INFORMATIKA

Toplinska karta u ilustraciji podatka

Tvrtko Tadić[1]

U ovom članku opisat ćemo metodu ilustracije podataka koristeći toplinsku kartu (engl. *heat map*). Toplinska karta koristi se za prikaz gustoće određenih pojava, poput gustoće stanovništva, poslovnih subjekata ili drugih **točaka interesa**[2]. Kao primjer ćemo analizirati javno dostupne podatke o kriminalu na području grada Seattlea te odgovoriti na pitanje: u kojim dijelovima grada ima najmanje provala? Ovo je pitanje koje je osobito zanimalo autorove prijatelje. Za modeliranje gustoće podataka koristit ćemo **Gaussovsku jezgru**, koja se primjenjuje za modeliranje širenja topline i drugih sličnih fenomena.

Podatci o kriminalu u Seattleu. Točke interesa

Novi roditelji u Seattleu žele se preseliti u mirniji dio grada te ih zanima u kojem dijelu ima najmanje kriminala. Podatci o svim kriminalnim događajima u gradu Seattleu dostupni su na službenim web-stranicama grada. Skup podataka [2] sadrži informacije o više od 1.15 milijuna različitih događaja, uključujući prometne nesreće, krađe i ubojstva, prikupljenih od 2008. godine. Podatci su organizirani u 17 stupaca, od kojih u nastavku navodimo najvažnije:

naziv	**opis**	**ime stupca**	**tip podatka**
broj izvještaja	broj pod kojim se vodi spis o događaju	`report_number`	tekst
broj prekršaja	više prekršaja može biti sadržano u istom izvještaju	`offense_idr`	tekst
vrijeme početka	datum i vrijeme početka prekršaja	`offense_start_datetime`	vremenska oznaka
prekršaj	ime / opis prekršaja	`offense`	tekst
mikrolokacija	ime mikro područja događaja	`mcpp`	text
dužina*	koordinata zemljopisne dužine događaja	`longitude`	broj
širina*	koordinata zemljopisne širine događaja	`latitude`	broj

* – podatci o koordinatama su zaokruženi na najbliži blok zgrada radi privatnosti, a u nekim osjetljivim slučajevima su izostavljeni

[1] Autor radi u Microsoft Corporation, Redmond, SAD; e-pošta: `tvrtko.tadic@math.hr`

[2] Engleski *points of interest*, često skraćeno u (stranoj) stručnoj literaturi kao *POI*.

Ovi podatci omogućuju detaljnu analizu kriminala u različitim dijelovima grada, što može pomoći u donošenju informirane odluke o preseljenju. Podatke koje smo preuzeli u *csv* formatu[3], učitat ćemo koristeći popularnu biblioteku Pandas u Pythonu:

```
import pandas as pd
prekrsaji = pd.read_csv("SPD_Crime_Data__2008-Present_20240921.csv")
```

Bacimo prvi pogled na podatke:

```
prekrsaji[['Offense Start DateTime', 'Offense', 'Longitude','Latitude']
].head(5)
```

```
Offense Start DateTime  Offense                        Longitude   Latitude
0 02/05/2020 10:10:00 AM Drug/Narcotic Violations    -122.385974  47.649387
1 02/03/2020 08:00:00 AM Theft of Vehicle Parts      -122.323399  47.675118
2 02/02/2020 08:30:00 PM Robbery                     -122.299552  47.666384
3 02/05/2020 01:17:00 AM Destruction of Property     -122.384865  47.642927
4 02/05/2020 12:51:21 AM Driving Under the Influence -122.366195  47.662193
```

Kako točaka ima dosta, uvijek je dobro pogledati neke jednostavne statistike o brojčanim podatcima.

```
prekrsaji[['Longitude','Latitude']].describe()
```

```
       Longitude     Latitude
count  1.144421e+06  1.144421e+06
mean  -1.168981e+02  4.551483e+01
std    2.521980e+01  9.809739e+00
min   -1.659980e+02  0.000000e+00
25%   -1.223475e+02  4.758025e+01
50%   -1.223288e+02  4.761420e+01
75%   -1.223091e+02  4.766332e+01
max    1.718570e+02  8.999999e+01
```

