UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

ZAGOTAVLJANJE KAKOVOSTI VIZUALIZACIJ PODATKOV Z ZEMLJEVIDI

Ljubljana, julij 2023 MIHA REJEC

IZJAVA O AVTORSTVU

Podpisani Miha Rejec, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Zagotavljanje kakovosti vizualizacij podatkov z zemljevidi, pripravljenega v sodelovanju s svetovalcem red. prof. dr. Jurijem Jakličem

I Z J A V L J A M

1. da sem predloženo delo pripravil samostojno;

2. da je tiskana oblika predloženega dela istovetna njegovi elektronski obliki;

3. da je besedilo predloženega dela jezikovno korektno in tehnično pripravljeno v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani, kar pomeni, da sem poskrbel, da so dela in mnenja drugih avtorjev oziroma avtoric, ki jih uporabljam oziroma navajam v besedilu, citirana oziroma povzeta v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani;

4. da se zavedam, da je plagiatorstvo – predstavljanje tujih del (v pisni ali grafični obliki) kot mojih lastnih – kaznivo po Kazenskem zakoniku Republike Slovenije;

5. da se zavedam posledic, ki bi jih na osnovi predloženega dela dokazano plagiatorstvo lahko predstavljalo za moj status na Ekonomski fakulteti Univerze v Ljubljani v skladu z relevantnim pravilnikom;

6. da sem pridobil vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označil;

7. da sem pri pripravi predloženega dela ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil soglasje etične komisije;

8. da soglašam, da se elektronska oblika predloženega dela uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;

9. da na Univerzo v Ljubljani neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve predloženega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja predloženega dela na voljo javnosti na svetovnem spletu preko Repozitorija Univerze v Ljubljani;

10. da hkrati z objavo predloženega dela dovoljujem objavo svojih osebnih podatkov, ki so navedeni v njem in v tej izjavi.

11. da sem preveril verodostojnost informacij, ki izhajajo iz zapisov na podlagi uporabe orodij umetne inteligence.

V Ljubljani, dne _____ Podpis študenta: _____

KAZALO

KAZALO TABEL

KAZALO SLIK

SEZNAM KRATIC

angl. – angleško

CSV – (angl. comma-separated values); vrednosti, ločene z vejico

DMS – (angl. Degrees-Minutes-Seconds); stopinje-minute-sekunde

EPSG – (angl. European Petroleum Survey Group); evropska raziskovalna skupina za nafto

ETL – (angl. Extract-Transform-Load); pridobivanje-preoblikovanje-nalaganje

ISO – (angl. International Organization for Standardization); mednarodna organizacija za standardizacijo

UVOD

Na svetu količina podatkov iz dneva v dan narašča. Podatkov je ogromno, ti pa sami po sebi niso nič vredni, če jih ne razumemo oz. iz njih ne razberemo uporabnih informacij. Zaradi nezmožnosti človekovega razumevanja takšne ogromne količine podatkov so uporabne metode in tehnologije, ki človeku omogočajo, da si te podatke lahko interpretira. Interpretacijo podatkov izvedemo z uporabo vizualizacij. Z vizualizacijo podatkov prikažemo informacije, ki so človeku razumljive in mu omogočajo lažje sprejemanje odločitev (Sinar, 2015). Pri vizualizaciji podatkov poznamo različne vrste grafov, tabel in zemljevidov. Za vizualizacijo prostorskih podatkov (angl. geospatial data) se uporabljajo zemljevidi. Vizualizacije z zemljevidi prikazujemo v geografskih informacijskih sistemih (angl. geographic information systems), ki poleg vizualizacije podatkov omogočajo tudi njihovo pridobivanje, shranjevanje in analizo (Zhu, Zhao, Liang & Qin, 2021).

Vizualizacija podatkov z zemljevidi je zelo uporabna, saj uporabnikom omogoča razumevanje prostorskih podatkov. Z njo lahko uporabniki iz velike količine prostorskih podatkov razberejo tiste informacije, ki so zanje pomembne (Yu, Zhang & Sarwat, 2018). Vizualizacija podatkov z zemljevidi podatke preslika na dejanski zemljevid in s tem uporabnikom omogoča, da si prostorske podatke interpretirajo bolj razumljivo. Z vizualizacijo z zemljevidi uporabnikom postanejo vidni tudi razni vzorci in povezave, ki bi bili brez tega opaženi veliko težje. Zaradi boljšega razumevanja podatkov uporabnik prejme več informacij in zato tudi lažje sprejema odločitve (Dong in drugi, 2020).

Kljub uporabnosti vizualizacije podatkov pa se lahko pojavijo težave pri zagotavljanju tega, da so vizualizacije kakovostne. Pomemben proces pred vizualizacijo podatkov je njihova priprava. Samo petina vseh podatkov, ki se shranijo, je strukturirana, to pa predstavlja težavo pri njihovi analizi (Barik in drugi, 2017). Med zajemanjem podatkov prihaja do napak, ki jih je treba odpraviti, preden začnemo z analizo. Nekakovostni podatki vodijo v informacije, s katerimi si po koncu analize ne moremo pomagati oz. nismo prepričani o njihovi kakovosti (Baur in drugi, 2015). Zaradi izvajanja vizualizacije na nekakovostnih podatkih in slabše kakovosti informacij tudi končne odločitve uporabnika niso optimalne (Sinar, 2015). Zato je za dobro vizualizacijo zelo pomembno, da podatke prej ustrezno pripravimo.

