SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Petra Martinjak

PROBABILISTIČKI (VJEROJATNOSNI) MODEL POSJETA KATEGORIJAMA PROSTORNIH OBJEKATA U POLAZNO-ODREDIŠNOJ MATRICI

Diplomski rad

Voditelji rada:
prof.dr.sc. Luka Grubišić
prof.dr.sc. Renato Filjar

Zagreb, veljača, 2019

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Mojim roditeljima koji su mi svojom ljubavi i požrtvovnosti omogućili da ostvarim sve svoje potencijale i bili mi podrška na svakoj stepenici puta.
Filipu, koji je dijelio sa mnom sve lijepe i sve teške trenutke i učinio ovo razdoblje sretnijim.
Kolegicama i kolegama s faksa na svoj pomoći i riječima ohrabrenja i podrške, te prijateljima, obitelji i svima ostalima koji su bili uz mene.
Veliko hvala i prof.dr.sc. Luki Grubišiču i prof.dr.sc. Renatu Filjaru na vodstvu i savjetima prilikom izrade ovoga rada.

Sadržaj

Uvod

Kada govorimo o migracijama ljudi unutar zadanog prostora, teško je predvidjeti kretanje pojedinca. Međutim, pokazalo se izvedivim i isplativim promatrati kretanja masa ljudi i na temelju takvih podataka i uočenih trendova u njima planirati buduće radnje. Takav pristup se koristi u prometu za planiranje prometne infrastrukture [27]. Obilježja tokova urbanih dnevnih migracija najčešće se zapisuju u obliku polazišno-odredišne matrice.

Tradicionalno, polazišno-odredišne matrice dobivaju se brojanjem ljudi i vozila na frekventnim područjima (prometna raskrižja, važne prometnice...) ili korištenjem podataka videonadzora. U novije vrijeme koriste se metode koje koriste podatke o korištenju javnih pokretnih (telekomunikacijskih) mreža za procjenu migracija [8]. Teleoperateri prikupljaju mnoštvo podataka s mobilnih uređaja, primjerice podatke o aktivnosti korisnika zasnovane na zapisima o govornim pozivima, slanju kratkih poruka (*engl. Short Messaging Service, SMS*) te ostalim podatkovnim transakcijama po internetskoj javnoj pokretnoj mreži. Svaki put kada korisnik napravi poziv, pošalje kratku poruku ili neki drugi skup podataka putem interneta, bilježi se identifikacijska oznaka pristupne točke radijske mreže na koju se korisnik spojio, te vrijeme početka i trajanje telekomunikacijske aktivnosti. Podaci o telekomunikacijskoj mreži (položaji pristupnih točaka radijske mreže) i telekomunikacijskoj aktivnosti mogu biti korišteni za realniju procjenu POM-e i opisu migracija primjerenije određivanje zona [26]. Određenje zona polazišno-odredišnih matrica u ovom radu je postignuto Voronoievom teselacijom oko pristupnih točaka radijske mreže.

OpenStreetMap (OSM) [11] je baza prostornih podataka otvorenog pristupa, nastala volonterskim doprinosima, s ciljem pružanja svima dostupnih prostornih podataka. Glavne karakteristike su joj
(1) detaljan i skalabilan opis prostora i objekata u prostoru na standardiziran način dostupan komercijalnim i slobodnim alatima za rad s prostornim podatcima
(2) usklađenost raznovrsnosti izvora, uz održavanu kvalitetu podataka
(3) široka dostupnost i omogućen pristup.
OpenStreetMaps je rezultat koordiniranog i usmjerenog rada volonterske zajednice koja održava najveću svjetsku bazu prostornih podataka, zasnovanu na volonterskom zapisivanju i dijeljenju podataka o trajektorijama (snimljenim satelitskim navigacijskim prijemnikom) te satelitskim, radarskim i ostalim snimkama prostora (kao podlogama baze). U

ovom radu, kategorizacija objekata biti će zasnovana na podacima dobivenim iz OSM-a.

Cilj istraživanja predstavljenoga u ovom diplomskog radu bio je iskoristiti povezivanje atributa prostornih podataka iz OSM baze s opažanjima uzorka migracija u svrhu dobivanja kvalitetnijih POM-ova koje se ne ograničavaju na raspoznavanje uzoraka migracija unutar transportne mreže, već izražavaju detaljnije spoznaje o migracijama unutar konteksta društveno-ekonomskih aktivnosti. Istraživanja predstavljena u ovom radu zasnivaju se na nekonvencionalnim izvorima podataka lokalnim urbanim migracijama: anonimiziranim zapisima o telekomunikacijskoj aktivnosti korisnika u javnoj pokretnoj mreži, atributim prostornih podataka iz baze prostornih podataka OSM te preciznim GNSS (Global Navigation Satellite System) opažanjima položaja kao elemenata migracijskih trajektorija podskupa ukupne populacije (taxi vozila opremljena GPS prijamnicima). Razmotreni podatci odnose se na izabrani vremenski period i područje grada Shenzhena, Kina.

U prvom poglavlju razrađena je motivacija za istraživanje i domena problema. U drugom poglavlju opisani su podaci koji se koriste te njihovo pribavljanje i obrada. U trećem poglavlju razrađena je kategorizacija podataka, a u četvrtom je opisana izrada regresijskog modela i primjena na eksperimentalnim podacima.

Rad je u potpunosti izveden u programskom jeziku R (verzija 3.4.4) koristeći okruženje RStudio. [12][22]

Poglavlje 1

Motivacija

1.1 Polazišno-odredišne matrice

Tokom posljednjih nekoliko desetljeća, s porastom broja automobila, pojavom novih i usavršavanjem tradicionalnih usluga javnog prijevoza te sukladno općem razvoju prometnih sustava, pojavljuje se potreba za novim informacijskim elementima vezanim uz upravljanje prometom i predviđanje prometa. U tu svrhu često se koristi polazišno-odredišna matrica kao način izražavanja koji opisuje tokove kretanja iz jednog područja u promatranom prostoru u drugo. Promatrani prostor dijeli se na zone koje mogu biti polazište i/ili odredište migracija.

Definicija 1.1.1 (Polazišno-odredišna matrica)**.** *Neka je $\mathcal{A}$ promatrano unaprijed određeno područje u ravnini, određeno u sebe zatvorenom krivuljom granice. Neka su $a_1, ..a_k$ zone unutar promatranog područja $\mathcal{A}$, pri čemu svaka zona predstavlja podskup točaka od $\mathcal{A}$, određenih jedinstvenim koordinatama položaja, određen u sebe zatvorenom krivuljom granice, uz uvjet da je presjek bilo kojih dviju zona prazan skup. Matricu $B = (b_{ij})$ dimenzije $n \times m$ koja je definirana tako da je b_{ij} ,$i = 1...n,\ j = 1, ...m$, broj migracija iz zone a_i u zonu a_j nazivamo polazno-odredišna matrica (POM).*

U (i, j) element polazišno-odredišne matrice (POM) upisuje se opaženi broj migracije iz i-te zone u j-tu zonu.

U izvornom tumačenju prometnih znanosti, vrijednosti elemenata POM-e dobivene su brojanjem prometa [5], tj. brojanjem vozila u određena doba dana i na određenim točkama (obično raskrižjima kao čvorovima cestovne mreže) na promatranom području [29]. Ovakav pristup prikupljanju podataka je složen, zamoran (stoga izaziva subjektivne pogreške individualnih promatrača kao izvora podataka), nepouzdan (podaci su neravnomjerne kvalitete) i skup. Tehnološki razvoj (posebno u području računarstva i određivanja položaja) omogućio je postupnu automatizaciju prikupljanja podataka, primjerice korištenjem po-

dataka iz videonadzora i mjernim osjetilima za automatsku detekciju vozila [3]. Podaci prikupljeni na ovakav način često i dalje nisu dovoljni za izradu polazišno-odredišnih matrica, stoga se podaci kombiniraju [4] te se pristupa različitim metodama procjene [6][13]. Uz to, ovakva metoda naglašava prostornu distribuciju intenziteta prometa, a manje na vremensku. Neka istraživanja [17] su ukazala na dodatne nedostatke načina kreiranja polazišno-odredišnih matrica tradicionalnim metodama: zasnivanje metoda na zastarjelim modelima POM-a, loša vremenska rezolucija podataka koja ne prikazuje dnevne promjene u prometu i ne uzimanje u obzir prirodnih dnevnih i sezonskih varijacija u prometu te promjena uslijed velikih događanja (utakmice, koncerti itd.)

Alternativa tradicionalnom pristupu pri izradi polazišno-odredišnih matrica je pristup koji koristi odnosne podatke iz alternativnih i neovisnih izvora, kao što su zapisi o telekomunikacijskim aktivnostima u javnoj pokretnoj mreži, uz potpuno očuvanje privatnosti individualnih korisnika javne pokretne mreže. Metodologija izrade i sposobnost detaljnije procjene migracija na dijelovima promatranog područja izvan javne prometne (cestovne) mreže ovako dobivenih POM-a validirana je u brojnim istraživanjima [21][7][15][8]. S porastom korištenja pokretnih uređaja raste količina podataka o pokretljivosti korisnika javne pokretne mreže, a time i količina podataka potencijalno iskoristivih za procjenu migracija i POM-e. Ti podaci se dobivaju bilježenjem identifikacijskih oznaka i položaja pristupnih točaka (baznih stanica) pristupne radijske mreže na koje se uređaji spajaju pri uspostavljanju poziva, slanja SMS poruka i slanja i primanja mobilnih podataka. Područja pokrivanja pristupnih točaka radijske mreže mogu biti definirana kao prostorne zone polazišno-određenih matrica. Jedan od načina kako se to može učiniti je Voronoievom teselacijom [10][9].

Definicija 1.1.2 (Teselacija)**.** *Teselacija plohe* Ω *definira se kao skup poligonalnih područja čija je unija cijela ravnina i čiji se unutrašnji dijelovi ne sijeku. Za teselaciju se kaže da je dobro poravnata ako se bilo koja dva područja susreću ili u zajedničkom vrhu ili u zajedničkom rubu ili uopće ne. Regularna teselacija je teselacija čija se poligonalna područja podudaraju s regularnim poligonom.*

Definicija 1.1.3 (Voronoieva teselacija)**.** *Voronoieva teselacija ili Voronoiev diagram zadan skupom točaka* $P = p_i$ *je teselacija čiji elementi* K_i *imaju svojstvo da su sve točke poligona stupnja* K_i *bliže točki* p_i *nego ijednoj drugoj točki skupa P.*
Voronoiev diagram se ponekad još naziva i Dirichletova teselacija.
Poligonalna područja od kojih se sastoji Voronoieva teselacija nazivaju se Voronoieve ćelije.

Osim položaja bazne stanice, precizno se mogu odrediti i trenutak početka i trajanje telekomunikacijske transakcije (korištenja usluge telekomunikacijske mreže) putem promatrane pristupne točke radijske mreže s baznom stanicom te kada i na koju sljedeću baznu

stanicu je uređaj bio spojen. Time je, osim prostorne, određena i precizna vremenska komponenta.

Može se pokazati kako [26] da polazišno-odredišne matrice dobivene iz telekomunikacijskih podataka iskazuju veću točnost od matrica dobivenih tradicionalnim metodama, te da mogu davati konzistentnije procjene migracija, pogotovo na područjima gdje je teško skupiti podatke za tradicionalne metode.

1.2 Primjena polazišno-odredišnih matrica

Najčešća primjena polazišno-odredišnih matrica je u planiranju prometne infrastrukture, primjerice kod odlučivanja o izgradnji novih cesta, broju traka, usmjeravanju prometa i postavljanju semafora. Zasnovano na kvalitetnim POM se mogu osmisliti načini korištenja cestovne mreže [27] koje omogućuju djelotvornije korištenje cestovne mreže i optimizaciju prometa, ili se mogu koristiti za planiranje prometnih strategija u slučaju masovnih događanja [28]. Međutim, osim u svrhu poboljlšanje prometnih tokova, POM-e se mogu primijeniti i u druge svrhe poput: identifikacije urbanih područja prema klasifikaciji aktivnosti (proizvodnja, trgovina, školstvo), ravnomjerni razvoj pojedinih gradskih četvrti putem poticanja društveno-ekonomske aktivnosti koja nedostaje razvojem cestovne i telekomunikacijske infrastrukture, ciljanim uređenjem okoliša i upravljenjem prometom itd.). Izvorne POM-e zasnovane na zapisima o telekom aktivnosti još uvijek ne sadrže kontekst koji bi omogućio razlučivanje društveno-ekonomskih aktivnosti koje potiču dnevne lokalne migracije. Iz samog broja i vremena putovanja nije jasna njihova svrha, zbog čega se dobiva nepotpuna slika o ljudskoj mobilnosti. Razumijevanje i poznavanje razloga zašto ljudi putuju ključno je za planiranje urbanih projekata, razvoj gradova, ekonomije i turizma [19][23][2].

1.3 Pristup problemu

Pokazalo se [1] da se uz pomoć POM-a dobivenih iz telekomunikacijskih podataka, te statističkih pokazatelja o društvenim trendovima stanovništva polazišno-odredišne migracije mogu podijeliti po svrsi. Statističke procjene ponašanja stanovništva dobivaju se dugotrajnim i skupim procesima i nisu uvijek dostupne, te se postavlja pitanje mogu li se podaci o razlogu i svrsi putovanja pribaviti na druge načine. Moguće je [24] iskoristiti podatke o društvenoj funkciji prostornih objekata na području interesa kako bi se ti objekti kategorizirali u skupine. Time se dobiva distribucija objekata po njihovoj funkcionalnosti. Vjerojatnost posjete objektu koji pripada nekoj od funkcionalnih skupina tada ovisi o tome koliko objekata te skupine ima u odnosu na objekte drugih skupina, tj. o distribuciji objekata po skupinama. Time se distribucija objekata po skupinama pretvara u vjerojatnosnu

distribuciju, koja se može iskoristiti za identifikaciju svrhe putovanja zabilježenih u polazišno-odredišnim matricama.

U ovom istraživanju biti će razrađena sama kategorizacija prostornih objekata na temelju podataka o njihovoj društevnoj funkciji i prikazan model koji ovisno o vremenskom okviru unutar dana predviđa vjerojatnost posjećenosti određenim kategorijama objekata.

