres-Sánchez, J., López-Granados, F., Castro, A. I., & Peña-Barragán, J. M. (2013). Configuration and Specifications of an Unmanned Aerial Vehicle (UAV) for Early Site Specific Weed Management. *PLoS ONE, 8*(3).

99. Truett, J., Cornfield, J., & Kannel, W. (1967). A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham. *Journal of Chronic Diseases, 20*(7), 511-24.

100. Ustinov, E. A. (2015). *Sensitivity Analysis in Remote Sensing.* Springer International Publishing AG.

101. Vukmirović, D. (1992). *Model hijerarhijskog klasifikovanja.* Beograd: Univerzitet u Beogradu, Ekonomski fakultet.

102. Vuković, N. (1977). *Generalizacija višestrukog i kolektivnog koeficijenta korelacije, Doktorska disertacija.* Novi Sad: Univerzitet u Novom Sadu, Prirodno-matematički fakultet.

103. Vuković, N. (1987). *Statistička analiza.* Beograd: Naučna knjiga.

104. Webb, A. R., & Copsey, K. D. (2011). *Statistical Pattern Recognition* (3rd ed.). Malvern, United Kingdom: John Wiley & Sons, Ltd.

105. Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd edition ed.). San Francisco, CA, USA: Morgan Kaufmann.

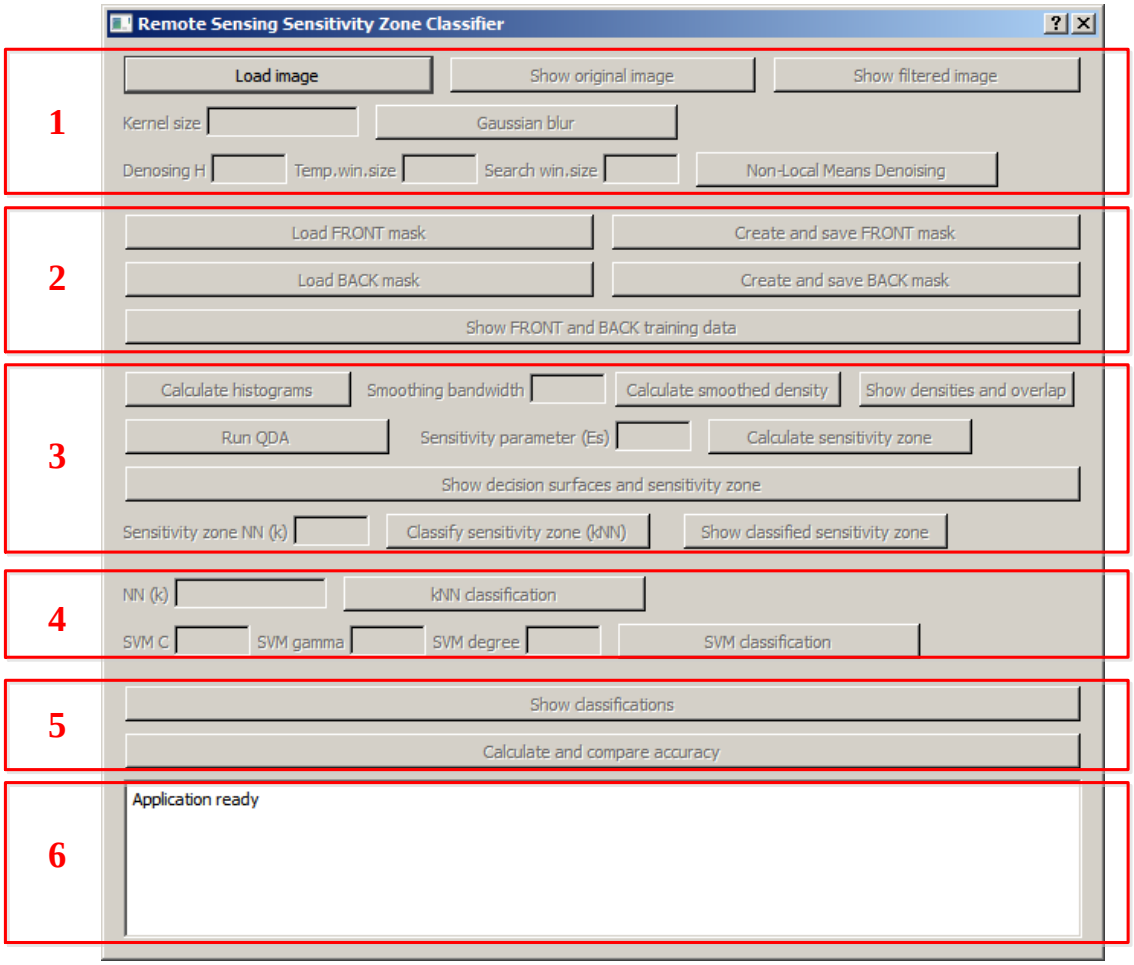
106. Xiang, T., & Gong, S. (2008). Spectral clustering with eigenvector selection. *Pattern Recognition, 41*, 1012-1029.

107. Zhang, C., & Kovacs, J. M. (2012). The application of small unmanned aerial systems for precision agriculture: a review. *Precision Agriculture, 13*, 693-712.