Če moramo biti na eni strani pozorni na podatke, s katerimi delamo, pa ne smemo pozabiti tudi na vizualizacijo. Od kakovosti vizualizacije je odvisno, kako so podatki predstavljeni in kakšne informacije sporočajo uporabnikom. Slaba vizualizacija lahko sporoča napačne informacije in je težje berljiva, zaradi česar si jo uporabniki narobe interpretirajo. S kakovostno vizualizacijo postanejo tudi informacije bolj jasne, uporabnikove odločitve pa bolj kakovostne (Wilke, 2019). Pri vizualizaciji podatkov z zemljevidi imamo na voljo več vrst zemljevidov. Vsaka vrsta služi svojemu namenu, zato zemljevida ne smemo izbirati na podlagi navade ali izbire prepustiti naključju, ampak moramo biti pozorni, da je izbrani zemljevid res najbolj primeren za naš primer uporabe. Napačna izbira zemljevida vodi v manj kakovostne vizualizacije in posledično v manj kakovostne informacije, ki so

uporabniku predstavljene. Vizualizacija podatkov z zemljevidi je ob slabi izvedbi nekakovostna, kar predstavlja manj kakovostne informacije in neoptimalne odločitve uporabnikov (Słomska-Przech & Gołębiowska, 2021).

V literaturi je mogoče najti parcialne napotke za pripravo kakovostnih vizualizacij z zemljevidi, prav tako splošna vodila za kakovostne vizualizacije, zato je tema magistrskega dela narediti celovit pregled zagotavljanja kakovosti vizualizacije podatkov z osredotočanjem na posebnosti vizualizacij z zemljevidi. Namen magistrskega dela je prispevati k razumevanju, kako zagotoviti kakovostno vizualizacijo podatkov z zemljevidi.

Cilji magistrskega dela so:

- na podlagi literature ugotoviti, kaj so merila kakovostne vizualizacije podatkov z zemljevidi;
- podati smernice za pripravo kakovostnih podatkov, ki vključuje njihovo pripravo glede na pomanjkljivosti, ki jih imajo, in njihovo konfiguracijo glede na njihov tip;
- podati smernice glede priprave kakovostne vizualizacije, ki vključuje izbiro ustreznega zemljevida, izbiro pravih podatkov in pravilno konfiguracijo vizualizacije.

Pri izdelavi magistrskega dela sem uporabil znanja in veščine s področja poslovne inteligence in poslovne analitike, ki sem jih pridobil s podiplomskim študijem poslovne informatike na Ekonomski fakulteti Univerze v Ljubljani. Magistrsko delo temelji na preučevanju domače in tuje literature ter svetovnega spleta.

V prvem delu sem s preučevanjem literature in informacij na spletu predstavil vizualizacijo podatkov z zemljevidi in z njo tesno povezane komponente na splošno. Predstavil sem tudi različne vrste zemljevidov.

Drugi del je prav tako zasnovan na sekundarnih virih. V njem sem definiral kakovost podatkov na splošno, kakovost prostorskih podatkov, kakovostne vizualizacije na splošno in kakovostne vizualizacije z zemljevidi.

V tretjem delu sem izvedel vizualizacijo podatkov z zemljevidi na primeru. Pri izvedbi sem si zamislil problem in ga reševal s pomočjo ugotovitev, ki sem jih dosegel z raziskavo v prejšnjih poglavjih. Postopek sem začel s pridobivanjem podatkov in nadaljeval z njihovo pripravo. Ko so bili podatki ustrezno pripravljeni, sem izbral 3 zemljevide, ki izbrani problem najboljše rešujejo. Za vsak zemljevid sem izbral ustrezen nabor podatkov in mu ustrezno prilagodil nastavitve. Na koncu sem implementiral tudi vizualizacijo teh zemljevidov. Pri celotnem postopku je bilo temeljno vodilo reševanje nastavljenega problema z upoštevanjem smernic za kakovostno vizualizacijo podatkov z zemljevidi.

1 VIZUALIZACIJA PODATKOV Z ZEMLJEVIDI

1.1 Vizualizacija podatkov

Na svetu obstaja ogromna količina podatkov, ki iz dneva v dan narašča. Zaradi ogromne količine podatkov je te težko razumeti. Podatke je zaradi njihove jasnosti treba prikazati na razumljiv način. Odličen način za prikaz podatkov je uporaba vizualizacij. Gre za grafični prikaz podatkov, ki v berljivi in razumljivi obliki, z uporabo različnih vizualnih elementov, uporabnikom prikaže informacije, na podlagi katerih lahko sprejemajo odločitve (Mallon, 2015).

