UPORABA STROJNEGA UČENJA ZA DOLOČITEV POPLAVLJENIH OBMOČIJ – PRIMER POPLAV V SELŠKI DOLINI LETA 2007

APPLICATION OF DATA MINING FOR DETERMINATION OF FLOODED AREAS – SELŠKA VALLEY 2007 FLOODS CASE STUDY

Peter Lamovec, Krištof Oštir

UDK: 004.431.2:556.166:528.7:(497.4 Selška dolina) **Klasifikacija prispevka po COBISS-u: 1.01**

V prispevku je obravnavana uporabnost postopkov strojnega učenja pri ugotavljanju poplavljenih območij v zgornjem delu Selške doline, ki jo je 18. 9. 2007 prizadelo hudourniško deževje. Hitro prepoznavanje poplavljenih območij je ključnega pomena za učinkovito reševanje in povračilo škode od zavarovalnic. Pri tem so zelo uporabni satelitski posnetki, saj omogočajo hitro določitev poplavljenih območij, tudi če so prizadeti zelo veliki predeli. Za prepoznavanje poplavljenih območij v Selški dolini so bile uporabljene tehnike strojnega učenja z različnimi vhodnimi podatki: satelitski posnetek SPOT (multispektralni in pankromatski), indeks NDVI, relief in njegovi izdelki (nadmorska višina, naklon, ukrivljenost), oddaljenost od vodotokov in raba tal. Učni vzorci, ki so bili uporabljeni za oblikovanje modela klasifikacije, so vsebovali 400, 255 oziroma 49 vzorčnih točk.

KLJUČNE BESEDE

strojno učenje, odločitveno drevo, klasifikacija, satelitski posnetki, poplave

This paper discusses the usefulness of machine-learning procedures for determining flooded areas in the upper part of the Selška valley. The area was affected by torrential rains on 18.9.2007. Rapid identification of flooded areas is essential for effective implementation of rescue operations and damage assessments. In this case, satellite images are very useful because they enable quick identification of flooded areas even in very large areas. To determine the flooded areas, machine learning techniques were applied to different input data. SPOT satellite image (multispectral and panchromatic), NDVI index, relief and its derivatives (altitude, slope, curvature), distance from rivers and land use were used. The learning samples consisted of 400, 255 and 49 sample points, which were used to build three different classification models.

KEY WORDS

machine learning, decision trees, classification, satellite images, floods

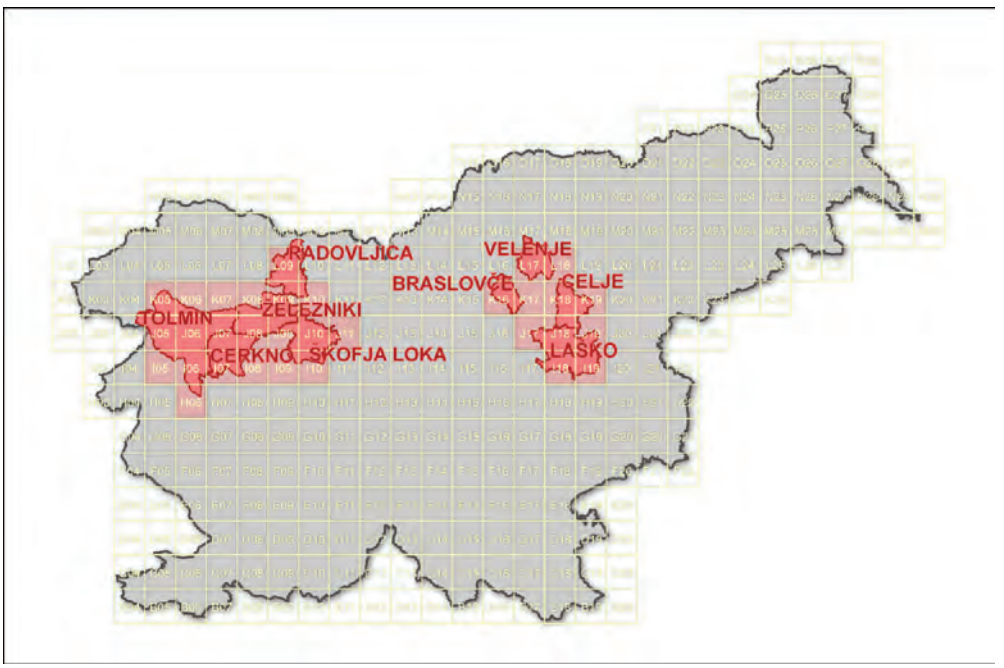
1 UVOD

Naravne nesreče so vse pogostejše in vse bolj obsežne. So posledica človekovega ravnanja ali pa zgolj običajnih naravnih pojavov. Njihovo preprečevanje, izogibanje ter omejevanje je pomembno za ohranjanje kakovosti življenjskega okolja. Širjenje življenjskega prostora in poseganje na manj primerna območja za različne človekove dejavnosti povečujeta potrebo po kakovostnem določanju ogroženih območji. V Sloveniji so med naravnimi nesrečami pogoste zlasti poplave in plazovi. Preprečevanje oziroma zmanjševanje njihovih posledic je zato nujno (Vodišek, 2009).

Pri reševanju, analiziranju in ukrepanju ob naravnih nesrečah imajo zelo pomembno vlogo prostorski podatki in sodobna tehnologija GIS. Za določanje ogroženih in poškodovanih območij so uporabne zlasti metode, ki omogočajo zajem prizadetega območja v celoti in v čim krajšem času po naravni nesreči. V zadnjem desetletju so se podatkom, pridobljenim s klasičnimi terestričnimi opazovanji in opazovanji iz zraka, pridružili še podatki daljinskega zaznavanja, predvsem satelitski posnetki. Prednost slednjih je zlasti zajem zelo obsežnih območij na zemeljskem površju v zelo kratkem času.

V članku je obravnavan primer poplav iz septembra 2007, ki so prizadele del zahodne in vzhodne Slovenije. Na prizadetih območjih je bilo poškodovanih na stotine hiš. Narasle reke in potoki so odplavile veliko mostov in avtomobilov. Neurje, v katerem je v intervalu od 6 do 12 ur padlo več kot 100 l/m^2 dežja (ARSO, 2008), je zahtevalo tudi šest smrtnih žrtev. Največja škoda je bila ugotovljena v občini Železniki. Druga prizadeta območja zahodne Slovenije so bila naselja Cerkno, Bohinjska Bistrica in Kropa ter soteska Baška grapa. Na vzhodu so bila zelo prizadeta območja ob reki Paki, Bolski in ob spodnjem toku reke Savinje (Pehani et al., 2008).