Uočimo da raspolažemo s 1.14 milijuna točaka. Iz prethodnog sažetka također možemo primijetiti da postoji veliki raspon vrijednosti za geografsku dužinu i širinu. Vrijednosti geografske dužine nalaze se u intervalu $[-166, 1\,144\,421]$, dok su vrijednosti geografske širine u intervalu $[0, 1\,144\,421]$.

Kod ovako velikog broja podataka često se javljaju **greške**. Ponekad se koordinate unesu pogrešno, dok se u drugim slučajevima, ako su nepoznate, unosi vrijednost 0.

Promatranjem bilo koje karte lako je uočiti da se Seattle nalazi unutar pravokutnika definiranog sa $[47.45, 47.75] \times [-122.45, -122.20]$. Stoga ćemo iz skupa podataka ukloniti sve točke čije koordinate ne pripadaju ovom pravokutniku.

```
prekrsaji = prekrsaji[(prekrsaji['Latitude'] > 47.45)
& (prekrsaji['Latitude'] < 47.75) & (prekrsaji['Longitude'] < -122.20)
& (prekrsaji['Longitude'] > -122.45)]
```

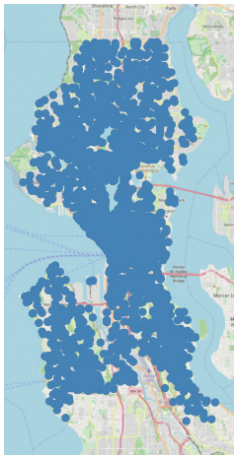
[3] *csv* je skraćenica na engleskom jeziku za *comma-separated values*. Ovo je jedan od standardnih formata koji se koristi za spremanje podataka. U takvim datotekama podatci su razdvojeni *separatorom*, koji je najčešće zarez, ali može biti i neki drugi znak.

Sada se broj podataka smanjio na 1.09 milijuna. Iako je to manje nego prije, još uvijek se radi o velikom broju podataka. Broj točaka koje trebamo prikazati na karti je prevelik – kada bismo ih sve nacrtali, karta bi bila toliko ispunjena da se ništa ne bi moglo razaznati.

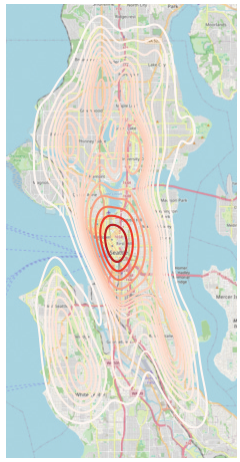
Stoga ćemo najprije uzeti slučajni uzorak od 4000 točaka i nacrtati ih na karti. Za ovaj zadatak koristit ćemo biblioteku Geopandas.

```
import geopandas as gpd
from shapely.geometry import Point

# kreiranje tocke (Point)
prekrsaji['geometry'] = prekrsaji.apply(lambda row:
Point(row['Longitude'], row['Latitude']), axis=1)

#pretvaranje podataka u geopandas okvir podataka
geo_prekrsaji = gpd.GeoDataFrame(prekrsaji, crs="EPSG:4326")

#crtanje karte
import geoplot as gplt
import geoplot.crs as gcrs
import matplotlib.pyplot as plt

geo_prekrsaji_uzorak =  geo_prekrsaji.sample(4000)
ax = gplt.webmap(geo_prekrsaji_uzorak, projection=gcrs.WebMercator())
gplt.pointplot(geo_prekrsaji_uzorak, ax=ax)
```

Rezultat je prikazan na slici 1. Na ovoj slici prikazano je svega 4000 točaka, no karta je već gotovo potpuno popunjena. Mnoge točke se međusobno preklapaju, pa bi bilo korisno na neki način izraziti *gustoću* tih točaka. U tome će nam pomoći toplinska karta.