108. Zhang, Y., Slaughter, D. C., & Staab, E. S. (2012). Robust hyperspectral vision-based classification for multi-season weed mapping. *ISPRS Journal of Photogrammetry and Remote Sensing, 69*, 65-73.

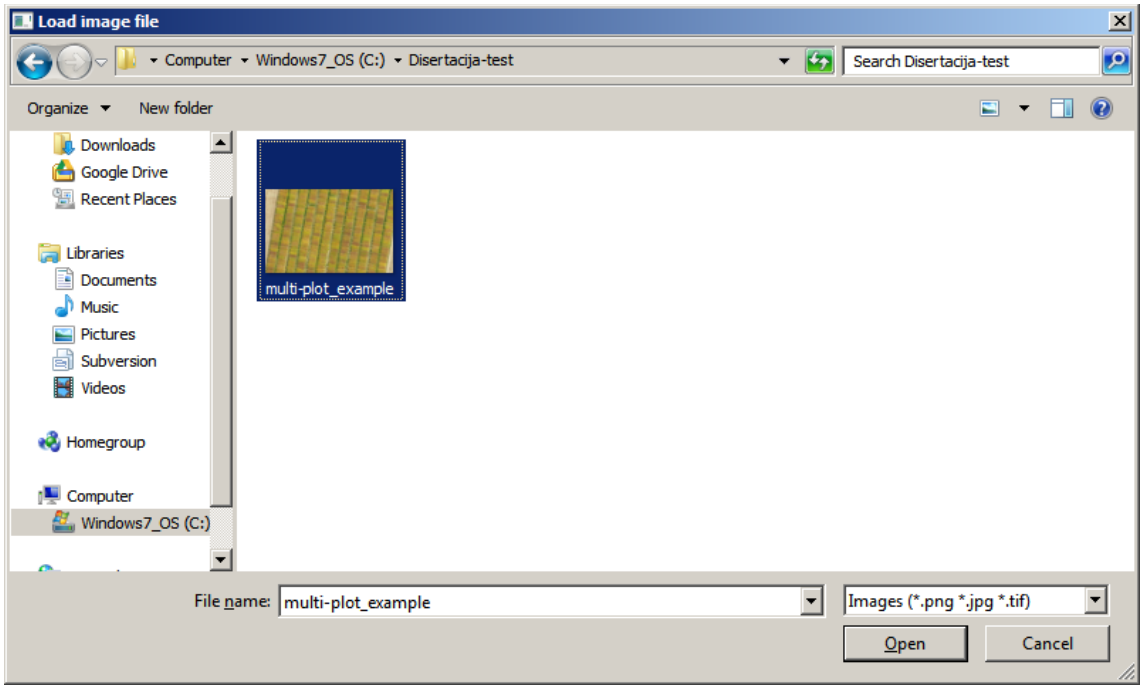
PRILOG A – PRIKAZ RAZVIJENOG SOFTVERSKOG REŠENJA

Tokom izrade ove doktorske disertacije i sprovođenja istraživanja koje ona opisuje, razvijeno je softversko rešenje koje implementira sve algoritme i korake u obradi podataka kako bi se dobili svi potrebni rezultati i međurezultati od značaja za istraživanje. Softver je implementiran pomoću programskog jezika Python (v3.4), a korišćene su i sledeće biblioteke: NumPy (v1.14), OpenCV (v3.4), Matplotlib (v2.2), Scikit-learn (v0.19), PySide (v1..2) i Qt (v4.8). Softver je nazvan „Klasifikator pomoću zone osetljivosti u daljinskom uzorkovanju“ (eng. *Remote Sensing Sensitivity Zone Clasifier*). Glavni prozor softvera (Slika A.1), koji se dobija nakon pokretanja, sastoji se od 6 segmenata, koji su opisani u nastavku. Izlazi iz softvera su uglavnom grafički prikazi i već su prikazani u poglavlju 7, tako da zbog sažetosti slike neće biti ponovo prikazivane u ovom prilogu, ali će biti ukazano koji rezultati se generišu u kom segmentu.

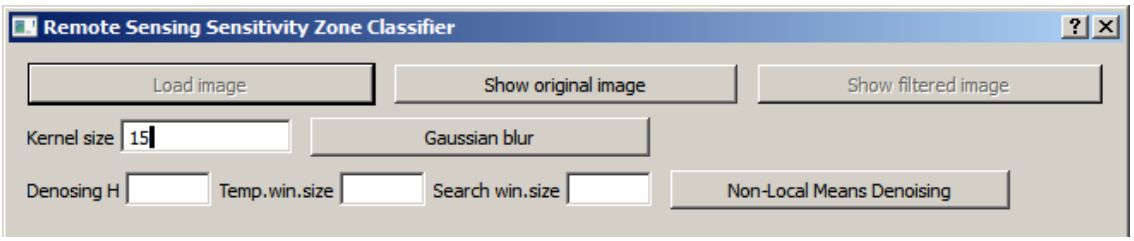


Slika A.1: Osnovni prozor softverskog rešenja

Segment 1 omogućuje učitavanje slike (Slika A.2) nad kojom sprovodimo analizu, zatim opciono predprocesiranje slike pomoću Gausovog zamućivanja ili Otklanjanja šuma pomoću nelokalnih sredina (Slika A.3). Takođe omogućuje prikaz originalne i na taj način obrađene slike (prikaz slike, poput na Slici 7.1.1).