Poglavlje 2

Podaci

2.1 OpenStreetMap

Ukratko o OSM projektu

Projekt OpenStreetMap (OSM) [11] pokrenuo je 2004. godine Steve Coast, u želji da učini prostorne podatke besplatnima i svima dostupnima. Inicijalno je projekt bio usredotočen na područje Ujedinjenog Kraljevstva, ubrzo se proširivši na ostatak svijeta. Danas se OSM podaci koriste u mnogim programskim podrškama i uslugama, te od strane mnogih kompanija i institucija. Nekoliko primjera: Apple, Foursqare, Flickr, Uber, TripAdvisor, Wolfram Alpha, DuckDuckGo, igre Ingress i PokemonGO. Jedno od češćih pitanja vezanih uz OpenStreetMap je koja je razlika između njegovih usluge zasnovanih na protokolima i bazama prostornih podataka i usluga GoogleMapsa i zašto ne koristiti GoogleMaps usluge. Odgovor na to pitanje leži u tome što iako su usluge GoogleMapsa do određene količine korištenih podataka besplatne kroz GoogleMaps API, podaci koji se u tim uslugama koriste i dalje podliježu autorskim pravima organizacija koje su ih prikupile. Zbog toga se podaci prikupljeni uz pomoć GoogleMapsa smatraju izvedenim djelom i podliježu istim autorskim pravima. Za razliku od GoogleMapsa, OSM projekt je projekt zajednice na volonterskoj bazi. Njegov glavni pokrovitelj je OpenStreetMaps Foundation koji se brine i o financijskoj potpori (primjerice za održavanje servera), ali on nije vlasnik OSM podataka. Stoga se podaci prikupljeni kroz OSM smiju legalno koristiti u daljnje svrhe.

Punjenju i osvježavanju OSM prostornih podataka, pročišćavanju i kontroli kvalitete OSM podataka može pridonijeti svatko, unošenjem zabilježenih putanja kretanja, zasnovanim na procjenama položaja GNSS prijamnikom, ručnim mapiranjem objekata ili bilo kojom drugom tehnikom preslikavanja prostornih podataka. Projekt prikupljanja prostornih podataka je suradničke prirode, pa je moguće da unatoč provjerama kvalitete postoji mali skup podataka koji nisu točni. Ovaj problem nije ograničen samo na OSM, naime, i

drugi, komercijalni izbori prostornih podataka također imaju malu količinu netočnih podataka kako bi se osigurale od krađe i neovlaštenog korištenja. Točnost podataka u OSM bazi prostornih podataka postiže se upravo pomoću kolaborativne prirode projekta. Velikoj većini korisnika u interesu je imati točnu bazu (prostornih) podataka. U slučaju da netko i unese netočan podatak, slučajno ili namjerno, ostali korisnici imaju uvid u najnovije promjene i mogu provjeriti i po potrebi ispraviti netočne podatke. Još jedna prednost takvog načina razvoja jest što omogućava brzo osvježavanje podataka, zbog čega se OSM podaci osvježenim prostornim podacima primjereno opisuju trenutno stanje prostora.

Struktura podataka

Objekti su u OSM bazi prostornih podataka određeni vrstom objekta (element baze) i opisom elementa (sadržanom u podatku žiga)

Definicija 2.1.1 (Element)**.** *Element OSM baze prostornih podataka je osnovna komponenta konceptualnog podatkovnog modela fizičkog svijeta u OpenStreetMapsu. Element može biti čvor (engl. node), put (engl. way) ili relacija (engl. relation). Svaki element može (ali i ne mora) imati pridružen opisni element žiga (engl. tag).*

Definicija 2.1.2 (Čvor)**.** *Čvor (engl. node) predstavlja jedinstvenu točku u prostoru koja sadrži razne atribute među kojima i identifikacijski broj (engl. node id), zemljopisnu širinu (engl. latitude) i dužinu (engl. longitude). Čvor može opisivati konkretan fizički objekt u prostoru ili oblik puta (dijela trajektorije ili prometne infrastrukture).*

Definicija 2.1.3 (Put)**.** *Put (engl. way) je uređena lista koja se sastoji od 2 do 2000 čvorova. Obično je opisan žigom ili je dio relacije. Ako je prvi čvor u putu ujedno i zadnji, kažemo da je put zatvoren. U suprotnom kažemo da je put otvoren.*

Definicija 2.1.4 (Relacija)**.** *Relacija (engl. relation) je uređena lista koja se sastoji od čvorova, puteva i/ili drugih relacija. Relacija definira logičku ili geografsku povezanost između svojih članova. Članovi mogu imati dodatni opisni podatak koji služi opisivanju njihove uloge unutar relacije.*

Definicija 2.1.5 (Žig)**.** *Žig (engl. tag) je opisni, tekstualni element koji služi kako bi se definirale značajke elementa. Sastoji se od ključa i vrijednosti. Ključ se koristi kako bi se definirala tema, kategorija ili tip elementa. Vrijednost definira specifičnost elementa za zadani ključ. I ključ i vrijednost mogu biti bilo koji niz znakova, ali u praksi se obično koristi zajednička konvencija.*

Iz definicija je vidljivo da među vrstama elemenata postoji hijerarhijski odnos. Putevi se sastoje od čvorova, a relacije se sastoje od puteva i čvorova. Ipak, svaki od elemenata

	id	visible	timestamp	version	changeset	user	uid	lat	lon
12268	2069628987	true	2012-12-16 11:47:03	1	14291696	Antonio Eugenio Burriel	24070	22.74788	113.8954
12269	2069628989	true	2012-12-16 11:47:03	1	14291696	Antonio Eugenio Burriel	24070	22.74796	113.8997
12270	2069628991	true	2012-12-16 11:47:03	1	14291696	Antonio Eugenio Burriel	24070	22.74804	113.9032
62035	2704217959	true	2014-03-06 17:10:26	1	20953737	MarsmanRom	44514	22.74478	113.8951
62036	2704217939	true	2014-03-06 17:10:25	1	20953737	MarsmanRom	44514	22.74487	113.8949
62049	4398783683	true	2016-09-13 11:54:45	1	42123601	MarsmanRom	44514	22.74485	113.8952
62050	4398783673	true	2016-09-13 11:54:45	1	42123601	MarsmanRom	44514	22.74501	113.8950
62057	3053049783	true	2014-09-01 06:16:15	1	25152826	MarsmanRom	44514	22.75000	113.8987
62066	3115407801	true	2014-10-06 19:45:50	1	25904042	MarsmanRom	44514	22.73990	113.9053
62067	3103491113	true	2014-09-30 09:44:47	1	25763710	MarsmanRom	44514	22.73989	113.9051
62068	4751095855	true	2017-03-23 15:47:59	1	47100294	Wahsaw	2237091	22.73987	113.9049
62175	3115407804	true	2014-10-06 19:45:50	1	25904042	MarsmanRom	44514	22.74637	113.8940
62176	3115407787	true	2014-10-06 19:45:49	1	25904042	MarsmanRom	44514	22.74686	113.8943
62177	3115407798	true	2014-10-06 19:45:49	1	25904042	MarsmanRom	44514	22.74714	113.8946

Slika 2.1: Atributi čvora unutar jedne ˇcelije POM-e

može biti i samostalan, tj. ne pripadati elementu ”iznad” sebe. Primjerice, iako se svaki put sastoji od čvorova, postoje čvorovi koji nisu dio nijednog puta.

Budući da čvorovi često predstavljaju stvarne fizičke objekte, oni će služiti kao skup objekata na kojima će biti izvedena kategorizacija i koji će služiti za razvoj vjerojatnosnog modela. U obzir neće biti uzeti svi čvorovi, nego samo oni koji predstavljaju relevantne objekte, kako će biti prikazano u poglavlju 3.2.

Izdvajanje podataka i segmentacija područja

Kao što je već navedeno, područje interesa iz kojeg će biti ekstrahirani podaci je grad Shenzhen u Kini. Problem izdvajanja podataka možemo podijeliti na 3 koraka. Prvi korak je dohvaćanje podataka iz OSM-a za cijelo područje interesa. Drugi je podjela područja interesa na Voronoieve ćelije, koje su ujedno i zone polazišno-odredišnih matrica. Treći korak je smještanje dohvaćenih OSM objekata u Voronoi ćelije kojima (geografski) pripadaju. Rješenje sva 3 koraka razrađeno je u [24], a u nastavku slijedi kratki opis.

OSM pruža opciju izravnog skidanja podataka na pravokutnom području omeđenom željenim geografskim duljinama i širinama na dva načina: putem web-sučelja za pristup OSM bazi prostornih podataka i pomoću funkcije **get_osm** u programskom okruženju za statističko računarstvo R. Međutim, područje Shenzhena je preveliko za dohvat podataka na takav direktni način. Zbog toga je područje podijeljeno na manje kvadrate, te su podaci dohvaćeni za svaki kvadrat posebno (pomoću funkcije get_osm). U smislu tipova podataka u programskom okruženju za statističko računarstvo R, OSM objekti su liste čvorova,

puteva i relacija, koji unutar sebe sadrže daljnje liste i *data.frame* tipove podataka. Zbog toga je osnovnim postupcima rukovanja podacima moguće međusobno pridružiti podatke s manjih područja u skup podataka za cijelo područje.

Iz Open Cell ID [20] baze podataka, koja sadrži podatke o položajima pristupnih točaka radijske mreže (baznih stanica) baznih stanica, dohvaćaju se koordinate položaja baznih stanica za područje Shenzhena. Oko tih baznih stanica radi se razdioba prostora na voronoieve ćelije koja definira pokrovno područje za svaku baznu stanicu. Ovaj korak se još naziva i voronoi teselacija. Na slici 2.2 prikazane su bazne stanice i voronoieve ćelije definirane oko njih. Tokom ovog koraka pronalaze se i sidrišta koje definiraju voronoieve ćelije. Te točke će kasnije biti korištene i za izdvajanje podataka o kretanju taxi vozila za točno određenu ćeliju (poglavlje 3.3).

Koristeći geolokacije i identifikacijske brojeve čvorova, generiran je *data.frame* (R objekt koji predstavlja podatkovnu matricu koja može imati podatke različitih tipova) prostornih točki koji je rastavljen po Voronoi ćelijama. Za svaki takav podskup prostornih točaka, pomoću identifikacijske oznake prostornog OSM čvora broja moguće je ponovno pristupiti opisnim atributima čvora. Tako su OSM podaci za cijelo područje interesa podijeljeni na podskupove podataka za svaku Voronoi ćeliju. Prikaz OSM podataka i njihovih Voronoi ćelija na području Shenzhena nalazi se na slici 2.3.

2.2 GPS podaci taksija

O podacima

Putanje kretanja taksi vozila definirane su skupovima procjena položaja satelitskim navigacijskim sustavom GPS i preuzeti su sa [14]. Podaci se odnose na grad Shenzhen u Kini i isključivo su za korištenje u akademske svrhe. Zbog privatnosti, sve informacije o datumima i vremenu su izuzete, a identifikacijske oznake po kojima bi se moglo identificirati stvarne osobe zamijenjene su serijskim brojevima.

Struktura podataka

Veličina skupa podataka o putanjama taksi vozila iznosi približno 1.9 GB. Skup podataka o putanjama taksi vozila se sastoji od 46927855 redaka. Svaki redak je sljedećeg oblika:

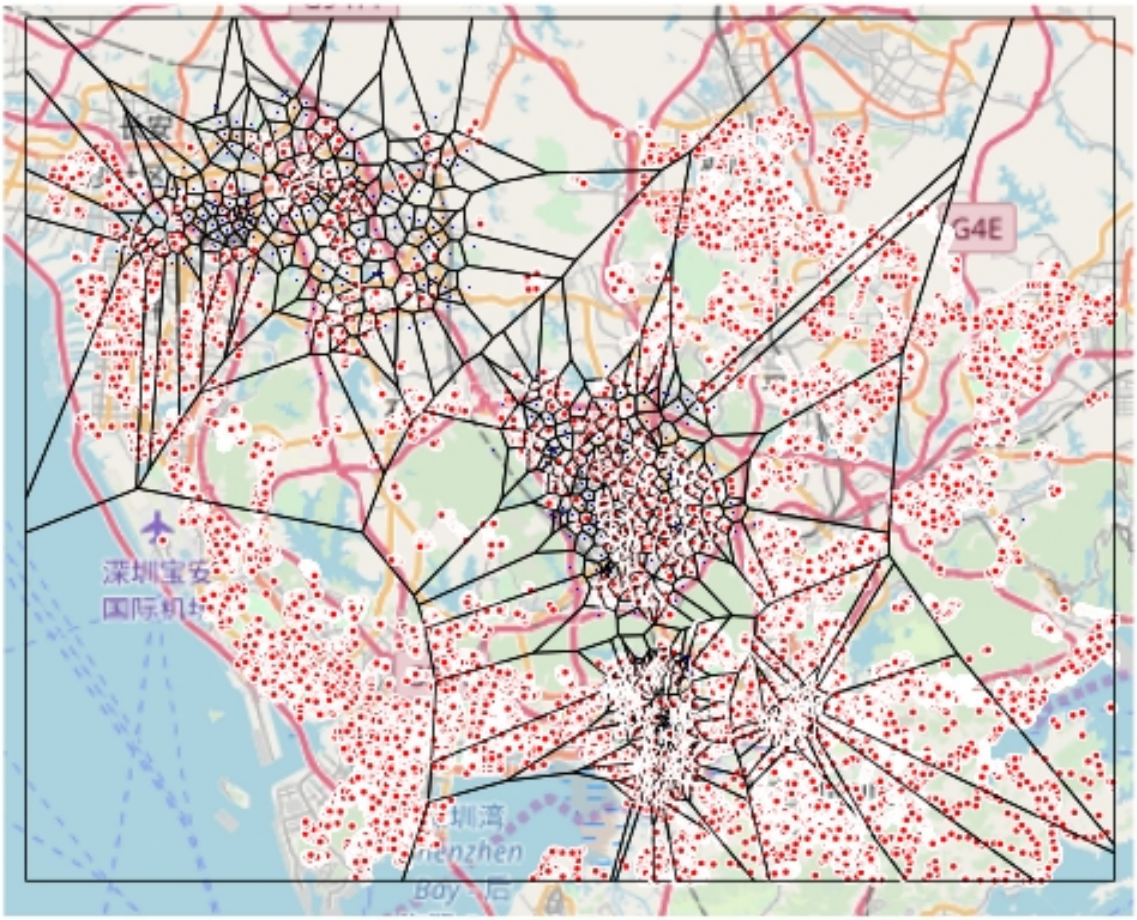
$$TaxiID, Time, Latitude, Longitude, Occupancy, Status, Speed \tag{2.1}$$

Primjer jednog retka: 22223,08:49:25,114.116631,22.582466,0
Taxi ID je redni broj taksija, *Time* predstavlja vremenski trenutak (sat, minute i sekunde) u kojem je podatak zabilježen, *Latitude* je geografska širina položaja taksi vozila u promatranom trenutku Time, *Longitude* je geografska dužina položaja taksi vozila u promatranom

Slika 2.2: Voronoi teselacija vezana za položaje pristupnih točaka radijske mreže s položajima baznih stanica kao sidrištima voronoievih ćelija

trenutku Time. *Occupancy* Status je vrijednost 0 ili 1 koja označava da je taksi prazan (0) ili da u njemu ima putnika (1). *Speed* je brzina kojom se taksi giba u promatranom trenutku Time.

Navedeni podaci opisuju putanje taksija. U nastavku teksta pod pojmom taksi podatka podrazumijeva se jedan redak oblika (2.1) u bazi.