Slika 1. Točke prikazuju lokacije prekršaja u Seattleu



Slika 2. Vidimo procjenu gustoće prekšaja u Seattleu

Većina softverskih alata ima ugrađenu metodu za ovakve analize, poznatu kao metoda **procjene gustoće jezgrom**[4]. Rezultat ove metode vidljiv je na slici 2. Kako bismo dobili ovu sliku, izvršili smo sljedeće linije koda:

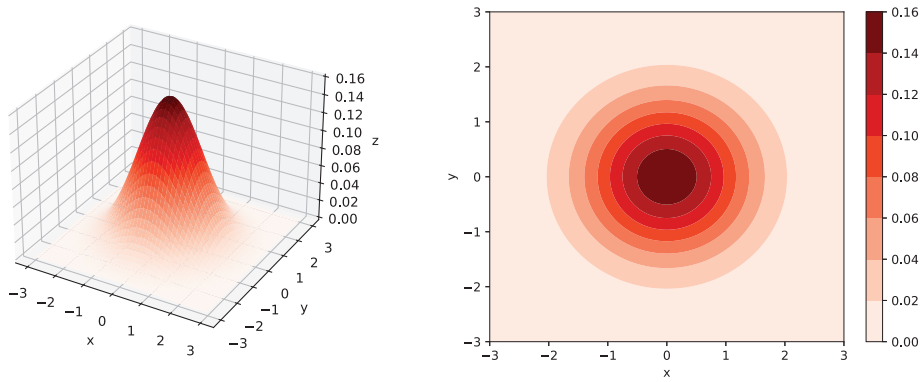
```
ax = gplt.webmap(geo_prekrsaji_uzorak, projection=gcrs.WebMercator())
gplt.kdeplot(geo_prekrsaji_uzorak, n_levels=15, cmap='Reds', ax =ax)
```

Metoda procjene gustoće Gaussovskom jezgrom

U ovom ćemo odjeljku, koristeći pojednostavljeni model, objasniti kako se dobiva procjena gustoće koristeći Gaussovsku jezgru. Općenitiji pristup i dodatne detalje čitatelj može pronaći u članku na Wikipediji posvećenom ovoj temi [7].

Procjena gustoće jezgrom (KDE) je statistička metoda koja se koristi za procjenu gustoće nekog skupa podataka. Kao što smo pokazali na primjeru prekršaja u Seattleu, ideja je da, ako imamo skup točaka u ravnini, želimo prikazati gdje se nalazi najviše točaka, odnosno područja najveće *gustoće*.

KDE nam pomaže vizualno predstaviti ovu gustoću stvarajući glatku funkciju koja ističe središta tih područja. Karta prikazana na slici 2 primjer je dvodimenzionalne projekcije takve procijenjene gustoće. Pojedinačan doprinos jedne točke ilustriran je na slici 3.



Slika 3. Trodimenzionalna slike funkcije $K(x,y) = \frac{1}{2\pi}\exp\left(\frac{-x^2-y^2}{2}\right)$ *i projekcija na dvije dimenzije*

Funkcija

$$K(x,y) = \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)$$

naziva se **Gaussovska jezgra**. Ova funkcija koristi se za modeliranje širenja topline, o čemu zainteresirani čitatelj može pronaći više informacija u [6][5].

[4] Engl. *kernel density estimate* – KDE.

[5] Vidi odjeljak o fundamentalnom rješenju.