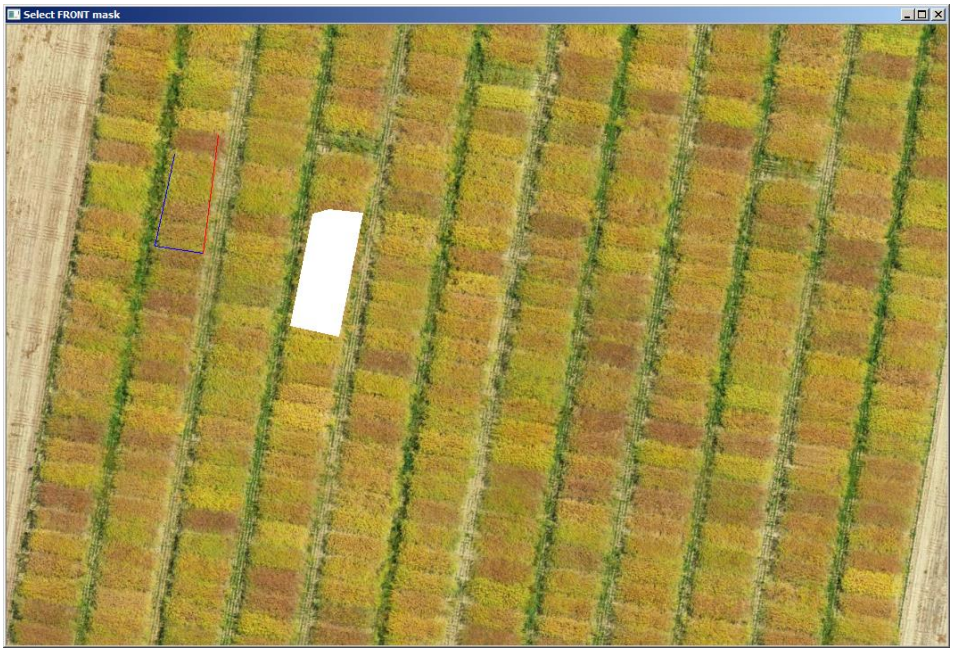


Slika A.2: Učitavanje slike za klasifikaciju

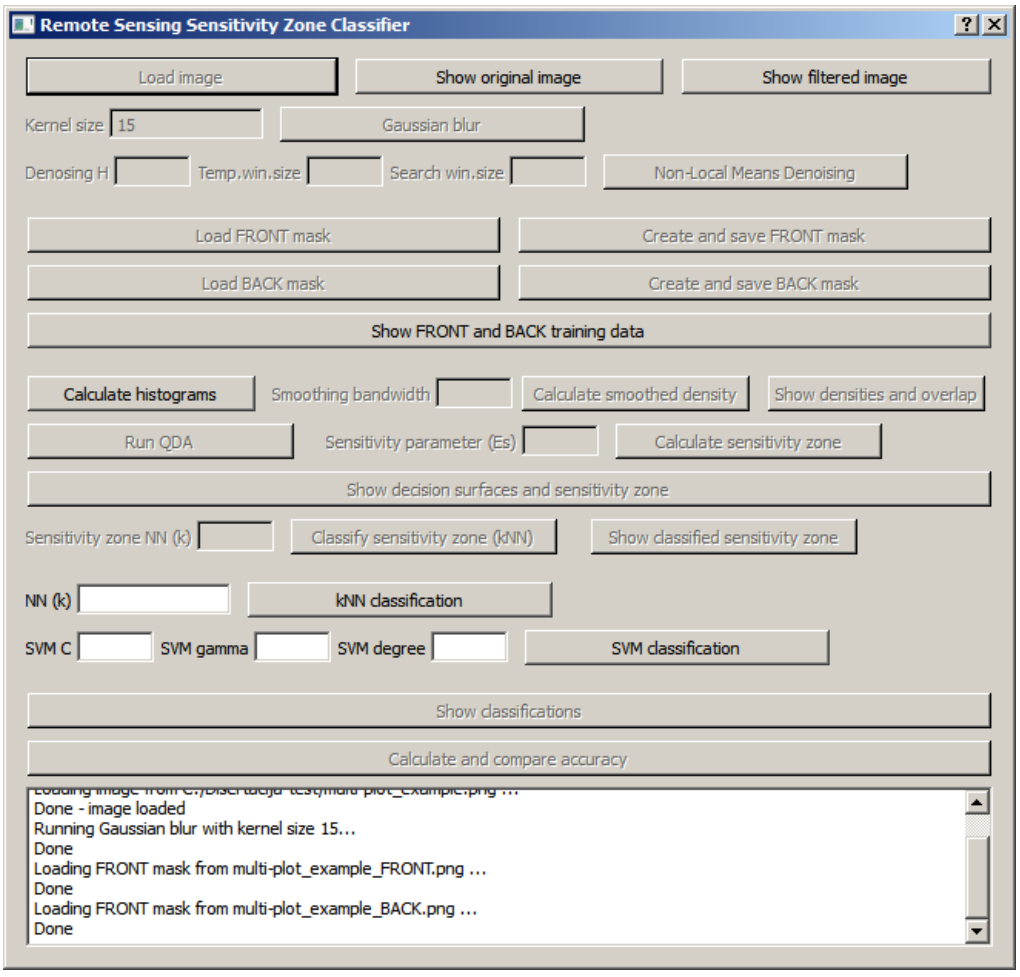


Slika A.3: Pokretanje Gausovog zamućivanja sa veličinom jezgra 15

Segmenet 2 učitavanje ranije pripremljenih maski za trening podatke dveju klasa (npr. useva i pozadine) – trening podaci se dobijaju uzimanjem podataka slike za oblasti koje su označene na masci. Ukoliko maske nisu pripremljene ranije, mogu se kreirati i sačuvati ovde pomoću posebno implementiranog modula za brzo označavanje i čuvanje delova slike (Slika A.4). Takođe, nakon učitavanja ili kreiranja maski za trening podatke (oblasti slike), omogućen je njihov prikaz, kao što je dato na Slici 7.1.2. Slika A.5 prikazuje stanje korisničkog interfejsa nakon učitavanja slike, obrade Gausovim zamućivanjem i nakon učitavanja/kreiranje maski za trening.

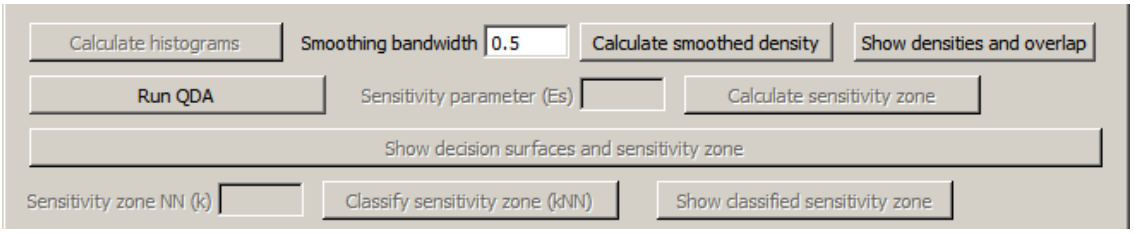


Slika A.4: Obeležavanje dva segmenta kao uzorka useva (već obeležen, belom bojom, i jedan čije je obeležavanje u toku – plave i crvena linija). Obeležavanje se vrši mišem.