Slika 2.3: OSM objekti unutar Voronoi ćelija

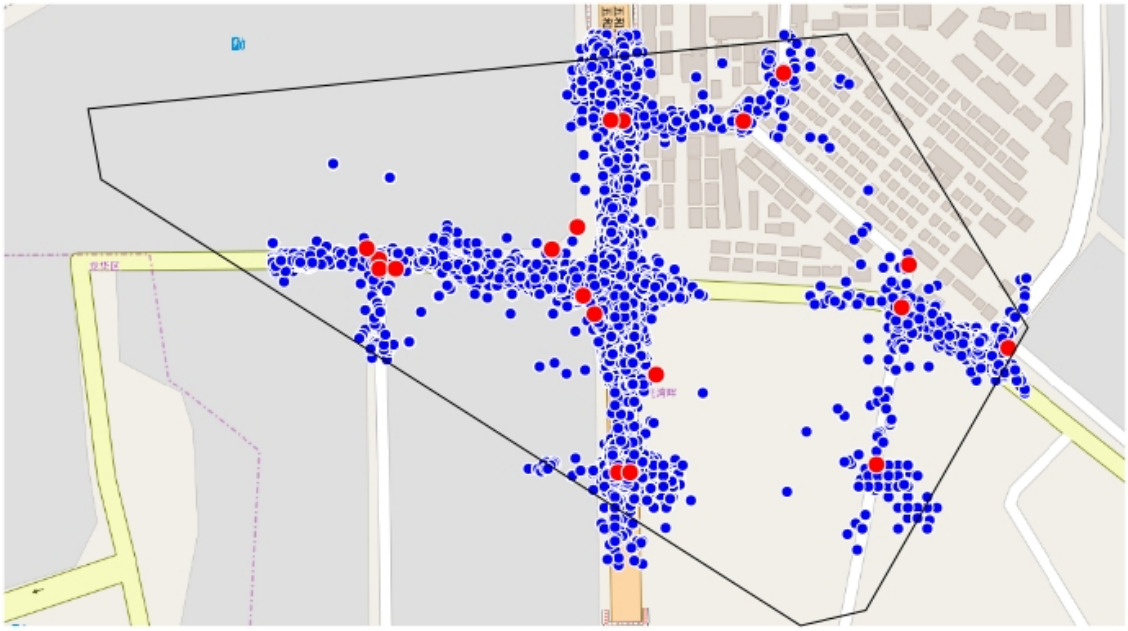
Priprema podataka

Izvorni set podataka je realtivno velik što ne omogućava izravno očitanje u R programsko okruženje za statističko računarstvo. Osnovno okruženje R djeluje tako da skup podataka kojeg procesira mora u cijelosti biti učitano u radnu memoriju računala, što ograničava veličinu skupa podataka kojeg R može procesirati. Međutim, za daljnji rad nisu potrebni svi podaci iz izvornog seta. *Taxi ID* nije bitan podatak, jer svaki taksi podatak gledamo kao individualni objekt, neovisno o tome o kojem točno taksiju se radi. Također, *Occupancy Status* nam ne daje informacije o tome koliko ljudi se nalazi u taksiju, nego samo da li ih ima, a također ne znamo ni je li podatak snimljen neposredno prije ili poslije dolaska na

odredište, pa nam taj podatak ne daje nikakve informacije. Nadalje, budući da se radi o putanjama podaci se zapisuju tokom čitavog puta, a ne isključivo u trenutku zaustavljanja. Kako bi se pronašli podaci nastali kada se taksi doista zaustavi na odredištu, promatraju se oni reci u kojima je brzina (*Speed*) jednaka 0.
Za dio filtracije sirovih podataka koji uključuje izdvajanje samo vrijednosti varijabli (stupaca) *Time*, *Latitude*, *Longitude*, *Speed* koristi se *bash*. *Bash* je skriptni jezik koji se koristi za interpretiranje korisničkih naredbi. To je interpretativni jezik koji obrađuje velike skupove podataka brže od programskog okruženja za statističko računarstvo R, te se tako izbjegava učitavanje velikog skupa podataka u radnu memoriju računala u R. Tako pretprocesiran set podataka učitava se u R pomoću funkcije **fread** iz programske knjižnice **data.table** koja se ponaša slično kao *read* funkcija iz osnovne knjižnice, ali je brža, može učitati veće datoteke, i pogodnija je od *read* funkcije. Skup podataka u R se učitava kao *data.table* objekt, ali zatim se pretvara u *data.frame* i izvršava se drugi dio filtriranja, filtriranje redaka sa brzinom 0. Završni skup podataka sastoji se od 18438300 redaka (136.5 MB).

Prikazani pristup pripremi podataka ima određene nedostatke. *Occupany Status* je značajna informacija koja može znatno utjecati na interpretaciju rezultata. Ako je primjerice u taksiju jedna osoba, to predstavlja jedan posjet objektu u blizini tog taksi objekta (više o kategorizaciji podataka nalazi se u poglavlju 3). U slučaju da su u taksiju 3 osobe, radi se o 3 posjeta objektu u blizini tog taksi objekta. Također, moguće je da taksi ima brzinu 0 zato što stoji na semaforu ili u gužvi. Međutim, iz dostupne strukture podataka, nije moguće izvući detaljnije zaključke bez dublje analize koja izlazi iz opsega ovog rada.

Slikom 2.4 prikazan je podskup taksi podataka unutar jedne voronoieve ćelije zajedno s OSM čvorovima koji sadrže žigove.



Slika 2.4: Taksi i OSM čvorovi koji sadrže žigove unutar jedne Voronoi ćelije. Taksi objekti su prikazani plavom bojom, a OSM čvorovi crvenom. Neki objekti (sivi objekti koji većinom predstavljaju zgrade ili kuće) nisu obilježeni kao čvorovi s žigom jer u OSM-u nije svaka zgrada označena kao čvor, a svaki čvor ne mora imati žig.

Poglavlje 3

Kategorizacija podataka

U ovom poglavlju biti će opisan postupak pripreme podataka za razvoj vjerojatnosnog modela. Kao što je već navedeno, razlikujemo dvije osnovne vrste podataka: podatke dobivene iz OSM-a koji će u ovom radu predstavljati stvarne fizičke objekte svrstane u neku kategoriju, te podatke dobivene iz GPS podataka o kretanju taksija koji će predstavljati posjete fizičkim objektima. Prije nego možemo razviti model za vjerojatnost posjeta kategorijama objekata, potrebno je definirati što su kategorije objekata, kako se pridjeljuju objektima i kako se konkretnom posjetu pridružuje objekt i kategorija. Također će biti razvijeni algoritmi za kategorizaciju i filtraciju podataka, te će ti algoritmi biti primjenjeni na stvarne podatke opisane u poglavlju 2.

3.1 Kategorije

Jedan od ciljeva ovog rada je omogućiti bolji uvid u svrhu kretanja ljudi. Kako bi to bilo moguće, treba pogledati razloge i motivaciju iza urbanih migracija. Premda na migracije stanovništa mogu utjecati mnogi faktori (generalna urbanizacija područja, političke prilike, etničko porijeklo, povijesni razlozi itd.) u ovom radu bavimo se migracijama na dnevnoj bazi (unutar 24h) pa su stoga i razlozi koje promatramo kao relevantne, oni svakodnevne prirode (npr. odlazak na posao, u školu, povratak kući, izlazak s prijateljima, odlazak do dućana). Uzimajući u obzir aktivnosti koje ljudi uobičajeno rade tokom prosječnog dana, analizom literature i razgovorom s ekspertima ustanovljeno je kako ih za potrebe ovog projekta možemo podijeliti u 6 osnovnih kategorija, koje ujedno predstavljaju i svrhu ili cilj ljudskog djelovanja:
Dom (Home), Posao (Work), Zdravlje (Health), Edukacija (Education), Zabava (Leisure) i Ostalo (Other).

Možemo reći da kategorija kojoj neka migracija pripada određuje prirodu te migracije i obrnuto. Ako je cilj neke migracije odlazak na posao, tada ona očito pripada kategoriji

Work. Ako je cilj izlazak u kazalište, tada migracija pripada kategoriji Leisure. Međutim, postavlja se pitanje kako za konkretnu migraciju odrediti kategoriju kojoj pripada ako a priori ne znamo njenu prirodu? Odgovor leži upravo u prostornim podacima. Većina migracija će za dolaznu točku imati neki položaj u blizini konkretnog fizičkog objekta. Ako pretpostavimo da je

(a). taj objekt cilj migracije

(b). taj objekt ima kategoriju

tada možemo reći da je kategorija te migracije ista kao i kategorija objekta koji je njen cilj.

U poglavlju 3.2 bavit ćemo se pitanjem kako odrediti kategoriju objekta, dok će u poglavlju 3.3 biti detaljnije opisana veza između posjeta objektu i kategorije objekta.

3.2 Kategorizacija i filtriranje prostornih objekata

U poglavlju 2.1 opisana je struktura OSM podataka. Poznato je da neki objekti, bilo da se radi o čvorovima, putevima ili relacijama, imaju žig. Premda zbog kolaborativne prirode samog OSM objekta žigovi mogu biti bilo kakav tekstualni objekt, u većini slučajeva žigovi ipak služe kao opis funkcije objekta, pogotovo ako se radi o žigovima na čvorovima. Upravo na ovoj osobitosti OSM žigova bazirat će se kategorizacija prostornih objekata. U nastavku će biti opisana tri načina na koje će OSM čvorovi biti razvrstani u 6 kategorija spomenutih u prethodnom potpoglavlju: Home, Work, Leisure, Education, Health i Other. Također će biti opisan algoritam koji primjenjuje sva tri načina razvrstavanja kako bi kategorizirao i filtrirao podatke iz poglavlja 2.1.

Kategorizacija pomoću vrijednosti (engl. tag value)

U ovom radu, kategorizacija OSM čvorova učinjena je pomoću vrijednosti pridijeljenih žigova. Premda iz definicije žiga ključ opisuje temu ili kategoriju objekta, te teme nisu uvijek u skladu s kategorijama relevantnim za ovo istraživanje. Također, postoje objekti čiji žigovi imaju isti ključ, ali na temelju vrijednosti ih želimo svrstati u različite kategorije. Stoga se prvo gleda „finiji" opis objekata, onaj baziran na vrijednosti žiga. Ovim načinom objekti se kategoriziraju u sve kategorije osim u kategoriju Other, iz razloga što Other kategorija predstavlja upravo one objekte koji nisu razvrstani u ostalih pet kategorija. Ideja je sljedeća: za svaku kategoriju (osim Other) postoji skup vrijednosti žigova koji pripada toj kategoriji. Taj skup zvat ćemo *skup vrijednosti kategorije*. Ako neki čvor ima žig čija se vrijednost nalazi u skupu vrijednosti neke kategorije, tada taj objekt pripada toj kategoriji. Skup vrijednosti svake kategorije određen je tako da u njega ulaze vrijednosti žiga pobrojane u poglavlju 3.1 koje svojim značenjem odgovaraju opisu te kategorije, u

skladu s opisima navedenim u 3.1. U nastavku se nalazi tablica 3.1 koja prikazuje skup vrijednosti za svaku kategoriju (osim Other).

Home	Leisure	Education	Health	Work
apartments	bar	college	clinic	commercial
house	pub	school	doctors	industrial
residential	cafe	university	hospital	building
apartment	restaurant	kindergarten	dentist	office
house_number	cinema	library	pharmacy	yes
	nightclub	books	chemist	government
	theatre	archive	hearing_aids	bank
	fast_food	music_school	herbalist	coworking_space
	playground	driving_school	medical_supply	works
	park	language_school	optician	
	department_store	research_institute	nursing_home	
	museum		social_facility	
	mall		blood_donation	
	clothes		first_aid_kit	
	boutique		ambulance_station	
	art			
	music			
	video			
	video_games			
	community_centre			

Tablica 3.1: Tablica skupa vrijednosti kategorija

Kategorizacija pomoću ključa (engl. tag key)

Kategorizacija pomoću vrijednosti ključa je dobra, ali ne uključuje sve objekte koji bi mogli (i trebali) spadati u ranije spomenute kategorije. Ranije je spomenuto da ne možemo isključivo kategorizirati objekte pomoću ključa. Nakon razmatranja, zaključeno je kako karakterizacija objekata pomoću ključa predstavlja primjeren pristup za potrebe ovog rada.

Konkretno, pomoću ključa će biti kategorizirane tri kategorije: Work, Leisure i Other. Za ostale kategorije nema ključeva koji bi zadovoljili uvjet da svi objekti s žigom takvog ključa pripadaju nekoj kategoriji. Analogno kategorizaciji pomoću vrijednosti, kod kategorizacije pomoću ključa definiramo skup ključeva kategorije. Ako neki čvor ima žig čiji ključ se nalazi u skupu ključeva neke kategorije, tada taj objekt pripada toj kategoriji. Tablica 3.2 prikazuje skupove ključeva kategorija.

Leisure	Work	Other
leisure	office craft	amenity building shop tourism

Tablica 3.2: Tablica skupa ključeva kategorija

Kategorizacija pomoću roditeljskog puta

U prva dva načina kategorizacije kategorizirani su bili čvorovi koristeći direktno njihove žigove. Na prvi pogled se čini da je tako moguće kategorizirati sve relevantne čvorove. Nažalost, u praksi to nije uvijek tako. Budući da pravila označavanja u OSM sustavu nisu stroga i više je načina na koje se određene stvari mogu označavati, samo gledanje žigova čvorova nije dovoljno. U ovom radu konkretan problem predstavljali su objekti Home kategorije. Stambene zgrade i kuće koje bi trebale pripadati u ovu kategoriju mogu biti označene žigom s vrijednosti *residential*. Međutim, takvih objekata u eksperimentalnim podacima gotovo da i nije bilo. Drugi način označavanja stambenih objekata jest označavanje cijelih puteva kao rezidencijalnih. Tako dolazimo do kategorizacije pomoću roditeljskog puta.