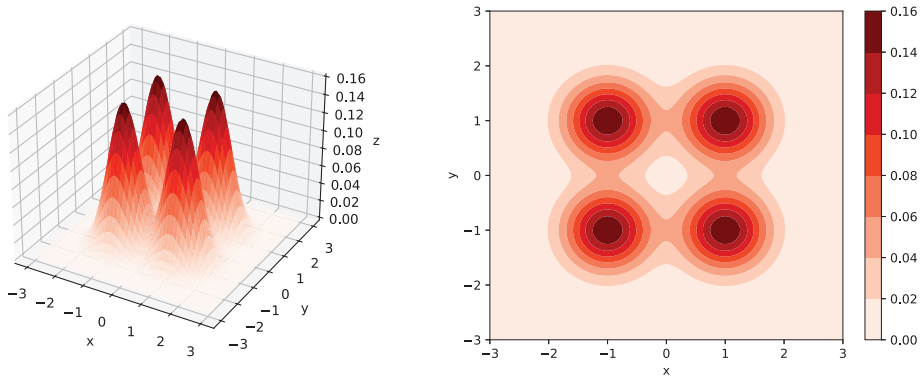
Počnimo od jedne točke u ravnini (x_j, y_j). Kao što smo naveli u podatcima o kriminalu u Seattleu, lokacije prekršaja nije točna nego zaokružena zbog privatnosti. Stoga *vjerodostojnost* lokacije obično prikazujemo pomoću funkcije

$$K_j^h(x,y) = \frac{1}{h_x h_y} K\left(\frac{x-x_1}{h_x}, \frac{y-y_1}{h_y}\right)$$
$$= \frac{1}{2\pi h_x h_y} \exp\left(-\frac{(x-x_j)^2}{2h_x^2} - \frac{(y-y_j)^2}{2h_y^2}\right),$$

gdje su $h_x > 0$ i $h_y > 0$ parametri glatkoće koji se zadaju. Funkcija $K_j^h(x,y)$ ima najveću vrijednost kada je $(x,y) = (x_j, y_j)$ i što je (x,y) dalje od (x_j, y_j) vrijednost se brzo smanjuje (kao što je ilustrirano na slici 3).

Uzimanjem prosjeka funkcija K_j^h za $j = 1, \ldots, n$, dobivamo **procjenu gustoće** $\hat{g}$:

$$\hat{g}(x,y) := \frac{K_1^h(x,y) + K_2^h(x,y) + \ldots + K_n^h(x,y)}{n}.$$



Slika 4. Procjena gustoće Gaussovom jezgrom za točke $(-1,-1)$, $(1,-1)$, $(1,-1)$ *i* $(1,1)$ *za* $\sigma_x = \sigma_y = 1/2$

Ostaje jedno važno pitanje: koju vrijednost odabrati za h_x i h_y? Postoje složene metode koje se mogu koristiti za određivanje ovih parametara. Međutim, u praksi se pokazalo da možemo postaviti:

$$h_x = \frac{1}{n^{1/6}}\sigma_x \quad \text{i} \quad h_y = \frac{1}{n^{1/6}}\sigma_y,$$

gdje su

$$\sigma_x = \sqrt{\frac{(x-x_1)^2 + \ldots + (x-x_n)^2}{n-1}} \quad \text{i} \quad \sigma_y = \sqrt{\frac{(y-y_1)^2 + \ldots + (y-y_n)^2}{n-1}}$$

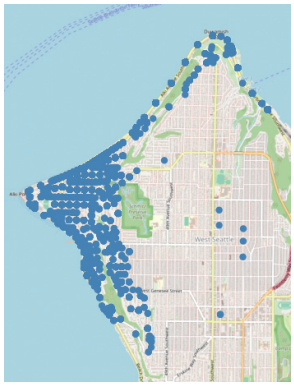
standardne devijacije vrijednosti x i y koordinata.

Za više informacija o standardnoj devijaciji, čitatelja upućujemo na članak [5] i knjigu [3].