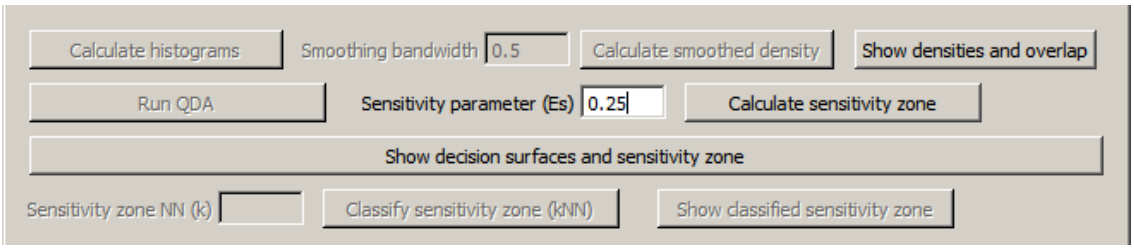


Slika A.5: Korisnički interfejs nakon učitavanja, predprocesiranja i trening podataka

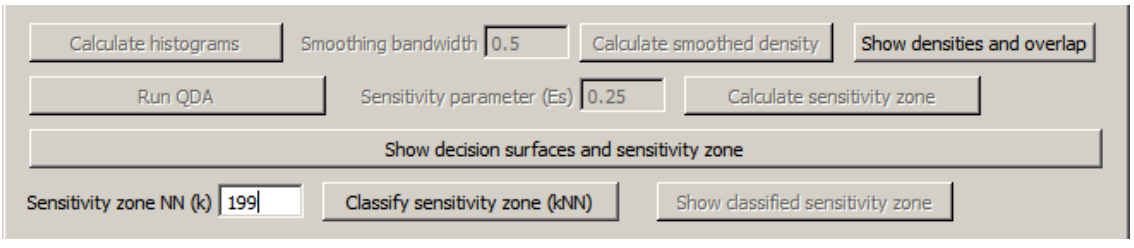
Segment 3 omogućuje pokretanje implementacije predložene metodologije. Prvo se računaju histogrami, zatim se opciono računa izglađeni histogram za zadati stepen izglađivanja (Slika A.6). Nakon toga mogu se dobiti prikazi histograma, zajedno sa izračunatim stepenom preklapanja, kao što je prikazano na Slikama 7.1.3 i 7.1.4. Potom se pokreće računanje posteriornih verovatnoća za piksele dveju klasa pomoću QDA i za zadati parametar ε_s se računa zona osetljivosti (Slika A.7). Nakon toga moguće je dobiti prikaze kako je dato na Slikama 7.1.5, 7.1.6 i 7.1.7. I konačno, sada se može pokrenuti klasifikacija piksela iz zone osetljivosti pomoću kNN metode, za zadati parametar k – broj suseda koje je potrebno naći (Slika A.8) i nakon toga može se dobiti prikaz kao na slici 7.1.8.