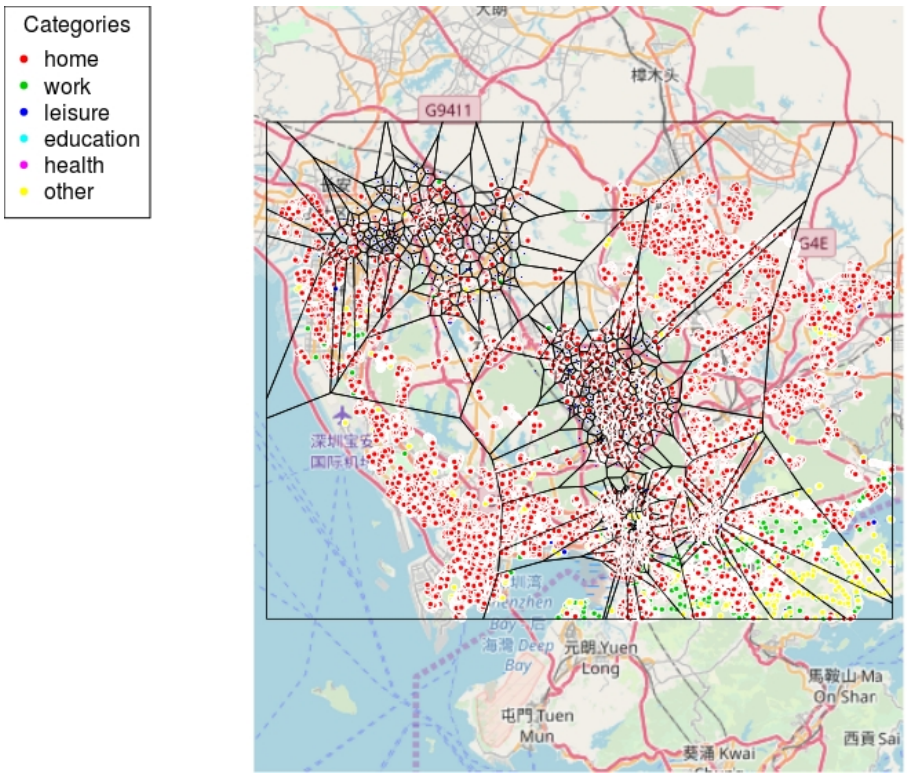
Ova kategorizacija koristi se samo za raspodjelu objekata u Home kategoriju. Ovaj put, promatraju se putevi unutar ćelije koji imaju žig s vrijednosti *residential*. Zatim se traže čvorovi na tim putevima koji se nalaze unutar ćelije (jer put može prolaziti kroz više ćelija). Čvorovi koji zadovoljavaju uvjete da se nalaze na rezidencijalnom putu i unutar ćelije, svrstavaju se u kategoriju Home.

Potrebno je naglasiti da nijedna metoda kategorizacije opisana u ovom radu neće uvijek i u potpunosti zahvatiti sve relevantne objekte, ali cilj je pokriti što veći broj objekata. Nadalje, metode kategorizacije uvelike ovise o konkretnim podacima. Ključevi i vrijednosti žigova koji se koriste u metodama spomenutim u ovom radu, a time i kategorije u koje se objekti svrstavaju pojedinom metodom, mogu se (i trebaju) mijenjati sukladno informacijama koje se želi dobiti iz podataka, kao i ovisno o prostoru na kojem se podaci prikupljaju i tome kako su na tom prostoru objekti unutar OSM-a označeni.

Algoritam za kategorizaciju prostornih objekata

U poglavlju 2.1 prikazana je ekstrakcija podataka iz OSM-a te preraspodjela prostornih OSM objekata u Voronoieve ćelije. Sada ćemo na te podatke primijeniti gore opisane metode kategorizacije.

Za svaku ćeliju postupak kategorizacije se primjenjuje posebno. Koriste se dvije funkcije: glavna u kojoj se odvija sama kategorizacija i pomoćna koja uklanja objekt (određen njegovom identifikacijskom oznakom) iz skupa objekata u ćeliji.
Pomoćna funkcija prima vektor identifikacijskih oznaka objekata koje je potrebno izbaciti, te osmar objekt u programskom okruženju za statističko računarstvo R (iz knjižnice *osmar*) iz kojeg se objekti izbacuju. Budući da osmar objekt predstavlja listu, objekti određeni identifikacijskim oznakama u ulaznom vektoru izbacuju se iz osmar objekta koristeći bazne mehanizme rada s listama u programskom okruženju za statističko računarstvo R. Točnije, objekti se izbacuju iz podliste u osmar objektu koja predstavlja čvorove. Zatim funkcija vraća promijenjeni osmar objekt, koji više ne sadrži neželjene objekte.
Unutar ćelije, objekti se prvo kategoriziraju redom u Home, Education, Leisure, Health i Work kategorije koristeći metodu kategorizacije pomoću vrijednosti žiga. Preostali objekti zatim se pomoću ključa žiga kategoriziraju u Work, Leisure i Other kategorije. Za objekte koji su ostali nekategorizirani koristi se metoda roditeljskog puta da ih se svrsta, ako je to moguće, u Home kategoriju. Detaljniji postupak prikazan je u algoritmu 1.



Slika 3.1: OSM objekti podijeljeni po kategorijama

Algoritam 1 Algoritam kategorizacije objekata

1 big_map = skup svih OSM objekata

2 Za svaku ćeliju:

    1 categories = (Home, Education, Leisure, Health, Work)

    2 map = skup OSM čvorova u ćeliji

    3 Za c u categories:

        i. values = skup vrijednost kategorije c

        ii. nađi objekte u map koji imaju vrijednost žiga iz skupa values

        iii. spremi longitude i latitude tih objekata

        iv. spremi broj objekata kategorije c

        v. izbaci objekte iz map

    4 Za c u (work,leisure,other):

        i. keys = skup ključeva kategorije c

        ii. nađi objekte u map koji imaju ključ žiga iz skupa keys

        iii. spremi longitude i latitude tih objekata

        iv. povećaj broj objekata kategorije c

        v. izbaci objekte iz map

    5 ids = identifikacijske oznake preostalih objekata u map

    6 w1 = u big_map pronađi čvorove s id-evima iz ids, te njihove roditeljske strukture

    7 w2 = u big_map pronađi puteve s vrijednošću ključa *residential*

    8 nađi puteve u presjeku w1 i w2

    9 nađi čvorove unutar tih puteva

    10 spremi longitude i latitude tih objekata

    11 povećaj broj objekata kategorije Home

Filtriranje prostornih objekata

U prethodnom koraku kategorizacije, rezultat je lista u kojoj su pohranjeni podaci o kategorijama prostornih objekata unutar svake ćelije, kao i podatak koliko objekata koje kategorije se nalazi u svakoj ćeliji. Za daljnje korake, nisu sve ćelije jednako relevantne. Neke ćelije imaju premalo kategoriziranih OSM objekata ili ih nemaju uopće. Takav slučaj se može pojaviti zbog prirode označavanja objekata u OSM-u i takve ćelije ne sadrže dovoljno podataka za daljnje zaključivanje. S druge strane, neke ćelije sadrže ogromnu količinu Home objekata, zbog metode roditeljskog puta koja se koristi za njihovo kategoriziranje, a koja je manje precizna od ostalih metoda. Premda čvorovi unutar residencijalnog puta u pravilu označavaju fizički objekt, ponekad mogu služiti samo kao "kostur" na kojem se gradi put. Stoga u stvarnosti nema toliko objekata Home kategorije, i korištenje ćelija s previše takvih objekata ne bi dalo točne rezultate. Stoga listu s podacima o kategorijama prostornih objekata filtriramo tako da nastavak postupka obuhvati:

- samo ćelije koje nemaju više od 100 Home objekata te
- imaju barem 5 ili više kategoriziranih objekata

Korištene knjižnice i funkcije

U procesu kategorizacije i filtriranja prostornih objekata korišteni su, uz osnovne knjižnice i osnovne funkcije iz programskog okruženja za statističko računarstvo R, sljedeće funkcije iz knjižnice **osmar**:

- **find(object, condition):** funkcija za traženje elemenata u objektu *object* koji zadovoljavaju uvjet *condition*
  object je osmar objekt
  condition je logički uvjet, traže se elementi ili reci koji ga zadovoljavaju. Mora se definirati na kojim elementima i podacima osmar objekta se traženje primjenjuje, u obliku element(data(condition)). Element može biti čvor, put ili relacija, a data je, ovisno o elementu, žig ili atribut.
  Povratna vrijednost je vektor identifikacijskih oznaka elemenata u objektu *object* koji zadovoljavaju uvjet *condition*
- **find_up(object, ids):** funkcija za traženje svih elemenata koji su u hijerarhiji iznad elemenata određenih identifikacijskim oznakama u *ids*
  object je osmar objekt
  ids je vektor identifikacijskih oznaka elemenata, neovisno o tome jesu li čvor, put ili relacija.
  Povratna vrijednost je vektor identifikacijskih oznaka elemenata

- **find_down(object, ids):** funkcija za traženje svih elemenata koji su u hijerarhiji ispod elemenata određenih id-evima u *ids*
  object je osmar objekt
  ids je vektor id-eva elemenata, neovisno o tome jesu li čvor, put ili relacija.
  Povratna vrijednost je vektor id-eva elemenata

3.3 Kategorizacija taksi podataka

U ovom poglavlju bit će objašnjen algoritam za kategorizaciju taksi podataka. Taksi podaci predstavljaju stvarne lokacije na koje ljudi dolaze. Upravo ti podaci služit će kako bi se predvidjela vjerojatnost posjeta pojedinim kategorijama objekata. U tu svrhu, potrebno je odrediti kategoriju kojoj pripada svaka lokacija iz taksi podataka, odnosno odrediti kategoriju objekta koji je pritom posjećen.

Definicija 3.3.1 (Kategorija taksi podatka)**.** *Kategorija taksi podatka, određenog geografskom širinom i dužinom, je jednaka kategoriji onog objekta koji je geografski najbliži taksi podatku.*

Haversineova formula

Pretpostavimo kako su zadana dva podatka određena geografskom širinom i dužinom (dva položaja određena na jedinstven način svojim koordinatama u zajedničkom unaprijed definiranom koordinatnom sustavu). Postavlja se problem određivanja udaljenosti ovih dvaju položaja ukoliko se oni nalaze na Zemljinoj površini (plohi geoda, geometrijskog tijela koje opisuje Zemlju). Moguće rješenje problema postiže se primjenom haversineove formule.

Definicija 3.3.2 (Haversineova formula)**.** *Haversineova formula je izraz koji računa najkraću udaljenost između dvije točke (zadane geografskom dužinom i širinom) na površini sfere.*

Računanje udaljenosti pomoću haversineove formule dano je na sljedeći način:

$$a = sin^2(\Delta\phi/2) + cos\phi_1 * cos\phi_2 * sin^2(\Delta\lambda/2)$$
$$c = 2 * atan2(\sqrt{a}, \sqrt{1-a})$$
$$d = R * c$$

gdje je
ϕ... zemljopisna širina (u radijanima)
λ ... zemljopisna duljina (u radijanima)

R ... Zemljin radijus ($\approx$ 6 371km)

Filtriranje podataka i algoritam za kategorizaciju

Za kategorizaciju taksi podataka koristit će se rezultati iz poglavlja 3.2, te taksi podaci opisani u poglavlju 2.2.

Algoritam se za svaku ćeliju primjenjuje posebno, međutim ne gledaju se sve ćelije nego samo odabrane (relevantne) ćelije dobivene postupkom filtriranja objašnjenom u poglavlju 3.2. Za pojedinu ćeliju prvo se dobavljaju taksi podaci samo za tu ćeliju, te se zatim primjenjuje funkcija *distanceCalculation* za izračunavanje udaljenosti i kategorije najbližeg objekta za svaki taksi podatak. Ova funkcija prima matricu taksi podataka za trenutnu ćeliju, listu matrica sa kategoriziranim OSM objektima za svaku ćeliju (nastalu kako je opisano u 3.2) te identifikacijsku oznaku trenutne ćelije. Rad funkcije detaljnije je opisan u algoritmu 2.

Rezultat ove funkcije je matrica udaljenosti, tj. matrica dimenzija $n \times 5$, gdje n predstavlja broj taksi objekata unutar pojedine ćelije, a stupci su redom: zemljopisna širina taksi podatka, zemljopisna dužina taksi podatka, udaljenost, kategorija i timeframe najbližeg prostornog objekta.

Budući da je cilj kategorizacije taksi podataka dobiti kategoriju objekata koje ljudi doista posjećuju, potrebno je voditi računa o tome da podaci budu smisleni. Ako je primjerice nekom taksi podatku najbliži prostorni objekt udaljen nekoliko kilometara, nije realno pretpostaviti da će putnik taksija doista posjetiti taj objekt. S druge strane, tako veliku udaljenost taksi objekta od prostornog objekta moguće je dobiti u podacima zbog više razloga: nisu svi objekti označeni na OSM-u, neki objekti se odbacuju prilikom kategorizacije prostornih objekata i zbog prirode taksi podataka. Kako bi se smanjila pogreška uzrokovana ovakvim scenarijima koji u stvarnosti nisu izgledni, matrica udaljenosti se filtrira tako da se odbacuju svi reci koji sadrže taksi podatke čiji je najbliži prostorni objekt udaljeniji od 100 metara.

Sve matrice udaljenosti spremaju se u listu matrica udaljenosti. Lista matrica udaljenosti koristit će se u koraku predviđanja posjeta kategorijama prostornih objekata. Prije toga se filtrira tako da se uzimaju one matrice koje imaju više od 200 redaka (svaki redak predstavlja detalje o jednom taksi objektu) i one koje sadrže podatke o taksi objektima iz više od jedne kategorije. Ovaj korak je potreban kako bi algoritmi za učenje koji će se kasnije koristiti dali bolje rezultate. Želimo se osigurati da za svaku matricu udaljenosti imamo dovoljno podataka i da ti podaci budu raznoliki (ako su svi iste kategorije, nemamo što predviđati).

Algoritam 2 Algoritam funkcije distanceCalculation

1. ako je matrica taksi podataka prazna, vrati prazan data frame
2. inicijaliziraj prazne vektore minimum, category i timeframe
3. inicijaliziraj matricu udaljenosti dist_matrix sa longitude i latitude vrijednostima svih podataka iz matrice taksi podataka
4. za svaki taksi objekt:
    1. dist_vector = izračunaj Haversine udaljenost taksi objekta od svih prostornih objekata
    2. pronađi najmanju udaljenost u dist_vector-u
    3. pronađi prostorni objekt koji ima najmanju udaljenost od taksi objekta
    4. pronađi kategoriju prostornog objekta koji ima najmanju udaljenost
    5. odredi timeframe taksi objekta
    6. vektorima minimum, category i timeframe dodaj (respektivno) najmanju udaljenost, kategoriju najmanje udaljenog objekta i timeframe najmanje udaljnog objekta
5. matrici dist_matrix dodaj stupce minimum, category i timeframe
6. dist_matrix = dist_matrix bez redaka gdje je minimum veći od 100
7. vrati dist_matrix

Korištene knjižnice i funkcije

U procesu kategorizacije i filtriranja taksi objekata korištene su većinom U procesu kategorizacije i filtriranja prostornih objekata korištene su većinom funkcije iz osnovnih knjižnica iz programskog okruženja za statističko računarstvo R. Također je za računanje haversineove udaljenosti bila korištena funkcija iz knjižnice **geosphere**:

- **distHaversine(p1, p2, r=6378137):** funkcija za računanje udaljenosti između dvije točke koristeći haversineovu formulu
  p1 je longitude/latitude točke ili točaka. Može biti vektor dva broja (koji predstavljaju longitude i latitude), matrica s dva stupca (prvi su longitude vrijednosti točaka, drugi latitude vrijednosti) ili SpatialPoints objekt
  p2 isto kao i *p1*

r je radijus sfere, po defaultu predstavlja radijus Zemlje i iznosi 6378137 metara
Povratna vrijednost je udaljenost u istoj mjernoj jedinici kao što je zadan *r* (po defaultu je mjerna jedinica metar). Ako su i *p1* i *p2* zadani kao vektori, povratna vrijednost će biti broj koji predstavlja udaljenost između točaka *p1* i *p2*. Ako je *p1* zadana kao vektor, a *p2* kao matrica, povratna vrijednost će biti vektor, u kojem je svaki broj udaljenost između točke *p1* i jedne od točaka u matrici *p2*.