Manj kot dan po neurju, 19. septembra 2007, je bil aktiviran mednarodni program Vesolje in velike nesreče (Space and Major Distasters), ki so ga vesoljske agencije ustanovile za učinkovito uporabo satelitske tehnologije pri naravnih in drugih nesrečah. Program je bil že uporabljen pri številnih poplavah, požarih, potresih, ciklonih, razlitjih nafte, vulkanskih izbruhih in drugih nesrečah. Njegova prednost je hitra pridobitev satelitskih posnetkov in njihova dostava ustanovam, ki se ukvarjajo z nadaljnjo obdelavo posnetkov. S programom so bila izvedena tudi

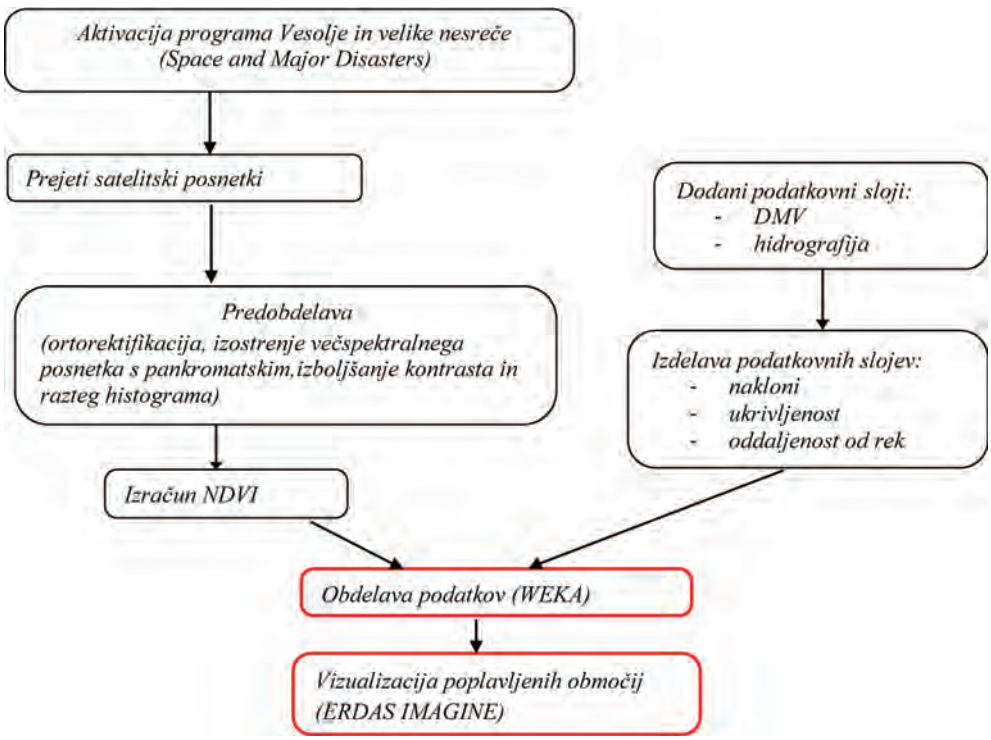


Slika 1: Območja, ki jih je neurje najbolj prizadelo in so bila opazovana ob aktivaciji programa Vesolje in velike nesreče (Inštitut za antropološke in prostorske študije, ZRC SAZU, 2007).

satelitska snemanja prizadetega območja na vzhodu in zahodu Slovenije (slika 1, označene so najbolj prizadete občine).

Prva snemanja prizadetih območij so potekala 21. 9. 2007 (tri dni po nesreči). Na šestnajstih snemanjih s satelitskimi sistemi SPOT, Envisat, Radarsat, IRS in Formosat je bilo skupaj narejenih 25 satelitskih posnetkov. Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti ZRC SAZU je posnetke prejel 25. 9. 2007 (sedem dni po nesreči) (Pehani et al., 2008). Izbrali so najbolj uporabne in jih obdelali. Končni rezultat je bila karta poplavljenih območij, ki so jo poslali Upravi RS za zaščito in reševanje nekaj dni po prejemu posnetkov.

Ker je hitrost pri izdelavi karte odločilnega pomena, smo na Inštitutu za antropološke in prostorske študije ZRC SAZU poleg drugih metod preizkusili tudi uporabo strojnega učenja na primeru določitve poplavljenih območij. V prvem koraku smo pridobili in pripravili vse podatke, za katere smo predvidevali, da lahko pomembno vplivajo na določitev poplavljenih površin. Uporabili smo multispektralni in pankromatski satelitski posnetek SPOT, vegetacijski indeks NDVI (ang. Normalized Difference Vegetation Index), relief in njegove izdelke (nadmorska višina, naklon, ukrivljenost), oddaljenost od vodotokov in rabo tal. Te podatke smo vključili v postopek strojnega učenja s programom WEKA (WEKA, 2010). Strojno učenje je omogočilo določitev dejavnikov in njihovih mejnih vrednosti, ki odločilno vplivajo na poplavljenost. Na koncu smo model, ki smo ga pridobili s strojnim učenjem na podlagi vzorčnih točk, uporabili za določitev



Slika 2: Shematski prikaz poti od aktivacije programa Vesolje in velike nesreče do izdelave karte poplavljenih območij

poplavljenih površin na celotnem obravnavnem območju. Ves postopek, od aktivacije programa Vesolje in velike nesreče do končnega izdelka – karte poplavljenih območij, je prikazan na sliki 2.

2 RUDARJENJE PODATKOV S TEHNIKAMI STROJNEGA UČENJA

Količina podatkov se zaradi sodobne opreme za njihovo pridobivanje in shranjevanje zelo hitro povečuje. Po drugi strani pa človeška sposobnost za njihovo obdelavo ostaja bolj ali manj nespremenjena (Maimon in Rokach, 2005). Tako se je razvilo veliko novih metod za obdelavo podatkov in odkrivanje pomembnih informacij iz večjih količin podatkov.

Rudarjenje podatkov (ang. data mining – DM) je glavni korak v celotnem procesu odkrivanja znanja iz podatkov (ang. knowledge discovery in databases – KDD). Temelji na uporabi računskih tehnik, tj. algoritmov, ki so realizirani kot računalniški programi za iskanje zakonitosti oziroma vzorcev v podatkih. Drugi koraki v postopku odkrivanja znanja iz podatkov so povezani s pripravo podatkov za rudarjenje in ovrednotenjem odkritih vzorcev (rezultatov rudarjenja) (Džeroski, 2001). Odkrivanje znanja iz podatkov je bilo na začetku opredeljeno kot poseben način iskanja prej nepoznanih in potencialno uporabnih informacij iz podatkov (Frawley, Piatetsky-Shapiro in Matheus, 1991). Novejša različica opredelitve pravi, da je odkrivanje znanja iz podatkov poseben proces prepoznavanja veljavnih, prej nepoznanih, potencialno uporabnih in končno razumljivih vzorcev/modelov iz podatkov (Fayyad, Piatetsky-Shapiro in Smyth, 1996).