Slika A.6: Pokretanje izglađivanja gustine – histograma Gausovim jezgrom sa stepenom izglađivanja 0,5.

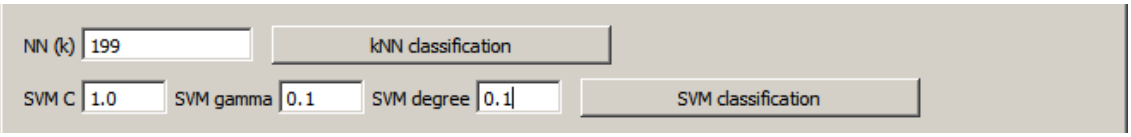


Slika A.7: Računanje zone osetljivosti za ε_s = 0,25.



Slika A.8: Klasifikacija piksela iz zone osetljivosti pomoću kNN metoda za k=199

Segment 4 omogućuje opciono pokretanje klasifikacije drugim metodama, kNN i SVM, kako bi se mogli uporediti rezultati klasifikovanja predloženom metodom sa rezultatima ostalih metoda u širokoj upotrebi. Pokretanje klasifikacije pomoću ove dve metode prikazano je na Slici A.9.



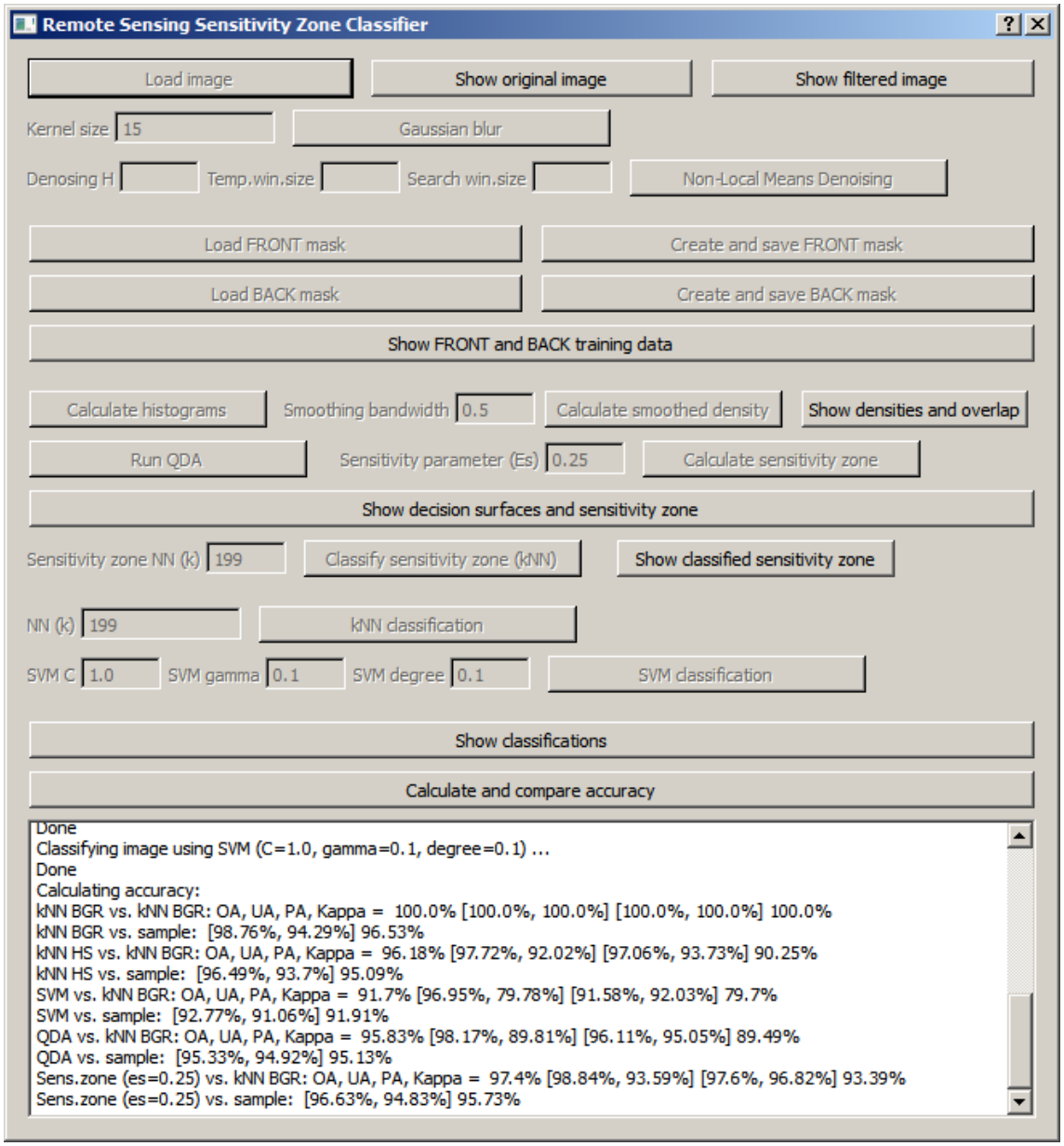
Slika A.9: Pokretanje kNN i SVM klasifikacije