Poglavlje 4

Vjerojatnosni pristup klasifikaciji podataka

4.1 Logistička regresija

Kada se govori o problemima koji se rješavaju statističkim učenjem [16], oni se mogu podijeliti na regresijske i klasifikacijske probleme. Kada su varijable koje se predviđaju kontinuirane naravi, tada se govori o regresiji. Kada su varijable diskretne naravi, govori se o klasifikaciji. Logistička regresija, premda sadrži riječ regresija u sebi, je metoda koja rješava klasifikacijske probleme. Klasifikacijske modele nadalje možemo podijeliti na binarne i multinomijalne. Binarni klasifikacijski modeli imaju za uvjet Bernoulijevu distribuciju izlaznih varijabli, dok multinomijalni imaju multinoulli distribuciju (multinoulli distribucija još se naziva i generalizirana Bernoulijeva distribucija). U literaturi se pojam logistička regresija često poistovjećuje s binomnom logističkom regresijom koja spada u binarne klasifikacijske modele. Osim toga, postoji i multinomijalna logistička regresija koja spada u multinomijalne klasifikacijske modele. Općenito, logistička regresija ne modelira direktno vrijednosti zavisne varijable, nego modelira vjerojatnost da zavisna varijabla pripada određenoj kategoriji. Zbog lakšeg razumijevanja, u potpoglavlju 3.1.1 će prvo biti opisana binomna logistička regresija, a u potpoglavlju 3.1.3 će ona biti poopćena na multinomijalnu logističku regresiju. Na kraju će biti prikazano kako je multinomna logistička regresija primjenjena na podacima opisanim u prijašnjim poglavljima.

Binomna logistička regresija

Neka je

$$(x_1, y_1), ..., (x_n, y_n)$$

skup opaženih podataka. Izlazne ili zavisne varijable označavat ćemo s $y_i, i = 1, ..., n$. Ulazne ili nezavisne varijable su $1 \times k, k \in \mathbb{N}$, vektori i označavat ćemo ih s $x_i, i = 1, ..., n$. Na ovom skupu ćemo graditi klasifikator. Izlazne varijable mogu poprimiti vrijednosti $c_1, ..., c_J$. Budući da se radi o binarnom modelu, vrijedi $J = 2, c_1 = 1, c_2 = 0$. Također pretpostavljamo da postoji J funkcija $f_1, ..., f_J$ takvih da je

$$P(y_i = c_j|x_i) = f_j(x_i, \theta), i = 1, ..., n, j = 1, ..., J$$

Primijetimo da uvjetna vjerojatnost ne ovisi samo o opaženim izlaznim vrijednostima, nego i o vektoru parametara θ.

Sada nas zanima oblik funkcija f_1 i f_2. Budući da se radi o vjerojatnostima, funkcije moraju biti nenegativne i davati 1 u sumi. Formalno, to možemo zapisati:

$$f_j(x_i, \theta) \geq 0, j = 1, ..., J$$

$$\sum_{j=1}^{J} f_j(x_i, \theta) = 1,$$

za svaki par (x_i, θ).
Vjerojatnost da neka izlazna varijabla y poprima vrijednost 1, uz ulaznu varijablu x možemo izraziti kao

$$P(y = 1|x) = p(x)$$

Vrijednosti $p(x)$ se nalaze u intervalu između 0 i 1. Postavlja se pitanje kako modelirati vezu između $p(x)$ i x. Koristeći model linearne regresije, tu vezu bismo mogli prikazati kao:

$$p(x) = \beta_0 + \beta_1 x$$

Problem s ovakvim pristupom je u tome što će kada x poprima vrijednosti blizu nuli, $p(x)$ poprimati negativne vrijednosti, a za jako velike x, $p(x)$ će poprimati vrijednosti veće od 1. Drugim riječima, vrijednosti $p(x)$ neće se nužno nalaziti u intervalu između 0 i 1. Kako bi se ovaj problem izbjegao, $p(x)$ treba modelirati pomoću funkcije koja uvijek daje vrijednosti između 0 i 1. Primjer takve funkcije je logistička funkcija

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{4.1}$$

Primijetimo da za vrlo male vrijednosti x logistička funkcija poprima vrijednosti blizu nule, ali nikad ispod, dok se za velike vrijednosti x funkcija približava vrijednosti 1, ali ju nikad ne prelazi. Graf logističke funkcije uvijek je oblika slova S te stoga bez obzira na vrijednost x uvijek daje smislene predikcije. Nakon malo računa dobije se

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0+\beta_1 x} \tag{4.2}$$

Vrijednost $\frac{p(x)}{1-p(x)}$ naziva se *omjer šansi(odds)* i poprima vrijednosti između 0 i ∞. Šansa blizu 0 označava vrlo malu, a šansa blizu ∞ vrlo veliku vjerojatnost varijable y.
Kada logaritmiramo obje strane jednakosti, dobivamo

$$log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x \tag{4.3}$$

Lijeva strana jednadžbe naziva se *logit*. Vidljivo je da logistički regresijski model ima logit linearan u x. U linearnom regresijskom modelu, β_1 određuje prosječnu stopu promjene u y povezanu s jediničnim porastom vrijednosti x. Drugim riječima, povećanje x za jednu jedinicu mjere povećat će $p(x)$ za β_1. Za razliku od toga, u logističkom regresijskom modelu povećanje vrijednosti x za jednu jedinicu promijenit će logit za β_1 odnosno pomnožiti će omjer šansi sa e^{β_1}. Međutim,budući da veza između x i $p(x)$ nije linearna, promjena vrijednosti $p(x)$ izazvana jediničnom promjenom u x ovisit će o trenutnoj vrijednosti varijable x. Jedino što se sa sigurnošću može reći jest sljedeće: bez obzira na vrijednost x,

- ako je β_1 pozitivan, tada će porast vrijednosti x uzrokovati porast $p(x)$
- ako je β_1 negativan, tada će porast vrijednosti x uzrokovati smanjenje $p(x)$

ako je β_1 pozitivan, tada će porast vrijednosti x uzrokovati porast $p(x)$.

Ako sada logističku funkciju izrazimo u općenitijem obliku, kao

$$S(t) = \frac{1}{1+e^{-t}}$$

možemo primijetiti da vrijedi

$$P(y_i = 1|x_i) = p(x_i) = S(x_i\beta)$$

gdje je y_i izlazna varijabla, x_i ulazna varijabla, a β vektor koeficijenata.
Također vrijedi

$$P(y_i = 0|x_i) = 1 - S(x_i\beta)$$

Iz svega ovoga je sada vidljivo da uz definiciju $\theta = \beta$ vrijedi

$$\begin{aligned} f_1(x_i, \theta) &= P(y_i = 1|x_i) = S(x_i\beta) \\ f_2(x_i, \theta) &= P(y_i = 0|x_i) = 1 - S(x_i\beta) \end{aligned} \tag{4.4}$$

Metoda maksimalne vjerodostojnosti

Koeficijenti β_0 i β_1 su nepoznati i potrebno ih je procijeniti na temelju podataka za treniranje. Pretpostavit ćemo također da se procjena temelji na neovisnim i jednako distribuiranim

podacima. Kod linearne regresije u tu svrhu se koristi metoda najmanjih kvadrata, ali kod logističke regresije uobičajena metoda je metoda maksimalne vjerodostojnosti (maximum likelihood). Kod metode maksimalne vjerodostojnosti, pokušavamo pronaći $\hat{\beta}_0$ i $\hat{\beta}_1$ takve da ubacivanje tih procjena u 4.1 daje približno 1 za one vrijednosti koje poprimaju kategoriju 'da' i približno 0 za one koje poprimaju kategoriju 'ne'. Ovaj problem može se formalizirati koristeći funkciju vjerodostojnosti [18].

Definicija 4.1.1 (Funkcija vjerodostojnosti)**.** *Neka je* $(x_1, ..., x_n)$ *opaženi uzorak za slučajnu varijablu* X *s gustoćom* $f(x|\theta)$*, gdje je* $\theta = (\theta_1, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$ *nepoznati parametar. Definiramo funkciju vjerodostojnosti* $L : \Theta \to \mathbb{R}$ *sa*

$$L(\theta) := f(x_1|\theta)...f(x_n|\theta), \theta \in \Theta.$$

Vrijednost $\hat{\theta} = \hat{\theta}(x_1, ..., x_n) \in \Theta$ *za koju je*

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

zovemo procjena metodom maksimalne vjerodostojnosi.
Statistika $\hat{\theta}(X_1, ..., X_n)$ *je procjenitelj metodom maksimalne vjerodostojnosti (MLE)*

Sada možemo dobiti funkciju vjerodostojnosti za opažanje (y_i, x_i) :

$$L(\beta; y_i, x_i) = [S(x_i\beta)]^{y_i}[1 - S(x_i\beta)]^{1-y_i} \tag{4.5}$$

Označimo $n \times 1$ vektor svih izlaznih varijabli s y, a $n \times k$ matricu svih ulaznih varijabli s X. Uz pretpostavku da su opažanja neovisna i jednako distribuirana, funkcija vjerodostojnosti cijelog uzorka je jednaka umnošku vjerodostojnosti pojedinih opažanja, odnosno:

$$L(\beta; y, X) = \prod_{i=1}^{n} [S(x_i\beta)]^{y_i}[1 - S(x_i\beta)]^{1-y_i} \tag{4.6}$$

U nastavku dajemo izvod logaritmirane funkcije vjerodostojnosti za logistički model:

$$
\begin{aligned}
l(\beta; y, X) &= ln(L(\beta; y, X)) \\
&= ln(\prod_{i=1}^{n} [S(x_i\beta)]^{y_i} [1 - S(x_i\beta)]^{1-y_i}) \\
&= \sum_{i=1}^{n} [y_i ln(S(x_i\beta)) + (1 - y_i) ln(1 - S(x_i\beta))] \\
&= \sum_{i=1}^{n} [y_i ln(\frac{1}{1 + e^{-x_i\beta}}) + (1 - y_i) ln(1 - \frac{1}{1 + e^{-x_i\beta}})] \\
&= \sum_{i=1}^{n} [y_i ln(\frac{1}{1 + e^{-x_i\beta}}) + (1 - y_i) ln(\frac{1 + e^{-x_i\beta} - 1}{1 + e^{-x_i\beta}})] \\
&= \sum_{i=1}^{n} [y_i ln(\frac{1}{1 + e^{-x_i\beta}}) + (1 - y_i) ln(\frac{e^{-x_i\beta}}{1 + e^{-x_i\beta}})] \\
&= \sum_{i=1}^{n} [ln(\frac{e^{-x_i\beta}}{1 + e^{-x_i\beta}}) + y_i (ln(\frac{1}{1 + e^{-x_i\beta}}) - ln(\frac{e^{-x_i\beta}}{1 + e^{-x_i\beta}}))] \\
&= \sum_{i=1}^{n} [ln(\frac{e^{-x_i\beta}}{1 + e^{-x_i\beta}} \frac{e^{x_i\beta}}{e^{x_i\beta}}) + y_i (ln(\frac{1}{1 + e^{-x_i\beta}} \frac{1 + e^{-x_i\beta}}{e^{-x_i\beta}}))] \\
&= \sum_{i=1}^{n} [ln(\frac{1}{1 + e^{x_i\beta}}) + y_i (ln(\frac{1}{e^{-x_i\beta}}))] \\
&= \sum_{i=1}^{n} [ln(1) - ln(1 + e^{x_i\beta}) + y_i (ln(1) - ln(e^{-x_i\beta})))] \\
&= \sum_{i=1}^{n} [-ln(1 + e^{x_i\beta}) + y_i x_i\beta]
\end{aligned}
$$

Dakle, logaritmirana funkcija vjerodostojnosti glasi:

$$
l(\beta; y, X) = \sum_{i=1}^{n} [-ln(1 + e^{x_i\beta}) + y_i x_i\beta] \tag{4.7}
$$

Da bismo procijenili koeficijente β_0 i β_1 tj. vektor koeficijenata β potrebno je pronaći procjenitelj metodom maksimalne vjerodostojnosti $\hat{\beta}$. $\hat{\beta}$ dobiva se kao rješenje problema:

$$
\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, l(\beta; y, X) \tag{4.8}
$$

Analitičko rješenje ovog maksimizacijskog problema postoji za neke specifične slučajeve, ali općenito ne postoji, nego se rješenje može pronaći samo numerički. Štoviše, ovaj problem uopće ne mora imati rješenje u slučaju da se model savršeno podudara s opaženim klasama. Ipak, takva situacija je izuzetno rijetka i u gotovo svim slučajevima postoji rješenje [25].

Multinomna logistička regresija

Notacija kod multinomne logističke regresije ista je kao i kod binomne. Opet je (x_1, y_1),..., (x_n, y_n) skup opaženih podataka za koji vrijede iste oznake kao i kod binomne logističke regresije. Jedina razlika je skup klasa koje izlazne varijable mogu poprimiti, $c_1, ..., c_J$. U binarnom modelu vrijedilo je $J = 2, c_1 = 1, c_2 = 0$, dok u multinomnom klasa može biti i više. Posljedica toga je da sada umjesto jednog vektora koeficijenata β, svaka klasa c_j ima svoj vektor koeficijenata β_j. Iz toga sada slijedi:

$$f_j(x_i; \theta) = P(y_i = c_j | x_i) = \frac{e^{x_i \beta_j}}{\sum_{l=1}^{J} e^{x_i \beta_l}} \tag{4.9}$$

za $j = 1, ..., J$, uz $\theta = [\beta_1 ... \beta_J]$. Primijetimo da izlazne varijable y_i imaju multinoulli distribuciju, s vjerojatnostima $f_1(x_i; \theta), ..., f_J(x_i; \theta)$

Zadržavajući oznake iz binomnog modela, poopćiti možemo i funkciju vjerodostojnosti cijelog uzorka:

$$L(\theta; y, X) = \prod_{i=1}^{n} \prod_{i=1}^{J} [f_j(x_i; \theta)]^{y_{ij}} \tag{4.10}$$

te logaritmiranu funkciju vjerodostojnosti:

$$l(\theta; y, X) = \sum_{i=1}^{n} \sum_{i=1}^{J} ln(f_j(x_i; \theta)) y_{ij} \tag{4.11}$$

Daljnji postupak pronalaska procjenitelja $\hat{\theta}$ metodom maksimalne vjerodostojnosti isti je kao i za binomni model.