Pri delu z velikimi količinami podatkov si pomagamo s tehnikami strojnega učenja (ang. machine learning). Strojno učenje je področje umetne inteligence, na katerem se razvijajo tehnike, ki omogočajo računalnikom oziroma strojem, da se učijo na primerih. Močno se opira na statistiko, vendar se v nasprotju z njo nekoliko bolj ukvarja s samimi algoritmi in računskimi operacijami (Polanec, 2006). Osnovna ideja strojnega učenja je, da stroj (računalnik) naučimo »razumeti« določeno vrsto podatkov. Za učenje uporabimo manjše število znanih primerov (vzorec), na podlagi katerih bo stroj znal analizirati celotno množico podatkov. Človek samostojno ni sposoben analizirati tolikšne količine podatkov oziroma bi bilo to časovno preveč potratno.

Ločimo dva načina strojnega učenja, to sta nadzorovano in nenadzorovano učenje. Pri nadzorovanem učenju modeliramo funkcijo na podlagi učne množice vzorcev. Učna množica vzorcev mora imeti poleg učnih vzorcev (vhodnih podatkov) opredeljene tudi oznake razredov (želeni izhod sistema). Izhod sistema je lahko zvezno področje vrednosti (govorimo o regresiji) ali enolična oznaka razreda, ki mu pripada dani vzorec (govorimo o klasifikaciji). Pri nenadzorovanem učenju imamo podane samo značilke vzorcev, nimamo pa podanih razredov, ki jim pripadajo. Naloga učnega algoritma je določiti razrede, tj. skupine vzorcev, ki so si med seboj čim bolj podobni (Polanec, 2006).

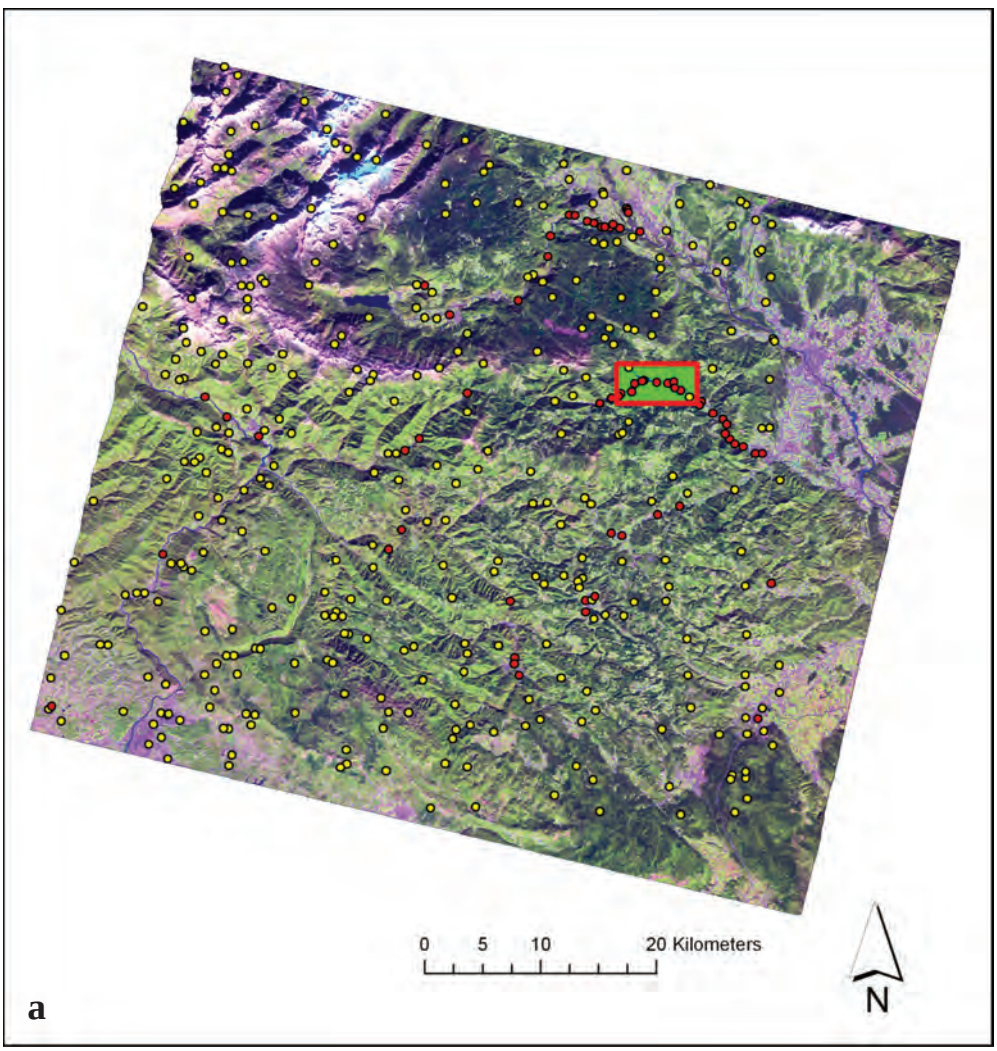
V članku je predstavljen primer nadzorovanega učenja – klasifikacije. Pri klasifikacijah gre za napovedovanje vrednosti enega polja na podlagi vrednosti drugih polj. Ciljno polje je imenovano razred (odvisna spremenljivka). Druga polja so atributi (neodvisne spremenljivke). Razred je pri klasifikaciji diskretna spremenljivka (končno število nominalnih vrednosti). Rezultat naloge je model, ki ga lahko uporabimo za napovedovanje razredov novega niza podatkov (Džeroski, 2001).