Segment 5 najzad omogućuje prikaz svih prethodno izvršenih klasifikacija, čime se dobija prikaz kao na Slikama 7.1.9 i 7.1.10. I svakako ono što je među najvažnijim elementima, dugme na samom dnu služi za pokretanje izračuna tačnosti svih prethodno izvršenih klasifikacija, u odnosu na kNN klasifikaciju po 3 kanala (RGB) i u odnosu na snabdeveni uzorak (trening podatke). Time će softver (u delu za tekstualne poruke – Segment 6) prikazati podatke koji su dati u Tabeli 7.1.4.

Segment 6 služi da softver prikazuje poruke korisniku u delu za ispis teksta. Tu će biti prikazan svaki početak obrade, sa detaljima ulaznih parametara, zatim poruke o greškama ili neispravnim unosima (npr. da određeni parametar nije propisno uet) i na kraju tekstualne rezultate obrade podataka (kao što je izračunata tačnost klasifikacije). Nakon sprovedenih svih koraka korisnički interfejs će se naći u stanju kako je prikazano na Slici A.10.

Ono što je takođe važno napomenuti je da tokom izvrašavanja softver upisuje mnogo detaljnije poruke u automatski generisanu log datoteku, čiji naziv sadrži vremenski potpis kreiranja (godina-mesec-dan_časova-minuta-sekundi). Iz tako generisanog loga se potom mogu očitati dodatni detalji obrade, odakle su, između ostalog, preuzeti podaci prikazani u Tabelama 7.1.2 i 7.1.3. Primer jednog segmenta generisane log datoteke dat je na Slici A.11.

Još jedna važna karakteristika softverskog rešenja jeste da sve međurezultate, za čije računanje je potrebno netrivijalno vreme obrade od strane računara, zapisuje i čuva u datotekama, kako se iste obrade, za iste ulazne podatke i iste parametre obrade, ne bi morale izvršavati više puta, kad već daju iste rezultate. Svaka takva datoteka u svom nazivu sadrži vrednosti parametara, tako da prilikom pokretanja obrade se prvo traži datoteka sa takvim nazivom i ukoliko se pronađe – učitavaju se rezultati, ukoliko ne – izračunavaju se. Ovo značajno štedi vreme za sprovođenje testova, kada se testiranje vrši nad istim ulaznim podacima ali različitim parametrima, i sl.



Slika A.10: Stanje korisničkog interfejsa nakog izvršenih svih koraka

```
Start at 2018-May-20 06:48:43
Loading image: Advanta3.png... OK
img dimensions (696, 1055) 734280
blur filter_ks= 15
Pre-processed = PRE_GAUSSBLUR_KS-15

Load masks for training data (or create and save them if they do not exists):
Loading front mask from: Advanta3_FRONT_mask.png... OK
total_mask_front_pixels= 46889 6.39%
Loading back mask from: Advanta3_BACK_mask.png... OK
total_mask_back_pixels= 17228 2.35%
Total training sample= 64117 8.73%

Calculate histograms and densities:
Old hist sums (front, back):  46889 17228
New hist sums after scaling (front, back):  46889 46889
gauss_bw (gauss density kernel bandwidth)= 0.5
Trying to load Advanta3_PRE_GAUSSBLUR_KS-15_HSDENS_BW-0.5_FRONT.npy from a file... Not found, calculating and saving it now
Launching getDensGaussHS...
getDensGauss completed in 0:00:16.518370
Done
Trying to load Advanta3_PRE_GAUSSBLUR_KS-15_HSDENS_BW-0.5_BACK.npy from a file... Not found, calculating and saving it now
Launching getDensGaussHS...
getDensGauss completed in 0:00:06.560478
Done
Old dens sums (front, back):  46889 17228
New dens sums after scaling (front, back):  46889 46889

DA classification and sensitivity zone:
Total sample overlapped pixels= 3975
Total sample overlap rating =  4.43%
```

A.11: Primer dela zapisa u log datoteci

PRILOG B – UPOREDNI PRIKAZ REZULTATA STUDIJA SLUČAJA

Tabela B.1 prikazuje uporedni prikaz veličina slika, veličina uzoraka (useva i pozadine) i izračunatog stepena preklapanja uzoraka. Tabela B.2 daje uporedni prikaz vremena izvršavanja klasifikacije referentnim metodama a Tabela B.3 vremena izvršavanja klasifikacije metodom sa zonom osetljivosti, prema testiranim parametrima ε_s i zajedno sa dodatnim podacima o veličini zone osetljivosti i klasifikovanim pikselima iz zone osetljivosti. Tabela B.4 daje uporedni prikaz izmerene tačnosti, odnosno ukupnu tačnost i Kappa statistiku u odnosu na referentne podatke iz klasifikacije kNN RGB metodom i podatke uzoraka.