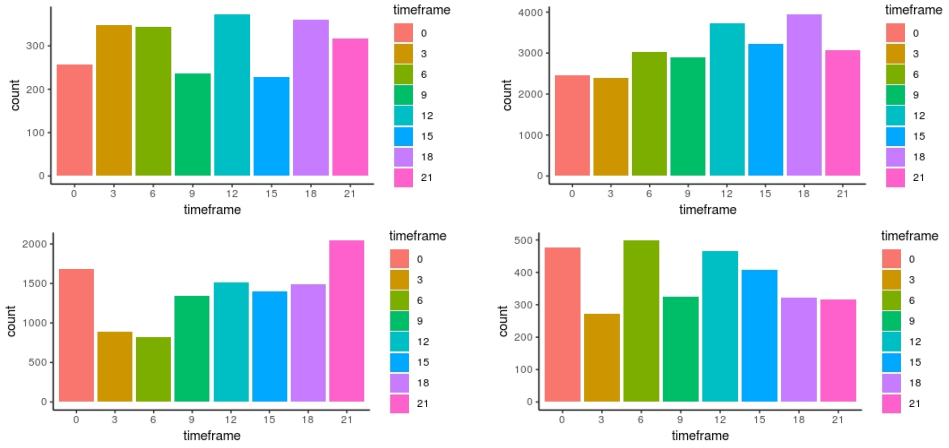
4.2 Primjena na eksperimentalnim podacima

U poglavlju 3.3 prikazan je postupak dobivanja liste matrica udaljenosti za relevantne ćelije. Svaka matrica udaljenosti sadrži podatke o taksi objektima unutar pripadajuće ćelije, udaljenosti taksi objekta od najbližeg kategoriziranog prostornog objekta, kategoriji tog prostornog objekta te vremenskom okviru taksi objekta. Ideja je zavisno o vremenskom okviru taksi objekta predvidjeti kategoriju kojoj taksi objekt pripada. Važno je napomenuti

da se longitude i latitude podaci o taksi objektu ne uzimaju u obzir.
Za ovakvo predviđanje koristit će se model multinomne logističke regresije, opisan u prethodnom poglavlju. Glavni razlog korištenja tog modela jest što on ne daje samo klase kojima taksi objekt pripada, nego i vjerojatnosti da pripada baš toj klasi. Drugim riječima, daje vjerojatnosnu distribuciju objekta po klasama.
Za svaku matricu udaljenosti, model ćemo raditi zasebno, jer se sastav ćelija međusobno jako razlikuje (broj objekata po klasama je bitno drugačiji, kao i broj klasificiranih objekata općenito). U daljnjem tekstu, dok ne bude navedeno drugačije, sve tvrdnje i postupci se odnose na podatke unutar jedne (proizvoljne) matrice udaljenosti.

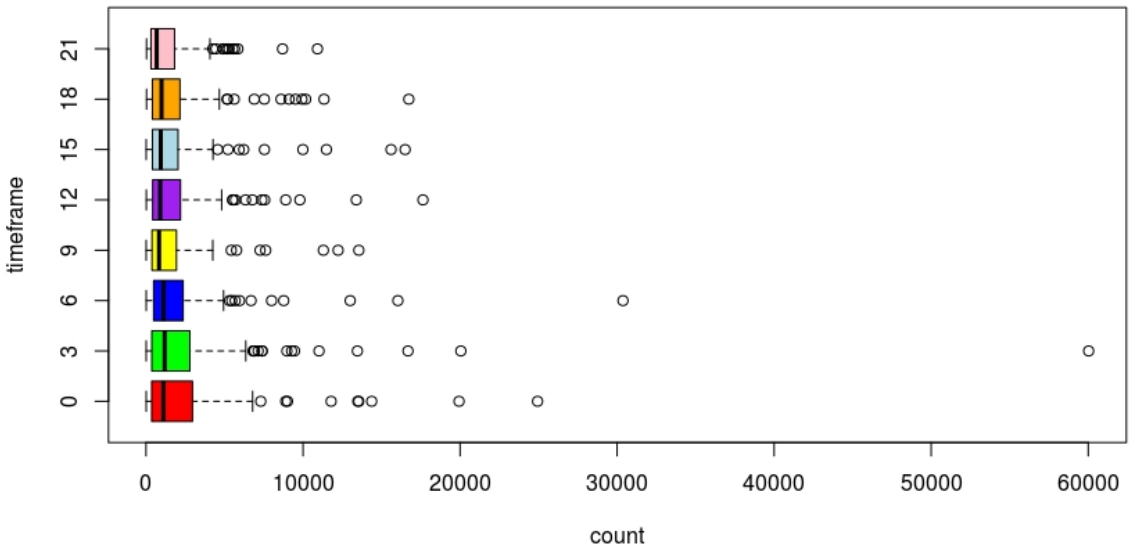
Priprema podataka

Podsjetimo se: matrica udaljenosti jest matrica dimenzija $n \times 5$, gdje je n broj taksi objekata, a stupci su redom: zemljopisna širina taksi podatka, zemljopisna dužina taksi podatka, udaljenost, kategorija i timeframe najbližeg prostornog objekta.
Označimo s $x_i, i = 1, ..., n$ timeframe i-tog taksi objekta, a s $y_i, i = 1, ..., n$ kategoriju i-tog taksi objekta. Sada je $(x_1, y_1), ..., (x_n, y_n)$ skup opaženih podataka. Na slici 4.1 prikazana je podjela ulaznih podataka po timeframeovima za četiri ćelije. Na slici 4.2 dan je boxplot



Slika 4.1: Raspodjela podataka po timeframeovima za različite ćelije

prikaz ulaznih podataka po timeframeovima za cijelo područje interesa. Varijable y_i mogu pripadati jednoj od šest klasa: Home, Education, Health, Leisure, Work ili Other. Mogli bismo te klase označiti brojevima od 1 do 6 te definirati skup klasa koje izlazne varijable mogu poprimiti kao $\{c_i = i\}, i = 1, ..., 6$. Međutim, ne sadrži svaka ćelija objekte svih kategorija. Stoga ćemo klase koje izlazne vrijednosti u ćeliji doista poprimaju označiti brojevima 1 do J, gdje je $J \leq 6$ broj klasa koje varijable doista poprimaju, a skup klasa koje



Slika 4.2: Boxplot prikaz svih podataka po timeframeovima

izlazne varijable poprimaju definirat ćemo kao $d_i = i, i = 1, ..., J$. Kako bi se to postiglo u programskom okruženju za statističko računarstvo R, korištenom u ovom radu, koriste se *faktori*. Faktor u programskom okruženju za statističko računarstvo R je tip podataka namijenjen prikazu kategoričkih podataka. Faktori se pohranjuju kao vektor cjelobrojnih vrijednosti s odgovarajućim skupom znakovnih vrijednosti koje se koriste pri prikazivanju faktora. Drugim riječima, faktori upravo označavaju klase (znakovne vrijednosti) različitim cjelobrojnim vrijednostima. Ako nije drugačije zadano, u programskom okruženju za statističko računarstvo R su te cjelobrojne vrijednosti upravo redom prirodni brojevi, počevši od jedinice.

Sljedeći korak je podjela skupa opaženih podataka na podatke za treniranje i testiranje. Podaci za treniranje služiti će za prilagodbu modela logističke regresije i oni čine 80% skupa opaženih podataka. Koji podaci iz skupa opaženih podataka će ući u skup za treniranje odabire se pomoću funkcije **createDataPartition**. Ova funkcija za faktor opaženih podataka, nasumično odabire podatke unutar pojedinih klasa tog faktora, kako bi se distribucija klasa u uzorku za treniranje zadržala. Više detalja o ovoj funkciji biti će navedeno u poglavlju 4.2.

Podaci koji ne ulaze u skup za treniranje, tvore skup za testiranje. Na skupu za testiranje će se provjeriti točnost, preciznost i osjetljivost modela, te izračunati greška. Tek na skupu za testiranje vidi se koliko je model 'dobar'.

Algoritam i korištene funkcije

Za svaku matricu udaljenosti, potrebno je izdvojiti stupce s podacima o timeframeu i kategoriji te ih pretvoriti u faktore, kako je opisano u potpoglavlju Priprema podataka. Zatim se na izlazne vrijednosti primjenjuje funkcija **createDataPartition** te se dobivaju podaci za trening i testiranje. Na trening podacima se funkcijom **multinom** gradi model multinomne logističke regresije, te se na testnim podacima prvo predviđa vjerojatnosna distribucija podataka po klasama, a zatim i klasifikacija svakog podatka. Uzima se da podatak pripada onoj klasi za koju ima najveću vjerojatnost da joj pripada. Za ova dva postupka koristi se funkcija **predict**. Osim toga računa se i matrica konfuzije te se rezultati spremaju u listu. Cijeli postupak (za jednu matricu udaljenosti) detaljnije je opisan u algoritmu 3:

Algoritam 3 Multinomijalna logistička regresija

```
1  dist_matrix$timeframe = faktor timeframe vrijednosti
2  dist_matrix$category = faktor category vrijednosti
3  trainingRows = createDataPartition(dist_matrix$category, p=.8, list = FALSE, times
   = 1)
4  training = trainingRows reci matrice udaljenosti
5  test = ostali reci matrice udaljenosti
6  multinomMOdel = multinom(category   ., data=training)
7  predicted_scores = predict (multinomModel, test, "probs")
8  predicted_class = predict (multinomModel, test)
9  confusion = table(predicted_class, test$category)
10 vrati list(predicted_scores,predicted_class,confusion)
```

Najvažnije funkcije koje su korištene su **createDataPartition** iz biblioteke **caret**, **multinom** i **predict** iz biblioteke **nnet**:

- **createDataPartition(y, times = 1, p = 0.5, list = TRUE, groups = min(5, length(y))):** stvara seriju test/training particija. Za bootstrap uzorke particije se biraju nasumičnim odabirom. Ako je y faktor, particije se biraju unutar pojedinih klasa faktora kako bi se zadržala distribucija klasa. Ako je y numerički, particije se biraju u podgrupama baziranim na percentilima.

y je vektor izlaznih varijabli
times je broj particija
p je postotak podataka koji idu u skup za treniranje
list je logički parametar, ako je TRUE rezultat će biti u listi, u suprotnom će biti matrica
groups ako je y numerički, broj prekida u kvantilima
povratna vrijednost je lista ili matrica koja sadrži redne brojeve redaka koji čine skup za treniranje

- **multinom(formula, data, weights, subset, ...):** gradi multinomijalni logistički model pomoću neuralnih mreža, pozivajući funkciju nnet.
  formula je izraz tipa *formula* (objekt u programskom okruženju R) za regresijski model, u obliku $response \sim predictors$. *response* treba biti faktor ili matrica s onoliko stupaca koliko ima klasa
  data je proizvoljni *data.frame* za interpretaciju varijabli iz parametra formula
  weights su opcionalne težine u gradnji modela
  subset je izraz koji govori koji podskup redaka treba uzeti za gradnju modela, po defaultu se uzimaju svi reci
  na.action je funkcija za filtriranje podataka koji nedostaju
  povratna vrijednost je nnet objekt s eventualnim dodatnim komponentama

- **predict(object, newdata, type = c(”raw”, ”class”, ”probs”), ...):** predviđa nove vrijednosti pomoću modela istreniranog na neuralnim mrežama
  object je objekt klase nnet koji je povratna vrijednost funkcije nnet
  newdata je matrica ili data.frame sa podacima iz testnog skupa
  type je vrsta izlaznog argumenta, može biti ”raw”, ”class” ili ”probs”
  ... dodatni argumenti iz drugih metoda
  povratna vrijednost je matrica vrijednosti koje vrati istrenirana mreža, ako je type = ”raw”, predviđene klase ako je type = ”class” i matrica vjerojatnosne distribucije po klasama ako je type = ”probs”

Rezultati

U listi matrica udaljenosti, izračunatoj u poglavlju 3.3, nakon filtriranja ostaje 144 matrica udaljenosti (za 144 ćelije) na koje se primjenjuje algoritam 3. Nakon toga, računaju se brojevi ispravno pozitivnih, ispravno negativnih, lažno pozitivnih i lažno negativnih rezultata za svaku ćeliju.

Definicija 4.2.1 (Ispravno pozitivan)**.** *Ispravno pozitivan (engl. true positive) je rezultat kada model točno predvidi pozitivnu klasu.*

Primjer ispravno pozitivnog rezultata bi bio objekt Home kategorije kojeg je model također svrstao u Home kategoriju. Broj svih objekata koje je model svrstao u kategoriju kojoj doista pripadaju je broj true positive rezultata.

Definicija 4.2.2 (Ispravno negativan)**.** *Ispravno negativan (engl. true negative) je rezultat kada model točno predvidi negativnu klasu.*

Primjer ispravno negativnog rezultata bi bio objekt kategorije koja nije Home kojeg model nije svrstao u Home kategoriju.

Definicija 4.2.3 (Lažno pozitivan)**.** *Lažno pozitivan (engl. false positive) je rezultat kada model netočno predvidi pozitivnu klasu.*

Primjer lažno pozitivnog rezultata bi bio objekt kategorije koja nije Home kojeg je model svrstao u Home kategoriju. Za neku kategoriju, broj svih objekata koje je model svrstao u nju, a ne pripadaju joj je broj lažno pozitivnih rezultata za tu kategoriju. Zbroj lažno pozitivnih rezultata za sve kategorije je broj lažno pozitivnih rezultata za cijelu matricu.

Definicija 4.2.4 (Lažno negativan)**.** *Lažno negativan (engl. false negative) je rezultat kada model netočno predvidi negativnu klasu.*

Primjer lažno negativnog rezultata bi bio objekt Home kategorije kojeg model nije svrstao u Home kategoriju. Za neku kategoriju, broj svih objekata koje model nije svrstao u nju, a ne pripadaju joj je broj lažno negativnih rezultata za tu kategoriju. Zbroj lažno negativnih rezultata za sve kategorije je broj lažno negativnoh rezultata za cijelu matricu.