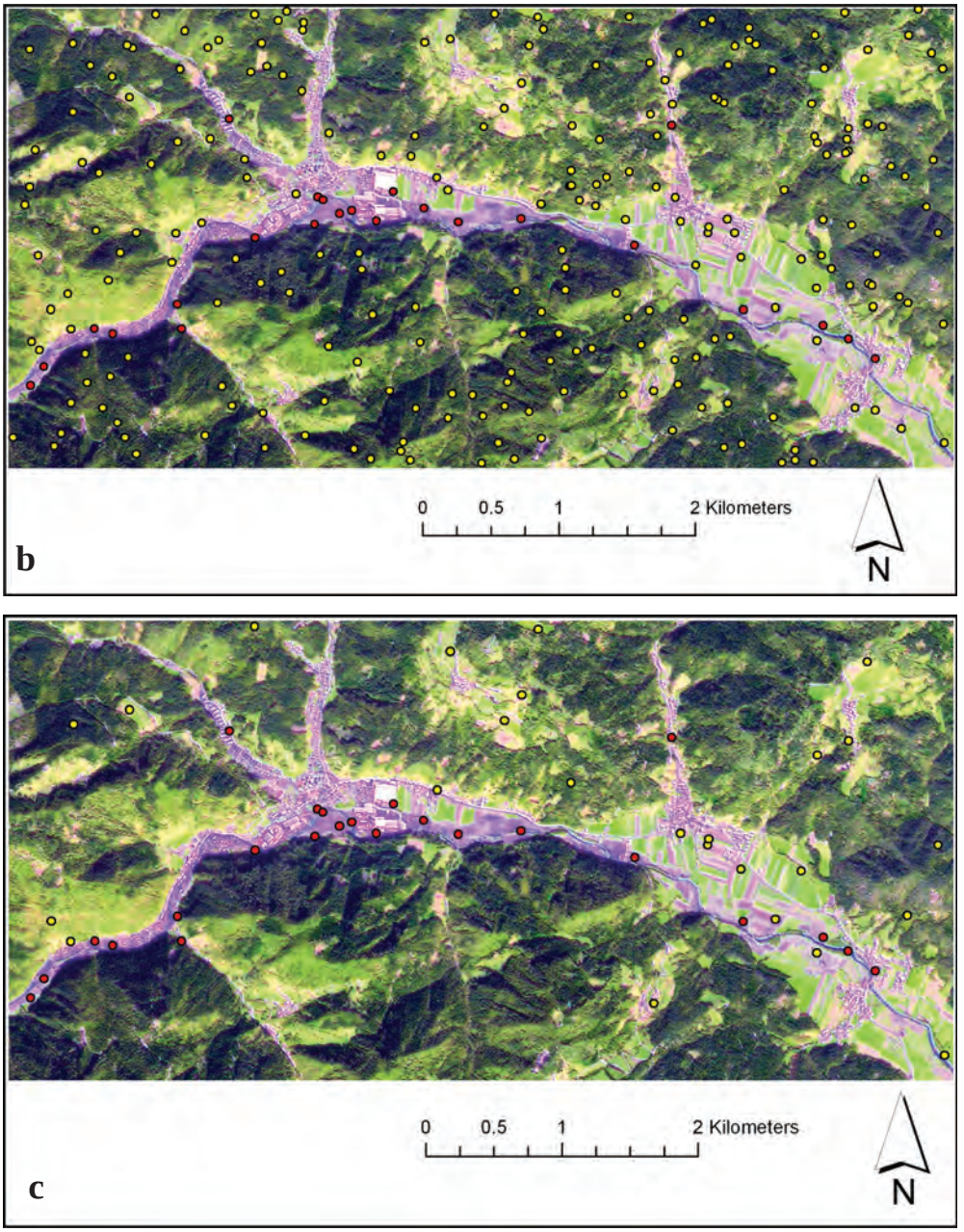
3 POSTOPEK DOLOČITVE POPLAVLJENIH POVRŠIN

Postopek klasifikacije poplavljenih površin je sestavljen iz naslednjih korakov:

- določitve obravnavanega območja,
- zbiranja podatkov za učenje sistema,
- opredelitve značilk vzorca,
- izvedbe algoritma učenja na učni množici vzorcev,
- ocene uspešnosti strojnega učenja.

Predstavljeni primer prikazuje postopek prepoznavanja poplavljenih površin na območju Železnikov v Selški dolini. Podatki, ki jih uporabimo za učenje sistema, morajo biti značilni za območje, na katerem bo model deloval. Zelo pomembna je uporaba dovolj velikega vzorca.





Slika 3: Prikaz razporeditev vzorčnih točk na satelitskem posnetku SPOT. V oklepajih je navedeno razmerje med točkami, ki ležijo na nepoplavljenih (rumeno obarvane točke) oziroma poplavljenih površinah (rdeče obarvane točke). Slika 3a: 400 (336 : 64) vzorčnih točk, razporejenih na širšem območju od obravnavanega; slika 3b: 255 (231 : 24) vzorčnih točk; slika 3c: 49 (25 : 24) vzorčnih točk.

Velikost vzorca precej vpliva na reprezentativnost populacije in posledično na uspešnost klasifikacije. Majhni vzorci so precej manj zanesljivi in običajno slabše opisujejo populacijo kot

veliki. V obravnavanem primeru smo uporabili vzorce treh različnih velikosti (vzorce z 400, 255 in 49 točkami) in z enakim naborom atributov. Prostorska razporeditev točk posameznega vzorca je prikazana na sliki 3. Prvi vzorec (slika 3a) vsebuje največje število točk (400), ki ležijo na širšem območju od obravnavanega. Vzorca na sliki 3b in 3c sta manjša in vsebujeta le točke, ki ležijo znotraj obravnavanega območja. Ta vzorca imate identične točke. Vzorcu z 49 točkami je v primerjavi z vzorcem z 255 točkami le zmanjšano število točk, ki ležijo na nepoplavljenih območjih. To storimo za pridobitev boljšega razmerja med točkami na poplavljenih in nepoplavljenih površinah.

Vhodni podatki in njihove značilke so običajno zbrani v tabeli. V vrsticah so predstavljeni posamezni primeri (objekti), v stolpcih pa njihove lastnosti (atributi). Številčni atributi z realnimi vrednostmi so zvezni, atributi z nominalnimi vrednostmi pa diskretni. V obravnavanem primeru so bili atributi pridobljeni iz satelitskega posnetka SPOT, digitalnega modela višin DMV 12,5 in vektorskega sloja vodotokov. Glavni del podatkov je bil pridobljen s satelitskim posnetkom SPOT, ki je visokoločljivostni satelitski sistem za opazovanje Zemlje iz vesolja. Uporabili smo multispektralni posnetek v ločljivosti 10 m in pankromatski posnetek v ločljivosti 2,5 m.

Pravilnost modela je precej odvisna od izbire atributov. Pri tem naj bi za oblikovanje modela uporabili čim bolj homogeno populacijo, ki zagotavlja čim bolj enakomerno zastopanost atributov različnih vrednosti. Merilo čim manjših razlik med deleži vzorčnih točk s posameznimi vrednostmi atributov v obravnavanem primeru ni najbolje izpolnjeno, kar omogoča izboljšave pri oblikovanju modela klasifikacije.

Točke vseh treh vzorcev so opisane z desetimi atributi. Vzorca, ki vsebujeta le točke z obravnavanega območja, imata dodaten atribut »Raba tal«. Atributi, z njihovimi statistikami (maksimalna vrednost, minimalna vrednost, povprečna vrednost in standardni odklon), so predstavljeni v preglednicah 1, 2 in 3.