Tabela B.1: Detalji veličine slike i uzoraka (trening podataka)

Primer	Visina slike	Širina slike	Ukupno piksela	Pikseli za trening-usev	Pikseli za trening-pozadina	Stepen preklapanja
Ogledna polja uljane repice	696	1.055	734.280	46.889 (6,39%)	17,228 (2,35%)	4,43% (3.975 piksela)
Kukurz u fazi V5	2.149	2.772	5.957.028	414 (0,01%)	839 (0,01%)	0,06% (1 piksel)
Plantaža manga	3.815	6.823	26.029.745	23.531 (0,09%)	66.469 (0,26%)	1,01% (1.331 piskel)

Tabela B.2: Uporedna vremena klasifikacije referentnim metodama (minuti:sekunde)

Primer	Vreme za kNN RGB metod	Vreme za kNN HS metod	Vreme za SVM metod	Vreme za QDA metod (bez z.o.)
Ogledna polja uljane repice	04:34	00:42	00:08	00:00
Kukurz u fazi V5	07:03	04:40	00:28	00:01
Plantaža manga	44:53	06:24	04:26	00:15

Tabela B.3: Vreme i broj piksela klasifikacije sa zonom osetljivosti

Primer	Parametar ε_s	Veličina zone osetljivosti	Vreme trajanja (minuti:sekunde)	Pikseli iz ZO klasifikovani kao usev
Ogledna polja uljane repice	0,25	27.799 (3,79%)	00:15	18.419 (66,26%)
	0,35	44.751 (6,09%)	00:15	29.411 (65,72%)
	0,40	57.242 (7,80%)	00:16	36.989 (64,62%)
Kukurz u fazi V5	0,25	283.789 (4,76%)	01:40	127.974 (45,09%)
	0,35	359.789 (6,04%)	02:18	178.600 (49,64%)
	0,40	425.929 (7,15%)	02:55	217.492 (51,06%)
Plantaža manga	0,25	545.366 (2,10%)	01:33	208.939 (38.31%)
	0,35	907.547 (3.49%)	02:54	353.076 (38.90%)
	0,40	1.217.861 (4.68%)	03:16	488.797 (40.14%)

Tabela B.4: Izmerena tačnosti klasifikacije

Primer	Primenjeni metod	Ukupna tačnost (ref. kNN RGB)	Kappa statistika (ref. kNN RGB)	Ukupna tačnost (ref. trening uzorak)
Ogledna polja uljane repice	kNN RGB	100,00%	100,00%	96,53%
	kNN HS	96,18%	90,25%	95,09%
	SVM	91,70%	79,70%	91,91%
	QDA bez ZO	95,83%	89,49%	95,13%
	Sa ZO (ε_s = 0,25)	97,40%	93,39%	95,73%
	Sa ZO (ε_s = 0,35)	98,03%	94,98%	96,06%
	Sa ZO (ε_s = 0,40)	98,45%	96,04%	96,20%
Kukurz u fazi V5	kNN RGB	100.00%	100.00%	97.04%
	kNN HS	56.28%	14.93%	91.29%
	SVM	74.71%	47.57%	81.16%
	QDA bez ZO	54.77%	12.10%	99.34%
	Sa ZO (ε_s = 0,25)	56.80%	16.11%	99.46%
	Sa ZO (ε_s = 0,35)	57.46%	17.41%	99.52%
	Sa ZO (ε_s = 0,40)	57.97%	18.44%	99.53%
Plantaža manga	kNN RGB	100.00%	100.00%	99.13%
	kNN HS	97.46%	94.10%	97.83%
	SVM	96.44%	91.81%	97.35%
	QDA bez ZO	97.04%	93.23%	98.17%
	Sa ZO (ε_s = 0,25)	97.97%	95.33%	98.40%
	Sa ZO (ε_s = 0,35)	98.35%	96.21%	98.45%
	Sa ZO (ε_s = 0,40)	98.54%	96.65%	98.52%

BIOGRAFIJA

Milan Dobrota je rođen 19.07.1981. u Kninu. Osnovno školovanje započinje u Zagrebu a zaršava u Batajnici, uz Vukovu diplomu. U Beogradu, srednju elektroteničku školu „Nikola Tesla“ pohađa od 1995. do 1999. godine, kada i upisuje Fakultet organizacionih nauka, smer Informacioni sistemi i tehnologije. Tokom studija bio je angažovan kao demonstrator vežbi na predmetu Principi programiranja, među osnivačima je Udruženja studenata informatike FONIS i bio je član Fudbalske ekipe fakulteta čitavim tokom studija.