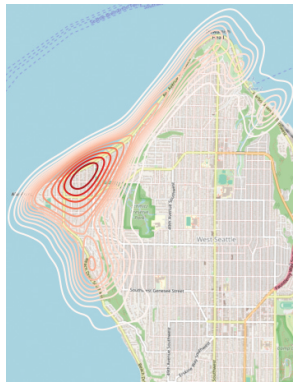
Vratimo se na primjer podataka s kojim smo započeli. Taj skup podataka sadrži informacije o dijelovima grada u kojima su se dogodili prekršaji. Međutim, dijelovi grada često su nejasno definirani i mogu se međusobno preklapati. Korištenjem ove metode možemo približno odrediti područje grada na koje se podatci odnose te koristiti ovu informaciju kao podršku pretragama.

Alki je dio grada Seattlea koji se nalazi uz more. Podatci o mikrolokacijama u skupu podataka o kriminalu u Seattleu daju popis lokacija koje se tvrdi da pripadaju tom dijelu grada.

Na slici 5 prikazane su sve lokacije za koje se tvrdi da su u dijelu grada koji se naziva Alki, zajedno s toplinskom kartom koju su te lokacije generirale. Možemo primijetiti da se većina točaka nalazi u blizini drugih, čineći **nakupinu**[6], dok su neke točke *izolirane*.

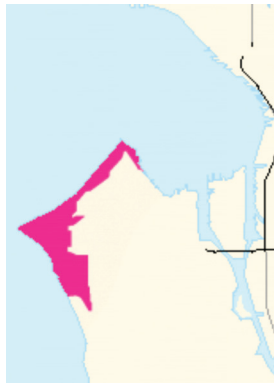


Slika 5. Točke označene s Alki



Slika 6. Toplinska karta točki

Slika 7. Primjer pocjene područja Alki prema procijenjenoj gustoći



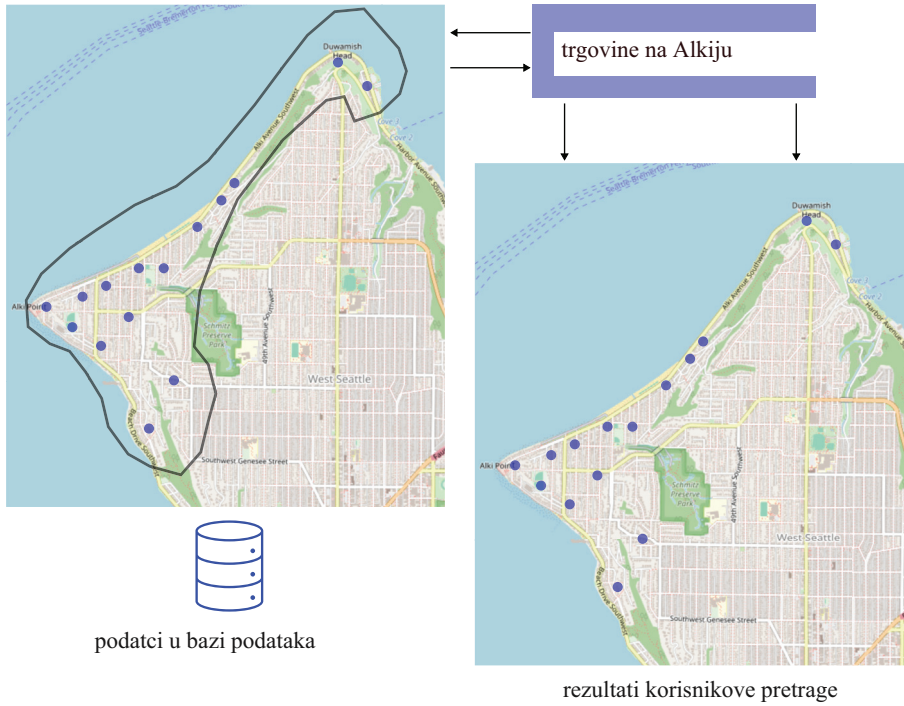
Slika 8. Područje Alki prema Wikipediji

[6] U literaturi na engleskom koristi se izraz *cluster*

Izolirane točke predstavljaju pogrešno prijavljene podatke, koje obično nazivamo **šum**[7].