Atributi – diskretni	**Št. razredov**	**Razred (frekvenca)**		
poplavljeno	2	0 (336), 1 (64)		
Atributi – zvezni	**Min.**	**Maks.**	**Povprečje**	**Stan. odklon**
B1	62	255	100	31
B2	42	255	86	41
B3	18	255	94	33
B4	19	221	85	33
NDVI	–0,54	0,42	0,05	0,21
višina	48,92	2348,8	704.5	447,3
naklon	0,01	68,28	17,07	13,92
ukrivljenost	0,05	56,31	9,81	9,26
oddaljenost od vodotokov	0	255	147	101

Preglednica 1: Uporabljeni atributi in njihove statistike za učni vzorec s 400 točkami

Vrednosti atributa »poplavljeno« označujejo nepoplavljena (0) oziroma poplavljena območja (1). Vrednosti atributa »raba tal« pa predstavljajo naslednje vrste rab: 1 – gozd, 2 – grmičevje, 3 – ekstenzivna travniška raba, 4 – obdelane kmetijske površine, 5 – pozidana in sorodna

zemljišča, 6 – reka. Oba atributa sta bila določena z vizualno interpretacijo multispektralnega satelitskega posnetka SPOT. Na njem je zaradi podobnega odboja elektromagnetnega valovanja od poplavljenih površin in površin z določenimi vrstami rabe težko zagotoviti enakomerno porazdeljenost vseh vrst rabe tako med poplavljenimi kot nepoplavljenimi vzorčnimi točkami. Odboj, podoben odboju na poplavljenih površinah, se pojavlja zlasti na obdelanih kmetijskih in urbanih površinah. V vzorec so uvrščene samo točke, za katere je vrednost atributa poplavljenosti nedvoumna. Nobena od poplavljenih vzorčnih točk ne leži na kmetijskih površinah. Tako je zagotovljena pravilnost učnega vzorca, ne pa tudi enakomernost porazdelitve atributa rabe tal v vzorcu. To velja za vzorčne točke na poplavljenih območjih, medtem ko je njihova enakomernost na nepoplavljenih površinah dobra. Neenakomerna porazdelitev vrednosti atributa med vzorčnimi točkami je prispevala k temu, da je bil atribut ocenjen kot nepomemben za klasifikacijo.

Atributi – diskretni	**Št. razredov**	**Razred (frekvenca)**		
poplavljeno	2	0 (231), 1 (24)		
raba tal	6	1(185), 2(10), 3(22), 4(6), 5(27), 6(5)		
Atributi – zvezni	**Min.**	**Maks.**	**Povprečje**	**Stan. odklon**
B1	46	245	95	32
B2	33	250	74	40
B3	38	180	86	29
B4	24	179	69	31
NDVI	–0,42	0,40	0,10	0,19
višina	401,3	947,0	579,3	122,3
naklon	0,31	46,26	22,05	12,58
ukrivljenost	0,16	47,61	10,63	8,45
oddaljenost od vodotokov	0	255	119	84

Preglednica 2: Uporabljeni atributi in njihove statistike za učni vzorec z 255 točkami

Atributi – diskretni	**Št. razredov**	**Razred (frekvenca)**		
poplavljeno	2	0 (25), 1 (24)		
raba tal	6	1(8), 2(5), 3(12), 4(5), 5(14), 6(5)		
Atributi – zvezni	**Min**	**Maks**	**Povprečje**	**Stan. odklon**
B1	59	245	131	42
B2	45	250	121	51
B3	38	180	99	29
B4	29	179	97	32
NDVI	–0,42	0,40	–0,08	0,26
višina	401,3	762,6	501,9	106,5
naklon	0,31	39,02	9,93	11,50
ukrivljenost	0,16	47,61	8,29	9,98
oddaljenost od vodotokov	0	255	73	83

Preglednica 3: Uporabljeni atributi in njihove statistike za učni vzorec z 49 točkami

Atributi »B1«, »B2«, »B3«, »B4« pomenijo vrednosti kanalov multispektralnega satelitskega posnetka SPOT. Kanal B1 zaznava odbito valovanje zelene svetlobe, kanal B2 zaznava rdečo

Kanal	Vrednosti EMV
B1	0,50–0,59 µm
B2	0,61–0,68 µm
B3	0,79–0,89 µm
B4	1,58–1,75 µm

Preglednica 4: Pregled kanalov multispektralnega posnetka SPOT in valovnih dolžin elektromagnetnega valovanja, ki jih zaznavajo posamezni kanali.

svetlobo, kanal B3 bližnjo infrardečo svetlobo in kanal B4 kratkovalovno infrardečo svetlobo. Njihove valovne dolžine so prikazane v preglednici 4.

Vegetacijski indeks »NDVI« je enostaven numerični indikator zelene vegetacije. Izračunan je na podlagi zaznanega odbitega valovanja rdeče in bližnje infrardeče svetlobe. Negativne vrednosti indeksa pomenijo vodne površine (vrednosti blizu –1), vrednosti blizu nič pomenijo neplodne površine, kot so skale, pesek, sneg (–0,1–0,1), nižje pozitivne vrednosti pomenijo grmičevje, travnike (0,2–0,4) ter višje pozitivne vrednosti zmeren in tropski deževni gozd (vrednosti blizu 1).

Enačba za izračun indeksa NDVI:

$$NDVI = \frac{bližnja\_IR - R}{bližnja\_IR + R}$$

Atribut »višina« določa nadmorsko višino. Podatki o višinah so bili pridobljeni iz digitalnega modela višin DMV 12,5. Iz višin sta izračunana dodatna atributa »naklon« in »ukrivljenost«, ki opisujeta površje. Nakloni so izraženi v stopinjah in lahko zavzemajo vrednosti med 0° in 90°. Vrednost 0° pomeni povsem ravno površje (gladina morja) in vrednost 90° navpično površje.

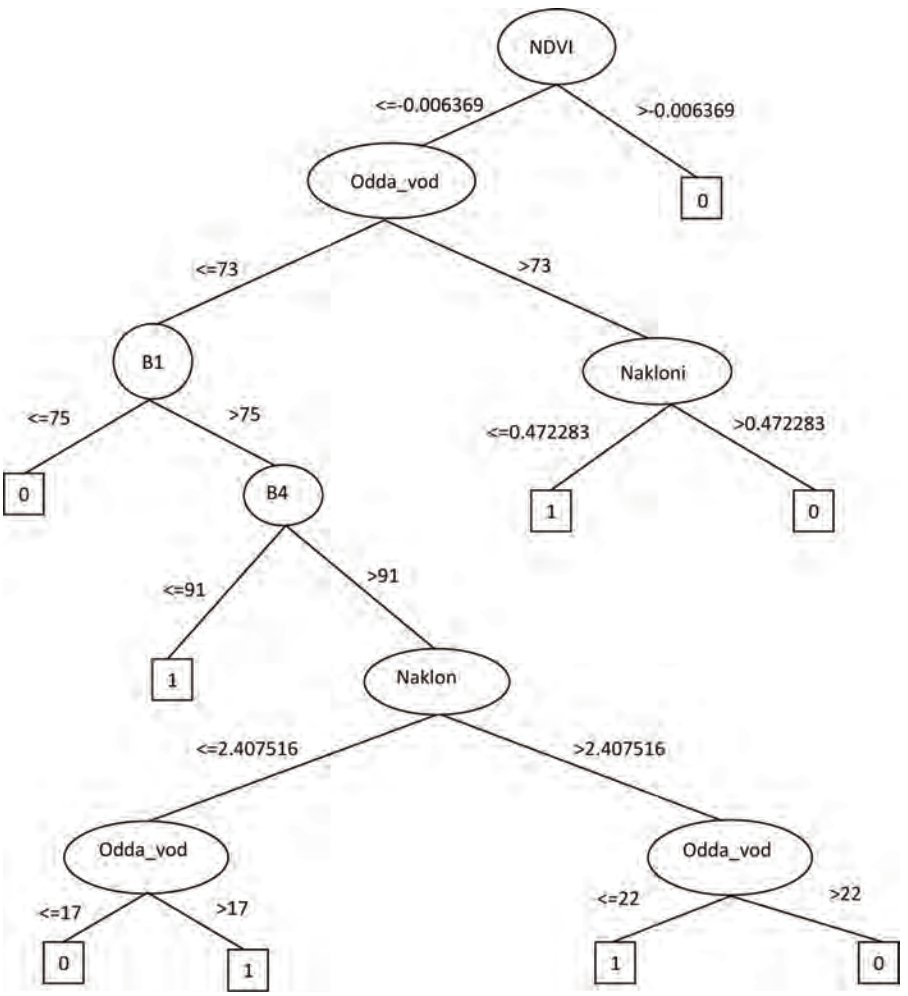
Atribut ukrivljenosti pomeni skupno ukrivljenost površja, ki je geometrična sredina med absolutnimi vrednostmi navpične in vodoravne ukrivljenosti površja. Površje z veliko skupno ukrivljenostjo je potencialno območje intenzivnih geomorfoloških procesov (Hrvatin in Perko, 2002).