Profesionalnu karijeru započinje sa 20 godina, kada je u periodu od 2001. do 2003. bio zaposlen u lokalnom ogranku kanadske kompanije Optiwave Systems Inc. na poslovima softver inženjera za razvoj sistema optičkih komunikacija. Nakon toga, do 2005. nastavlja bavljenjem razvojem softvera kao nezavisni preduzetnik, da bi se 2005. zaposlio u francuskoj kompaniji Interex, balkanskom ogranku francuskog lanca supermarketa Intermarché, na poziciji Analitičar poslovanja u centralnom odeljenju za informacione tehnologije. Godinu dana kasnije, 2006, preuzima funkciju Projektnog vođe u istom sektoru. U januaru 2007. diplomira na Fakultetu organizacionih nauka pod mentorstvom prof. dr Siniše Vlajića, sa prosečnom ocenom studija 8.16 i ocenom 10 na Diplomskom radu. Te godine, 2007, u istoj kompaniji preuzima poziciju direktora Sektora za informacione tehnologije za region Balkana, a od 2009. uzima i ulogu u regionalnom Upravnom odboru kompanije, kao jedini član bez francuskog porekla. Iste godine nastavlja i preduzetničku karijeru kao suosnivač kompanije Golive Information Systems Consulting. Godine 2011. upisuje Doktorske studije na Fakultetu organizacionih nauka u Beogradu, studijski program Informacioni sistemi i menadžment – Menadžment.

Godine 2012. napušta prethodna profesionalna angažovanja da bi osnovao kompaniju LOGIT, koja danas broji 40 inženjera, u kancelarijama u Beogradu, Ploiestiu (Rumunija) i Sarajevu (BIH), angažovanih na različitim internacionalnim softverskim, konsultantskim i istraživačko-razvojnim projektima. U kompaniji LOGIT je do danas i zaposlen kao direktor grupe. Početkom 2015. osniva „spin-off“ projekat i poslovni poduhvat, danas poznat pod brendom Agremo, koji se bavi razvojem softverskog rešenja za automatizaciju analiza slika prikupljenih daljinskim uzorkovanjem, prvenstveno pomoću bespilotnih letelica (dronova), za svrhe precizne poljoprivrede i efikasnijeg uzgoja useva. Rešenje je danas prepoznato kao jedno od svetskih lidera inovacije u oblasti analize podataka prikupljenih bespilotnim letelicama u poljoprivredi, na globalnom nivou.

Od 2015. je takođe angažovan i kao konsultant Evropske banke za obnovu i razvoj, u svojstvu samostalnog industrijskog savetnika a od iste godine do kraja 2016. i kao ko-direktor beogradskog ogranka Founder Institue-a, programa obuke preduzetnika i podrške startup kompanijama iz „Silikonske doline“, SAD. Tokom dosadašnje

profesionalne karijere učestvovao je na preko 27 velikih projekata, u svojstvu inženjera, menadžera ili konsultanta, od kojih je velika većina internacionalnog karaktera. Tokom dosadašnjeg toka Doktorskih studija u saradnji sa drugim autorima objavio više naučnih radova u zbornicima domaćih i međunarodnih konferencija, te domaćem kao i u vrhunskom međunarodnom naučnom časopisu.

PRILOG 1

Izjava o autorstvu

Potpisani: Milan Dobrota

broj indeksa: 5030/2011

Izjavljujem

da je doktorska disertacija pod naslovom:

„Statistički pristup definisanju zone osetljivosti u metodama daljinskog uzorkovanja“

- rezultat sopstvenog istraživačkog rada,
- da predložena disertacija u celini ni u delovima nije bila predložena za dobijanje bilo koje diplome prema studijskim programima drugih visokoškolskih ustanova,
- da su rezultati korektno navedeni i
- da nisam kršio autorska prava i koristio intelektualnu svojinu drugih lica.

U Beogradu, _____

Potpis doktoranda

_____

PRILOG 2

Izjava o istovetnosti štampane i elektronske verzije doktorskog rada

Ime i prezime autora: Milan Dobrota

Broj indeksa: 5030/2011

Studijski program: Informacioni sistemi i menadžment, Menadžment

Naslov rada: „Statistički pristup definisanju zone osetljivosti u metodama daljinskog uzorkovanja“

Mentor: Prof. dr Zoran Radojičić, redovni profesor, Univerziteta u Beogradu, FON-a

Potpisani: Milan Dobrota

Izjavljujem da je štampana verzija mog doktorskog rada istovetna elektronskoj verziji koju sam predao za objavljivanje na portalu **Digitalnog repozitorijuma Univerziteta u Beogradu**.

Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog zvanja doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu.

U Beogradu, _____

Potpis doktoranda

_____

PRILOG 3

Izjava o korišćenju

Ovlašćujem Univerzitetsku biblioteku "Svetozar Marković" da u Digitalni repozitorijum Univerziteta u Beogradu unese moju doktorsku disertaciju pod naslovom:

„Statistički pristup definisanju zone osetljivosti u metodama daljinskog uzorkovanja“

koja je moje autorsko delo.

Disertaciju sa svim prilozima predao sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju pohranjenu u Digitalni repozitorijum Univerziteta u Beogradu mogu da koriste svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons) za koju sam se odlučio.

1. Autorstvo
2. Autorstvo – nekomercijalno
3. Autorstvo – nekomercijalno – bez prerade
4. Autorstvo – nekomercijalno – deliti pod istim uslovima
5. Autorstvo – bez prerade
6. Autorstvo – deliti pod istim uslovima

U Beogradu, _____

Potpis doktoranda

_____