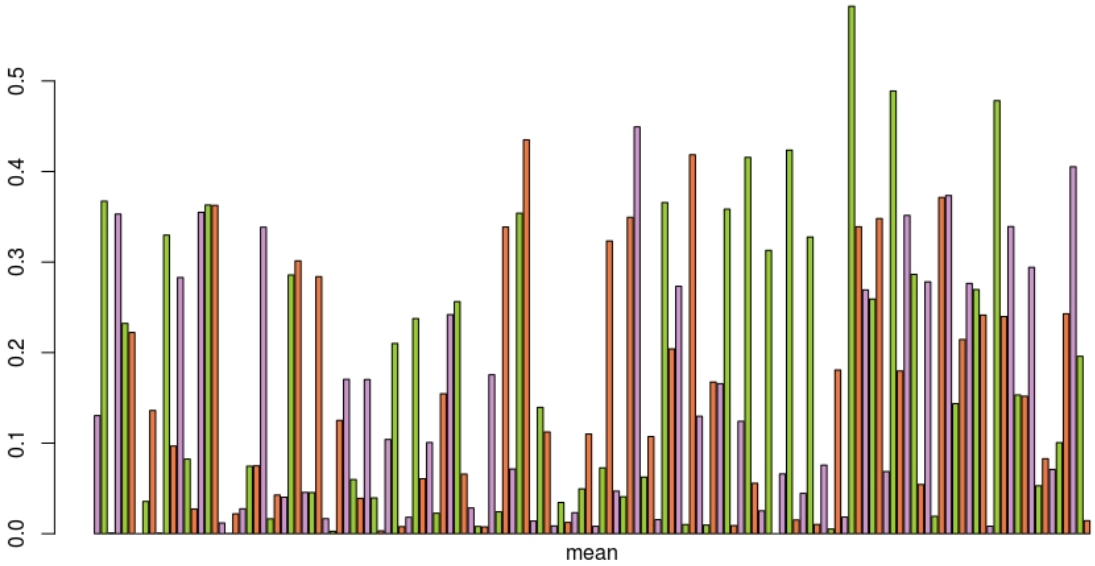
Pomoću brojeva ispravno pozitivnih, ispravno negativnih, lažno pozitivnih i lažno negativnih rezultata moguće je dobiti nekoliko mjere točnosti predviđanja klasifikatora. Oznake TP, TN, FP i FN definiramo na sljedeći način:

TP (ispravno pozitivni)... broj ispravno klasificiranih elemenata testnog skupa s pozitivnim atributom klase

FP (lažno pozitivni)... broj neispravno klasificiranih elemenata testnog skupa s pozitivnim atributom klase

TN (ispravno negativni)... broj ispravno klasificiranih elemenata testnog skupa s negativnim atributom klase

FN (lažno negativni)... broj neispravno klasificiranih elemenata testnog skupa s negativnim atributom klase



Slika 4.3: Stupčasti graf prosječne pogreške po ćelijama

Prva mjera je prosječna pogreška (*engl. mean*). To je mjera za prosječan broj krivo klasificiranih podataka. Na slici 4.3 nalazi se graf stupčaste podjele prosječne greške po ćelijama. Prosjek svih prosječnih grešaka iznosi 0.1595678, a medijan 0.1112110.

Sljedeća mjera je točnost modela (*engl. accuracy*) ili stopa prepoznavanja. Točnost modela je omjer podataka koje je model točno predvidio u cijeloj populaciji (skupu procesiranih podataka). Formalno, točnost se dobiva izrazom

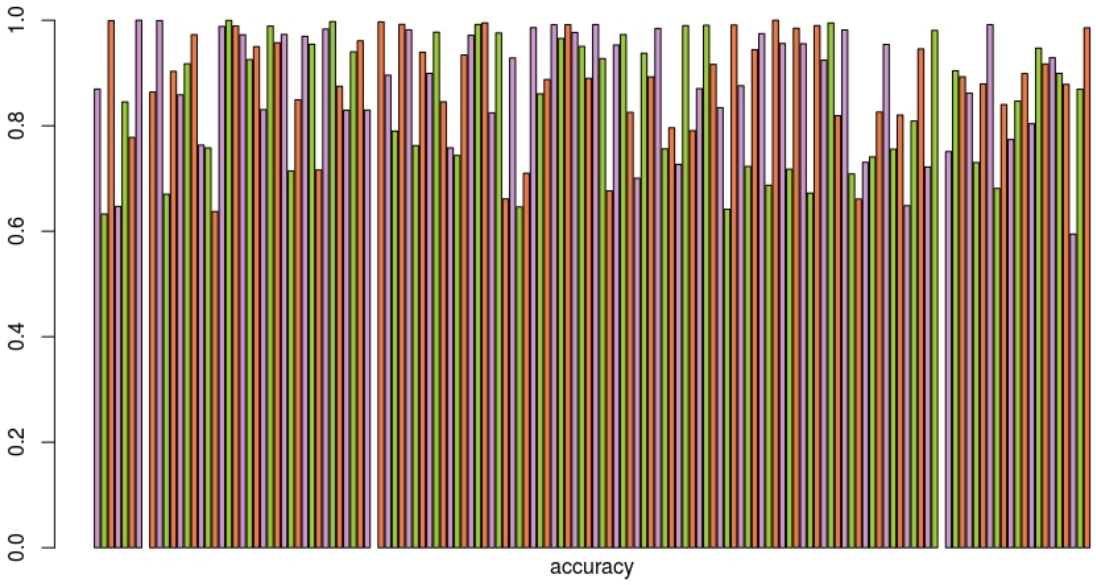
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.12}$$

Na slici 4.4 nalazi se graf stupčaste podjele točnosti po ćelijama. Prosjek svih točnosti iznosi 0.8693479, a medijan 0.8959108.

Kada je odnos kategorija neravnomjeran (u jednoj kategoriji je znatno veći broj objekata nego u ostalima) kao što je to slučaj u eksperimentalnim podacima gdje je često značajno veći broj objekata Home kategorije, tada točnost nije najbolja mjera za model. Stoga se koriste još dvije mjere: preciznost i osjetljivost.

Preciznost (*precision*) je mjera koja iskazuje koliko je pozitivnih procjena doista točno. Preciznost je definirana izrazom (4.13):

$$precision = \frac{TP}{TP + FP} \tag{4.13}$$



Slika 4.4: Stupčasti graf točnosti po ćelijama

Na slici 4.5 nalazi se graf stupčaste podjele preciznosti po ćelijama. Prosjek svih preciznosti iznosi 0.4769839, a medijan 0.5000000.

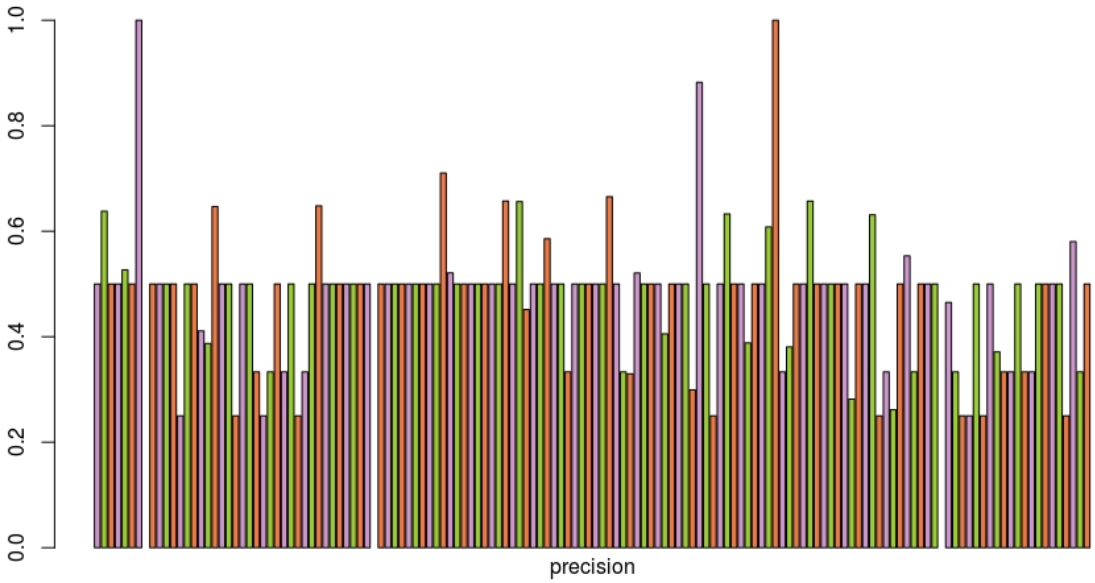
Osjetljivost (*recall*) je mjera koja iskazuje koliko je podataka neke kategorije model točno predvidio. Formula za računanje osjetljivosti je:

$$recall = \frac{TP}{TP + FN} \tag{4.14}$$

Na slici 4.6 nalazi se graf stupčaste podjele osjetljivosti po ćelijama. Prosjek svih osjetljivosti iznosi 0.8364470, a medijan 0.8758542.

Iz ovih rezultata je vidljivo da model multinomne logističke regresije na eksperimentalnim podacima ima malu preciznost, odnosno da u prosjeku pola podataka koje model kategorizira u neku kategoriju ne pripadaju toj kategoriji. S druge strane, model ima dobru osjetljivost, tj. dobro predviđa više od 80% podataka neke kategorije. Budući da problem kojim se bavi ovaj rad ne favorizira nijednu kategoriju niti po njegovoj prirodi ima smisla davati prednost preciznosti nad osjetljivošću ili obrnuto, biti će iskazana još jedna mjera. Radi se o mjeri F1. Mjera F1 koristi se kada se traži ravnoteža između osjetljivosti i preciznosti, a distribucija klasa je jako neravnomjerna. To je harmonijska sredina preciznosti i osjetljivosti, iskazana izrazom:

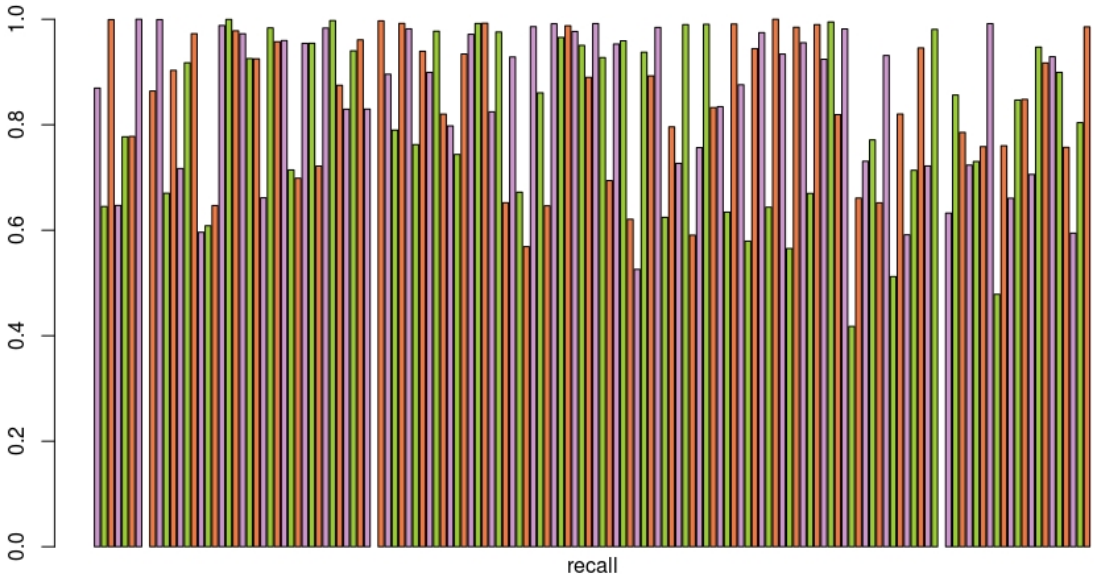
$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{4.15}$$



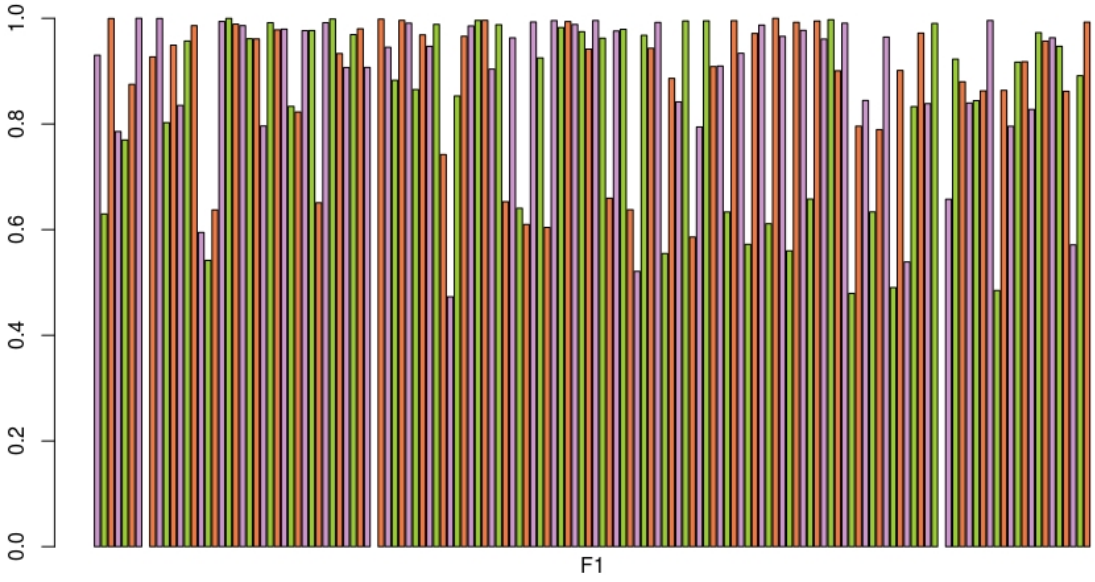
Slika 4.5: Stupčasti graf preciznosti po ćelijama

Na slici 4.7 nalazi se graf stupčaste podjele mjere F1 po ćelijama. Prosjek svih mjera F1 iznosi 0.8680865, a medijan 0.9338191.
Konačno, na slici 4.8 mogu se vidjeti boxplotovi svih spomenutih mjera i njihova usporedba.