»Oddaljenost od vodotokov« je bila izračunana na podlagi vektorskega sloja hidrografija – linijska topologija generalizirane kartografske baze GKB25. Izračun oddaljenosti je bil izveden s programskim orodjem ArcMAP. Rezultat je rastrski sloj ločljivosti 2 m, kjer vsaka celica pomeni oddaljenost od najbližjega vodotoka.

Izboru atributov sledi izvedba klasifikacije vzorčnih primerov. Ta del poteka povsem samodejno in omogoča stroju (računalniku), da se na podlagi vzorčnih (učnih) primerov nauči pravilno razporejati preostale primere. Med bolj pogoste načine izvedbe klasifikacije spadajo odločitvena drevesa, odločitvena pravila, naivni Bayesov klasifikator in klasifikator z najbližjimi sosedi (Polanec, 2006). V obravnavanem primeru je bila izbrana struktura odločitvenih dreves.

Odločitveno drevo je sestavljeno iz vozlišč in vej. Vozlišča ustrezajo atributom in veje, ki izhajajo iz vozlišč, ustrezajo podmnožicam vrednosti atributov (Kononenko, 1997). Vgrajeni algoritmi v programskih orodjih za strojno učenje samostojno ocenijo pomembnost posameznega atributa. V klasifikaciji sodelujejo le tisti atributi, ki so prepoznani kot pomembni, tj. odločilni pri razvrstitvi vzorčnih primerov v razrede. Za slednje algoritem določi njihove vrednosti na vejah. Sosledje tako dobljenih pogojev, ki so med sabo konjunktivno povezani, vodi v oblikovanje odločitvenega drevesa. Drevo se konča z listi, ki ustrezajo razredom. Vsak vzorčni primer se lahko uvrsti v en sam razred, do katerega vodi samo ena pot od korenine do lista odločitvenega drevesa.

Za ponazoritev si poglejmo primer klasifikacije na učnem vzorcu z 255 točkami (slika 5). Prvi uporabljen atribut (prvo vozlišče) v našem primeru je NDVI. Mejna vrednost, ki določa, ali bo posamezna točka uvrščena med poplavljena ali med nepoplavljena območja, je –0,086. Točke z višjo vrednostjo indeksa so uvrščene v razred nepoplavljenih območij. Podmnožica točk, katerih

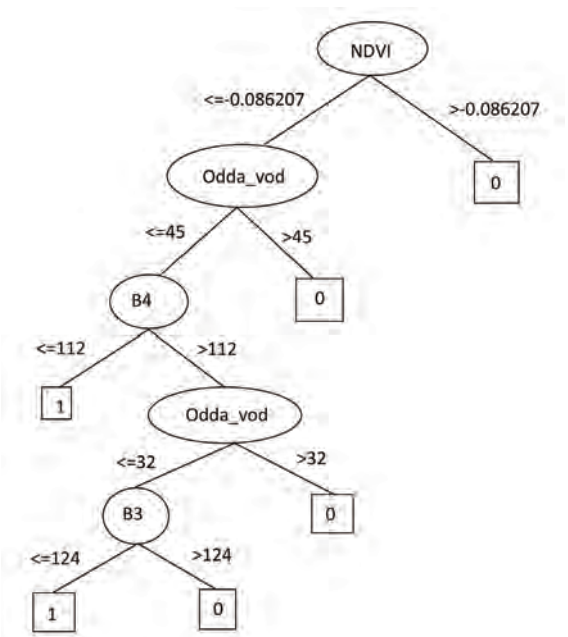


Slika 4: Prikaz modela klasifikacije poplavljenih površin z odločitvenim drevesom za vzorec s 400 točkami

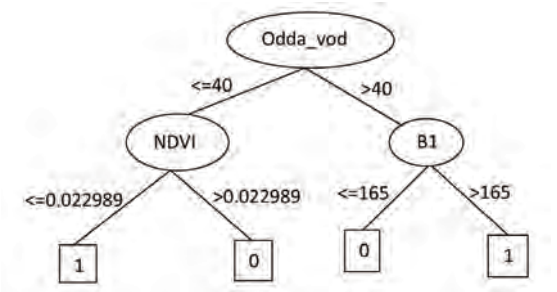
vrednost NDVI je manjša ali enaka –0,086, pa je postavljena pred dodaten pogoj oddaljenosti od vodotokov. Ta pogoj razdeli točke v dve novi podmnožici. To sta podmnožica točk, katerih oddaljenost od vodotokov je večja od 45 m, in podmnožica točk z manjšo ali enako oddaljenostjo od vodotokov. Medtem ko je prva podmnožica uvrščena v razred nepoplavljenih območij, je druga podmnožica izpostavljena novemu pogoju. Tako si sledijo dodatni pogoji, dokler niso vsi primeri v svojem razredu (Witten in Frank, 2005).

Običajno je učenju namenjen le del podatkov, drugi del podatkov je uporabljen za ovrednotenje uspešnosti modela (testiranje modela). Rezultat učenja je funkcija (model), ki preslika prostor atributov v razred. Prednost modela je, da opisuje samo dejstva v podmnožici podatkov (vzorcu). Model se izogiba prikazovanju manj pomembne vsebine, zaradi česar je iz njega enostavneje razbrati glavne informacije obravnavanega pojava.