Procijenjena gustoća pruža dobar indikator koji dijelovi pripadaju Alkiju, a koji ne. Područje s većim vrijednostima predstavlja dio Alkija, dok područje s manjim vrijednostima nije. Ne ulazeći u detalje, ovo nam može poslužiti za procjenu poligona koji obuhvaća to područje, o čemu čitatelj može saznati više u [4].

Ovako dobiveni mnogokut ne mora nužno biti prikazan, već se može koristiti u pozadini za pretrage, kao što je ilustrirano na slici 9.



Slika 9. Primjer mogućeg rješenja korištenja procijenjenog mnogokuta

Završni osvrt

Kao i mnoge druge metode, toplinske karte su moćni alati za analizu podataka koji omogućuju bolje razumijevanje i vizualizaciju kompleksnih skupova podataka, pružajući uvide koji bi inače ostali skriveni u sirovim brojkama.

Primjena toplinskih karata je široka, od ekologije, gdje se koriste za procjenu staništa životinja, do urbanog planiranja, gdje mogu pomoći u identifikaciji područja s visokom frekvencijom prometnih nesreća ili kriminalnih aktivnosti.

[7] Engl. *noise*

Ovdje su brojni detalji izostavljeni. Metoda procjene gustoće jezgrom dobro je razrađena i matematički dokazana metoda. Ovdje je korištena Gaussovska jezgra jer se pokazala kao uspješna metoda za vizualni prikaz podataka. Ipak, postoji mnogo drugih jezgara koje se mogu koristiti.

Nismo govorili puno o tome kako se dobiva procijenjeni mnogokut jer za to treba uvesti algoritme grupiranja[8] koji bi utvrdili koji dio pripada području koje želimo procijeniti, a koji ne. Za to bi nam trebao poseban članak.

Ovdje smo koristili Python prvenstveno jer se koristi u srednjim i osnovnim školama u nastavi informatike. Python bilježnica s primjerima korištenim u ovom članku dostupna je na adresi `web.math.hr/~tvrtko/mfl/toplinske_karte`. Programski jezik R za statistička računanja i vizualizaciju podataka ima bolje mogućnosti nego Python za karte i vizualizaciju procjene Gaussovskim jezgrama.

Podatci o kartama dolaze iz projekta OpenStreetMap [1]. To je projekt koji ima za cilj stvaranje slobodne i otvorene geografske baze podataka svijeta. Projekt je započeo 2004. godine, a korisnici diljem svijeta doprinose dodavanjem i uređivanjem podataka o cestama, stazama, zgradama i drugim geografskim značajkama.

Literatura

[1] *Openstreetmap*, `https://www.openstreetmap.org/`, Pristupljeno: 30. 11. 2024.

[2] *Seattle police department crime data: 2008-present*, `https://data.seattle.gov/Public-Safety/SPD-Crime-Data-2008-Present/tazs-3rd5/about_data`, Pristupljeno: 30. 11. 2024.

[3] Nikola Sarapa, *Vjerojatnost i statistika II. dio: Osnove statistike – slučajne varijable*, Školska knjiga, 1996.

[4] Beril Sirmacek and Cem Unsalan, *Urban area detection using local feature points and spatial voting*, IEEE Geoscience and Remote Sensing Letters, **7** (1): 146–150, 2010.

[5] Tvrtko Tadić, *Aritmetička sredina i standardna devijacija*, Poučak: časopis za metodiku i nastavu matematike, **18** (69): 4–18, 2017.

[6] *Wikipedia contributors, Heat equation, Wikipedia, the free encyclopedia*, 2024, [Online; accessed 10-November-2024].

[7] *Wikipedia contributors, Multivariate kernel density estimation, Wikipedia, the free encyclopedia*, 2024, [Online; accessed 10-November-2024].

[8] Engl. *clustering algortihms*.