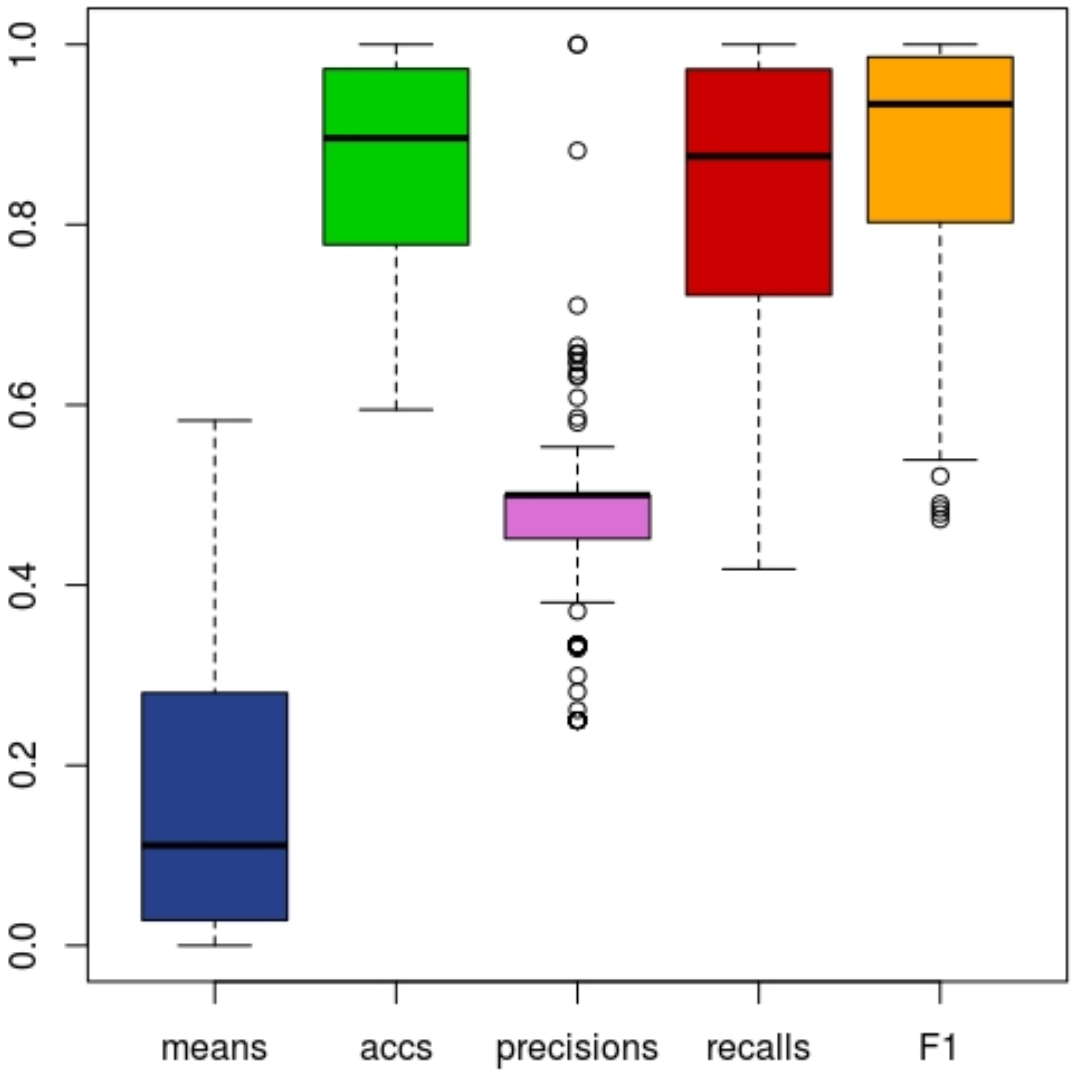
Valja napomenuti da se prilikom računanja TP, TN, FP i FN vrijednosti računaju za svaku kategoriju unutar ćelije posebno. Prilikom računanja statističkih pokazatelja kvalitete točnosti, također se računa statistički pokazatelj kvalitete točnosti za svaku kategoriju posebno i prosjek tih pokazatelja se uzima kao vrijednost statističkog pokazatelja za ćeliju. Posljedica toga je da se ni prosjek ni medijan F1 mjere po svim ćelijama ne može dobiti direktnim uvrštavanjem prosjeka preciznosti i osjetljivosti u 4.15.



Slika 4.6: Stupčasti graf osjetljivosti po ćelijama



Slika 4.7: Stupčasti graf mjere F1 po ćelijama



Slika 4.8: Boxplot prikaz svih statističkih pokazatelja kvalitete točnosti

Poglavlje 5

Zaključak

Migracije ljudi važan su podatak koji se koristi pri urbanom planiranju u prometne, ekonomske i druge svrhe. Tradicionalno se migracije bilježe polazišno-odredišnim matricama koje daju prostornu i vremensku komponentu urbanog kretanja. Međutim, svrha kretanja ljudi i razlozi putovanja nisu zabilježeni ovakvim pristupom. Ovdje je prikazana nova formalna (poopćena) metoda kategorizacije prostornih objekata i predviđanja vjerojatnosti posjeta kategorijama tih objekata, u svrhu omogućavanja točnijeg određivanja prirode urbanih migracija. Korišteni su podaci iz baze prostornih podataka OpenStreetMap o namjeni objekata na području grada Shenzhena u Kini te GPS podaci putanja taksija u istom geografskom području u izabranom periodu vremena. Prostorni objekti su kategorizirani pomoću žigova (atributa prostornih podataka koji opisuju objekte) unutar OSM-a koji opisuju namjenu objekata. Navedena kategorizacija je zatim proširena na taksi podatke, te je za svaku posjetu taksija unutar zadanog područja (Shenzhen, Kina) dobivena kategorija te posjete. Skup podataka o kategoriji taksi objekata podijeljen je na podskup za učenje i podskup za provjeru modela u odnosu 80% : 20%, te je zasnovano na podskupu za učenje razvijen model mulitnomne logističke regresije koji na temelju vremenskog okvira u kojem se posjeta dogodila predviđa kategoriju posjete. Kvaliteta modela provjerena je korištenjem podskupa izvornih podataka za provjeru i kvaliteta modela izražena je statističkim pokazateljima kvalitete točnosti, preciznosti, osjetljivosti i F1 mjerom.

Bibliografija

[1] L. Alexander, S. Jiang, M. Murga i M.C. González, *Origin–destination trips by purpose and time of day inferred from mobile phone data*, Transportation Research Part C: Emerging Technologies **58** (2015), 240 – 250, ISSN 0968-090X, http://www.sciencedirect.com/science/article/pii/S0968090X1500073X, Big Data in Transportation and Traffic Engineering.

[2] M. Alhazzani, F. Alhasoun, Z. Alawwad i M.C. González, *Urban Attractors: Discovering Patterns in Regions of Attraction in Cities*, CoRR **abs/1701.08696** (2016).

[3] C. Antoniou, M. Ben-Akiva i H. Koutsopoulos, *Incorporating Automated Vehicle Identification Data into Origin-Destination Estimation*, Transportation Research Record: Journal of the Transportation Research Board **1882** (2004), 37–44, https://doi.org/10.3141/1882-05.

[4] M. Ben-Akiva, *Methods to Combine Different Data Sources and Estimate Origin-Destination Matrices*, Transportation and Traffic Theory: Proceedings of the 10th International Syposium on Transportation and Traffic Theory (M. Gartner i N. H. M. Wilson, ur.), Elsevier, New York, 1987, str. 459–481.

[5] S. Bera i K.V.K. Rao, *Estimation of origin-destination matrix from traffic counts: The state of the art*, European Transport Trasporti Europei **49** (2011), 2–23.

[6] A. Cui, *Bus passenger Origin-Destination Matrix estimation using Automated Data Collection systems*, (2007).

[7] M. G. Demissie, F. Antunes, C. Bento, S. Phithakkitnukoon i T. Sukhvibul, *Inferring origin-destination flows using mobile phone data: A case study of Senegal*, 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), June 2016, str. 1–6.

[8] S. Dešić, M. Filić i R. Filjar, *Determination of origins and destinations for an O-D matrix based on telecommunication activity records*, 2017 40th International Conven-

tion on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2017, str. 424–427.

[9] R. Filjar, M. Filić, A. Lučić, K. Vidović i D. Šarić, *Anatomy of Origin-Destination Matrix derived from GNSS alternatives*, Coordinates (2016), https://mycoordinates.org/anatomy-of-origin-destination-matrix-derived-from-gnss-alternatives/, posjećena 2019-01-06.

[10] M.A. Florez, S. Jiang, R. Li, C.H. Mojica, R.A. Rios i M.C. González, *Measuring the impact of economic well being in commuting networks – A case study of Bogota, Colombia*, Transportation Research Board 96th Annual Meeting (2017), https://trid.trb.org/view/1438491.

[11] OpenStreetMap Foundation, *OpenStreetMap*, https://www.openstreetmap.org, posjećena 2018-12-16.

[12] R Foundation, *R project*, https://www.r-project.org/, posjećena 2018-10-08.

[13] M. Gaudry, *The Four Approaches to Origin-Destination Matrix Estimation for Consideration by the MYSTIC Research Consortium*, Bureau d'Economie Théorique et Appliquée, UDS, Strasbourg, Working Papers of BETA (2000).

[14] T. He, *Data Description for UrbanCPS*, https://www-users.cs.umn.edu/~tianhe/BIGDATA/, posjećena 2019-01-15.

[15] Md.S. Iqbal, C.F. Choudhury, P. Wang i M.C. González, *Development of origin–destination matrices using mobile phone call data*, Transportation Research Part C: Emerging Technologies **40** (2014), 63 – 74, ISSN 0968-090X, http://www.sciencedirect.com/science/article/pii/S0968090X14000059.

[16] G. James, D. Witten, T. Hastie i R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 6., Springer, 2013.

[17] J. Ma, H. Li, F. Yuan i T. Bauer, *Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data*, International Journal of Transportation Science and Technology **2** (2013), br. 3, 183 – 204, ISSN 2046-0430, http://www.sciencedirect.com/science/article/pii/S204604301630140X.

[18] A. Mimica i M. Ninčević, *Statistika primjeri i zadaci*, https://web.math.pmf.unizg.hr/nastava/stat/files/vjezbe_novo.pdf, posjećena 2019-01-15, kolovoz 2010.

[19] P.E. Murphy i C.P. Keller, *Destination travel patterns: An examination and modeling of tourist patterns on Vancouver Island, British Columbia*, Leisure Sciences **12** (1990), br. 1, 49–65, https://doi.org/10.1080/01490409009513089.

[20] OpenCellid, *Open Database of Cell Towers*, https://opencellid.org/, posjećena 2019-01-26.

[21] C. Pan, J. Lu, S. Di i B. Ran, *Cellular-Based Data-Extracting Method for Trip Distribution*, Transportation Research Record: Journal of the Transportation Research Board **1945** (2006), 33–39, https://doi.org/10.3141/1945-04.

[22] RStudio, *RStudio*, https://www.rstudio.com/, posjećena 2018-10-08.

[23] C. Song, Z. Qu, N. Blumm i A.L. Barabási, *Limits of Predictability in Human Mobility*, Science **327** (2010), br. 5968, 1018–1021, ISSN 0036-8075, http://science.sciencemag.org/content/327/5968/1018.

[24] I. Stupar, P. Martinjak, V. Turk i R. Filjar, *Socio-economic origin-destination matrix derivation through contextualization of material world*, 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2018, str. 0417–0421.

[25] M. Taboga, *StatLect*, https://www.statlect.com/fundamentals-of-statistics/maximum-likelihood-algorithm, posjećena 2019-01-02.

[26] R. Tolouei, S. Psarras i R. Prince, *Origin-Destination Trip Matrix Development: Conventional Methods versus Mobile Phone Data*, Transportation Research Procedia **26** (2017), 39 – 52, ISSN 2352-1465, http://www.sciencedirect.com/science/article/pii/S2352146517308669, Emerging technologies and models for transport and mobility.

[27] P. Wang, T. Hunter, A. Bayen, K. Schechtner i M.C. Gonzalez, *Understanding Road Usage Patterns in Urban Areas*, Scientific reports **2** (2012), 1001.

[28] K.I. Wong i S.A. Yu, *Estimation of origin–destination matrices for mass event: A case of Macau Grand Prix*, Journal of King Saud University - Science **23** (2011), br. 3, 281 – 292, ISSN 1018-3647, http://www.sciencedirect.com/science/article/pii/S1018364710001631, Special Issue on "Advances in Transportation Science".

[29] H. Yang i J. Zhou, *Optimal Traffic Counting Locations for Origin-Destination Matrix Estimation*, Transportation Research Part B: Methodological **33** (1998), 109–126.

Sažetak

Urbane migracije važan su podatak koji se koristi pri urbanom planiranju u prometne, ekonomske i druge svrhe. Migracije se bilježe polazišno-odredišnim matricama koje daju prostornu i vremensku komponentu urbanog kretanja. Tradicionalno, polazišno-odredišne matrice dobivaju se brojanjem ljudi i vozila na frekventnim područjima, a u novije vrijeme za njihovu se izradu koriste metode koje koriste podatke o korištenju javnih pokretnih (telekomunikacijskih) mreža za procjenu migracija. Takvim pristupom nije zabilježena svrha kretanja ljudi i razlozi putovanja. U ovom istraživanju je prikazana nova formalna (poopćena) metoda kategorizacije prostornih objekata i predviđanja vjerojatnosti posjeta kategorijama tih objekata, u svrhu omogućavanja točnijeg određivanja prirode urbanih migracija.

Uvodno poglavlje formulira ciljeve istraživanja i kratak pregled strukture rada. Prvo poglavlje detaljno opisuje domenu problema i motivaciju za istraživanje. U drugom poglavlju opisani su podaci koji se koriste te njihovo pribavljanje i obrada. U trećem poglavlju razrađena je kategorizacija podataka. U četvrtom poglavlju je opisana izrada regresijskog modela i primjena na eksperimentalnim podacima. Zaključno poglavlje daje zaključak istraživanja.

Korišteni su podaci iz baze prostornih podataka OpenStreetMap o namjeni objekata na području grada Shenzhena u Kini te GPS podaci putanja taksija u istom geografskom području. Prostorni objekti su kategorizirani pomoću žigova (atributa prostornih podataka koji opisuju objekte) unutar OpenStreetMapa koji opisuju namjenu objekata. Navedena kategorizacija je zatim proširena na taksi podatke.

Summary

Urban migration is an important information used in urban planning for traffic, economic and other purposes. Migrations are recorded with origin-destination matrices which give the spatial and temporal components of urban motion. Traditionally, origin-destination matrices are obtained by counting people and vehicles in frequency domains, and in recent times methods used for their development are using data on the use of public mobile (telecommunication) networks for estimating migration. That approach does not report purpose of people's movements and the reasons for traveling. In this research, a new formal (generalized) method of categorizing spatial objects and predicting probabilities of visits to the categories of these objects is introduced, with intention of enabling the more accurate understanding of the nature of urban migration.

The introductory chapter formulates research objectives and a brief overview of the work structure. First chapter details the problem domain and motivation for research. The second chapter describes the data used and their acquisition and processing. In the third chapter the categorization of data was elaborated. The fourth chapter describes the creation of regression model and application on experimental data. The concluding chapter gives the conclusion of the research.

Data from the OpenStreetMap spatial database on the use of objects in the Shenzhen city area in China and the GPS data of the taxiway in the same geographic area are used. Spatial objects are categorized by means of the spatial attributes describing objects within OpenStreetMap that describe the purpose of the objects. The categorization is then extended to the taxi data.

Životopis

Rođena sam 1994. godine u Zagrebu. Tamo završavam srednju školu i 2013. godine upisujem studij Matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Po završetku preddiplomskog studija, 2016. godine nastavljam obrazovanje na diplomskom studiju Računarstva i matematike. Tokom diplomskog studija volonterski sudjelujem u radu udruge Programerko osmišljajući i provodeći radionice programiranja za djecu u sklopu CodeClub programa.

2017. godine sudjelujem na ljetnom kampu Ericsson Nikola Tesla. Rad na projektu rezultirao je objavom i predstavljanjem zajedničkog rada na MIPRO (međunarodni skup za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku) konferenciji. Od ožujka 2018. zaposlena sam u ReversingLabsu kao Data Analyst. U slobodno vrijeme bavim se volonterskim radom na području dobrobiti životinja.