Pri modelu na sliki 5 se upoštevajo atributi NDVI, oddaljenost od vodotokov ter kanala B4 in B3 multispektralnega satelitskega posnetka SPOT. Ker se klasifikacija prične z atributom NDVI (korenina drevesa), lahko predvidevamo, da je ta atribut najpomembnejši. Iz drevesa lahko sklepamo, da bodo poplavljena območja, ki imajo negativen vegetacijski indeks (manjši ali enak –0,086), niso od vodotokov oddaljena več kot 45 m in imajo vrednost multispektralnega kanala B4 največ 112. Če je vrednost kanala B4 večja od 112, mora območje ustrezati še dvema dodatnima pogojema, da bo uvrščeno med poplavljene površine. Oddaljenost od vodotokov mora biti manjša ali enka 32 m in kanal B3 ne sme presegati vrednosti 124. Dobljeni model (slika 4) je smiseln, saj so negativne vrednosti indeksa NDVI značilne za vodne površine. Prav tako manjša oddaljenost od vodotokov pomeni večjo nevarnost poplav. Nižje vrednosti kanalov B3



Slika 5: Prikaz modela klasifikacije poplavljenih površin z odločitvenim drevesom za vzorec z 255 točkami



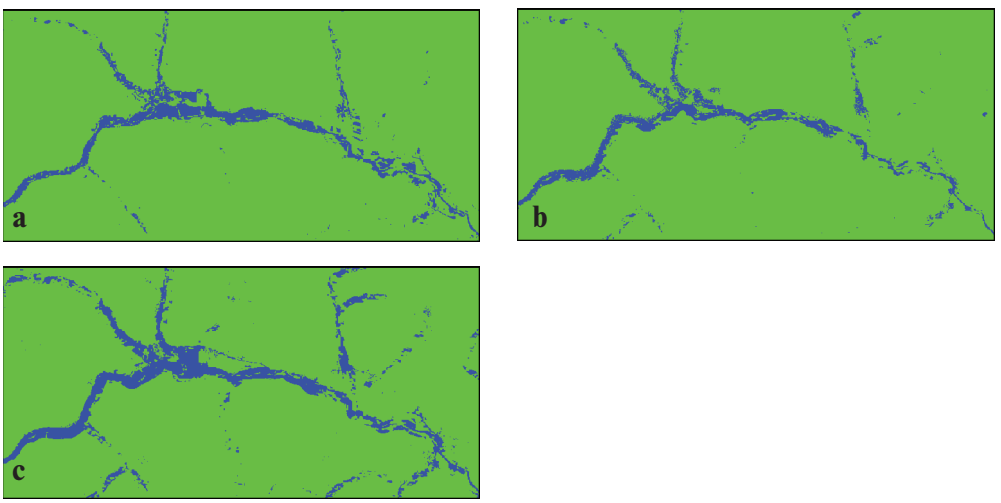
Slika 6: Prikaz modela klasifikacije poplavljenih površin z odločitvenim drevesom za vzorec z 49 točkami

in B4 na poplavljenih območjih pa je mogoče razložiti z manjšim odbojem bližnje infrardeče svetlobe na vodnih površinah.

Na koncu opravimo oceno uspešnosti strojnega učenja. Za zgrajeni algoritem nas zanima, kako uspešno bo reševal nove težave. Pri naučenih logičnih relacijah je treba oceniti, za koliko novih objektov bodo samodejno zgrajeni odnosi delovali pravilno (Kononenko, 1997). Za ocenjevanje uspešnosti samodejno zgrajenega znanja običajno ločimo razpoložljive podatke na učno in testno množico. Učna množica vsebuje podatke, ki so na voljo algoritmu za učenje. Testna množica pa se uporabi za ocenjevanje uspešnosti samodejno zgrajene teorije. Večina programov za strojno učenje ponuja možnost samodejne razdelitve podatkov na učno in testno množico.

4 PRIKAZ POPLAVLJENIH POVRŠIN

Prikaz poplavljenih površin – karta omogoča vsem zainteresiranim ob poplavah, da čim učinkoviteje opravijo svoje delo. Reševalcem omogoča pravilno porazdelitev reševalnih ekip po območju, zavarovalnicam pa lažjo ovrednotenje povzročene škode.



Slika 7: Primerjava klasifikacij poplavljenih območij, izdelanih z modeli, pridobljenimi iz različnega števila vzorčnih točk: 400 (a) 255 (b) in 49 (c).

V našem primeru smo za vizualizacijo poplavljenih območij uporabili program ERDAS IMAGINE (Erdas Imagine, 2010). Modele, ki smo jih pridobili na podlagi vzorčnih točk, smo uporabili nad celotnim obravnavnim območjem. V našem primeru je enoto klasifikacije pomenila ena rastrska celica. Pogoj za to je, da poznamo vrednosti atributov, ki nastopajo v modelu, za vsako rastrsko celico obravnavanega območja. Vsaka celica se tako uvrsti v razred, ki mu pripada glede na izbrani model. Rezultat uporabljenega modela je karta poplavljenih območij.

Na sliki 7 je prikazan rezultat klasifikacije poplavljenih površin z različnim številom vzorčnih točk. Najboljši rezultat smo dobili s klasifikacijo na podlagi 255 vzorčnih točk (slika 3b). Delež pravilno klasificiranih točk znaša 95 %. Uspešnost preostalih dveh klasifikacij je 92 % pri večjem in 84 % pri manjšem vzorcu.

5 RAZPRAVA IN SKLEPNE UGOTOVITVE

Predstavljeni postopek določitve poplavljenih površin je pokazal, da je strojno učenje v tovrstnih primerih uporabno. Pri analizi uporabljenih podatkov se je izkazalo za učinkovito, kar je razveseljivo, saj so količine teh podatkov običajno zelo velike, njihova analiza pa je posledično dolgotrajna. Metode strojnega učenja lahko bistveno pripomorejo k skrajšanju obdelave tovrstnih podatkov.

Za določitev poplavljenih območij smo uporabili algoritem klasifikacije. Algoritem samostojno oceni, kateri podatki in njihove vrednosti odločilno vplivajo na to, ali bo neko območje poplavljeno ali ne. Trije različni vzorci, na katerih je potekalo strojno učenje, so pokazali na nekatere pomanjkljivosti, ki dopuščajo še veliko prostora za izboljšanje rezultatov. Vzorci se v obravnavanem primeru niso izkazali za dovolj reprezentativne za celotno obravnavano območje. Vzorec mora biti dovolj velik in zastopan s primeri, ki opisujejo celotno populacijo z vsemi njenimi raznolikostmi. Nabor atributov je bil v predstavljenem primeru sicer zadosten, vendar so bili vzorci bodisi premajhni bodisi njihovi atributi niso dovolj homogeno zastopali celotne populacije. Delež vzorčnih točk na poplavljenih površinah je veliko manjši kot na nepoplavljenih površinah. Razmerji 336 : 64 ter 231 : 24 v prid vzorčnim točkam na nepoplavljenih površinah sta bistveno preveliki. To razmerje je boljše le v najmanjšem vzorcu (25 : 24), vendar pa je vzorec z 49 točkami za učenje premajhen, saj je v tem primeru vpliv posamezne vzorčne točke na oblikovanje modela klasifikacije prevelik. Dodatne izboljšave je mogoče doseči tudi s preizkušanjem uporabe različnega razenja odločitvenih dreves. Rezanje je mogoče uravnavati z določevanjem minimalnega števila vzorčnih primerov, ki jih mora vsebovati posamezen razred (list odločitvenega drevesa), in določevanjem stopnje zaupanja. V našem primeru smo uporabili privzete vrednosti, tj. št. min. primerov 2 ter faktor zaupanja 0,25.

Kljub naštetim pomanjkljivostim je treba predstaviti, katri atributi so se izkazali kot najpomembnejši za klasifikacijo v izdelanih modelih (glej odločitvena drevesa na slikah 4, 5, 6) in te ugotovitve upoštevati pri delu v prihodnje. Najpomembnejši vpliv na poplavljenost površin obravnavanega območja imajo vegetacijski indeks NDVI, kanali B1, B3 in B4, oddaljenost od vodotokov in naklon. Pri tem je treba omeniti, da je vegetacijski indeks NDVI izračunan iz vrednosti kanalov B2 (rdeča) in B3 (bližnja IR) multispektralnega posnetka SPOT. Iz tega

izhaja, da imajo pomembno vlogo pri določitvi poplavljenih površin vsi kanali multispektralnega posnetka SPOT – tudi kanal B2, ki nastopa pri klasifikaciji le posredno prek indeksa NDVI. Pričakovan je tudi pomemben vpliv oddaljenosti od vodotokov. Naklon je bil pomemben dejavnik le pri klasifikaciji pri vzorcu s 400 točkami. Območja, ki so daleč stran od rek in potokov, niso bila poplavljena. Prav tako ni poplavljeno površje z velikimi nakloni. Območja, ki so poplavljena kljub precej veliki oddaljenosti od vodotokov, imajo običajno manjši naklon. Dejavnika višine in ukrivljenosti nimata večjega vpliva na poplavljenost/nepoplavljenost območij. Neodvisnost poplavljenih površin glede na višino lahko utemeljimo s tem, da so poplavljena tako območja v zgornjem delu toka reke Selška Sora (območja z višjo nadmorsko višino) kot območja v spodnjem delu toka (območja z nižjo nadmorsko višino). Glede ukrivljenosti predvidevamo, da bi ta dejavnik igral pomembnejšo vlogo pri obravnavanju večjega območja. V našem primeru je obravnavana samo Selška dolina, kjer prevladuje podobna ukrivljenosti za celotno območje.

Upamo lahko, da bo članek pripomogel k razširitvi uporabe strojnega učenja na podobnih primerih. V spodbudo pri uporabi in izboljšavi predstavljenega postopka pa naj bo praktična uporabnost tovrstnih rezultatov. Poleg ugotavljanja škode po naravnih nesrečah je uporabna tudi pri možnostih za a-posteriori oziroma naknadne določitve poplavne ogroženosti in s tem preprečevanje najhujših posledic pri morebitnih ponovitvah. Predhodne oziroma apriorne analize poplavne ogroženosti so pomembne pri načrtovanju vseh posegov v prostor.

Literatura in viri:

ArcGIS (2010). Pridobljeno 22. 11. 2010 s spletne strani: http://www.esri.com/software/arcgis/arcgis10/ index.html.

Džeroski, S. (2001). Data Mining in a Nutshell. V: S. Džeroski in N. Lavrač (ur.), Relational Data Mining. Berlin: Springer, 3–27.

Erdas Imgine (2010). Pridobljeno 22. 11. 2010 s spletne strani: http://www.erdas.com/default.aspx.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data minig to knowledge discovery: An overview. V: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (ur.), Advances in Knowledge Discovery and Data Mining. Cambridge: MIT Press, 1–34.

Frawley, W., Piatetsky-Shapiro, G., Matheus, C. (1991). Knowledge discovery in databases: An overview. V: G. Piatesky-Shapiro, W. Frawley (ur.), Knowledge discovery in databases. Cambridge: MIT Press, 1–27.

Hrvatin, M., Perko, D. (2002). Ugotavljanje ukrivljenosti površja z digitalnim modelom višin in njena uporabnost v geomorfologiji. V: T. Podobnikar, D. Perko, M. Krevs, Z. Stančič, D. Hladnik (ur.), Geografski informacijski sistemi v Sloveniji 2001-2002. Ljubljana: ZRC-SAZU, 65–72.

Kononenko, I. (1997). Strojno učenje. Ljubljana: Založba FE in FRI.

Maimon, O., in Rokach, L. (2005). Introduction to knowledge discovery in databases. V: O. Maimon in L. Rokach (ur.), Data Mining and Knowledge Discovery Handbook. New York: Springer, 1–17.

Pehani, P., Kokalj, Ž., Marsetič, A., Oštir, K. (2008). Uporaba satelitskih posnetkov za analizo poplav septembra 2007. V: D. Perko, M. Zorn, N. Razpotnik, M. Čeh, D. Hladnik, M. Krevs, T. Podobnikar, B. Repe, R. Šumrada (ur.), Geografski informacijski sistemi v Sloveniji 2007–2008. Ljubljana: Geografski inštitut Antona Melika ZRC-SAZU, 117–128.

Polanec, K. Strojno učenje (2006). Pridobljeno 23. 11. 2009 s spletne strani: http://dat.si/publikacije/Article/Strojno-u--269-enje/66.

ARSO (2008). Visoke vode in poplave 18. septembra 2007 (26. 2. 2008). Pridobljeno 20. 11. 2010 s spletne strani: http://www.arso.gov.si/vode/poro%C4%8Dila%20in%20publikacije/Visoke%20vode%20in%20poplave%2018.%20 septembra%202007.pdf.

Vodišek, D. (2009). Opazovanje poplav s podatki daljinskega zaznavanja. Diplomska naloga. Ljubljana: Fakulteta za gradbeništvo in geodezijo.

Weka (2010). Pridobljeno 22. 11.2010 s spletne strani: http://www.cs.waikato.ac.nz/ml/weka/.

Witten, I. H., Frank, E. (2005). Data mining: practical machine learning tools and techniques. Second Edition. Elsevier: San Francisco.

Prejeto v objavo: 16. avgust 2010
Sprejeto: 30. november 2010

Lamovec Peter, univ. dipl. inž. geodezije
IAPŠ ZRC SAZU, Novi trg 2, 1000 Ljubljana,
e-pošta: plamovec@zrc-sazu.si

izr. prof. dr. Krištof Oštir, univ. dipl. inž. fizike
IAPŠ ZRC SAZU, Novi trg 2
1000 Ljubljana
e-pošta: kristof@zrc-sazu.si