Za razliko od branja tabele podatkov nam dobra vizualizacija prikaže celotno sliko, s katere lahko hitro vidimo, kam se podatki nagibajo in kje so odstopanja. Največja dodana vrednost nastopi pri večjih količinah podatkov, ki bi jih bilo brez vizualizacij skoraj nemogoče razumeti (Tableau, brez datuma b).

To je v današnjem času zelo pogost pojav, saj živimo v obdobju, ko obstaja ogromno podatkov, ki jih je vsak dan čedalje le več. Razvoj tehnologije ne omogoča le hranjenja večje količine podatkov, ampak tudi njihovo lažje zbiranje. Podjetja imajo v lasti veliko podatkov, ki jih s pomočjo vizualizacij raziskujejo in si jih razlagajo. Vizualizacija omogoča odkrivanje vzorcev in povezav med podatki, prikaz informacij v obliki, ki je razumljiva čim večjemu številu ljudi (tudi tistim, ki si morda s področjem podatkov niso najbolj blizu), ter izboljšanje odločitev v primerjavi s tistimi, ki bi jih sprejeli na podlagi surovih podatkov (Sinar, 2015).

Zato je proces vizualizacije podatkov vpet v veliko različnih panog. Gre za proces, ki se ga je priporočljivo naučiti, saj so zaradi dodatnega razumevanja in sprejemanja boljših odločitev tudi rezultati boljši (Tableau, brez datuma b).

Poznamo več vrst vizualizacij. Ko govorimo o vizualizacijah, se ne smemo omejiti samo na osnovne diagrame, tabele ipd. To področje ponuja veliko več vizualizacij in podvizualizacij, kjer je vsaka uporabna za svoje primere uporabe, navsezadnje pa vsaka stremi k temu, da podatke prikaže na razumljiv način. Ena od tipov vizualizacij so tudi zemljevidi (Tableau, brez datuma b).

1.2 Vizualizacija podatkov z zemljevidi

Ena od tipov vizualizacij podatkov je vizualizacija podatkov z zemljevidi. To je vizualizacija podatkov, ki hranijo informacijo o lokaciji. Poleg lokacije pa ti podatki zaradi analize vsebujejo tudi podatke drugih atributov, ki se nanašajo na to lokacijo (Marzouki, Lafrance, Daniel & Mellouli, 2017).

Informacija o lokaciji je lahko shranjena v različnih oblikah, npr. naslov, zemljepisna širina in zemljepisna dolžina ipd. Razumevanje lokacijskih podatkov bi bilo brez njihove

vizualizacije izredno težko in bi v veliko primerih vključevalo napake, to pa bi vodilo v napačno sklepanje in odločitve. Zato je vizualizacija podatkov z zemljevidi zelo pomembna za njihovo razumevanje (Tableau, brez datuma a).

1.3 Geografski informacijski sistem

Za analizo in prikaz prostorskih podatkov potrebujemo informacijski sistem, ki zna ravnati s takšno vrsto podatkov. Takšna vrsta informacijskega sistema se imenuje geografski informacijski sistem (angl. geographic information system). Ta poleg prikaza prostorskih podatkov skrbi tudi za njihovo zbiranje, hranjenje, obdelavo in analizo (Wei, 2012).

Gre za informacijski sistem, ki podatke, ki so povezani z neko lokacijo, preslika na zemljevid. Uporabnikom z grafičnim prikazom omogoča lažje in boljše sprejemanje odločitev (Ali, 2020).

V splošnem geografski informacijski sistem lahko razdelimo na 5 delov. To so (Ali, 2020):

- Podatki: Gre za geografske podatke (angl. spatial data) in podatke drugih atributov (angl. attribute data). Geografski podatki hranijo lokacijo, medtem ko podatki drugih atributov predstavljajo podatke, ki nas za določeno lokacijo zanimajo, npr. količina nečesa.
- Programska oprema: Aplikacija z grafičnim vmesnikom, ki uporabniku omogoča hranjenje in analizo podatkov ter njihov prikaz na zemljevidu. Uporabnik lahko prek uporabniškega vmesnika vnaša in ureja podatke, nad njimi izvaja poizvedbe in jih vizualizira.
- Strojna oprema: Na njej je naložen in se izvaja geografski informacijski sistem.
- Ljudje: Uporabniki geografskega informacijskega sistema.
- Metode: Za učinkovite rezultate uporabe geografskega informacijskega sistema mora imeti organizacija tudi dobro in jasno definirano strategijo pridobivanja ter urejanja podatkov. Sem spadajo metode, ki definirajo načine pridobivanja podatkov, npr. merjenje, kje bodo ti podatki shranjeni, npr. podatkovna baza, načine razširjanja podatkov za pridobitev novih informacij in njihovo ureditev za večjo kompatibilnost z zunanjimi podatki.

Ko govorimo o komponentah geografskega informacijskega sistema, ne smemo mimo komponent programske opreme oz. zemljevidov. Ti so lahko sestavljeni iz več komponent, vsi dobri sistemi pa morajo imeti vsaj 3 glavne (Sergieieva, 2021):

- Legendo: Ta uporabniku pove, kateri deli zemljevida prikazujejo katere podatke, npr. katero kategorijo, kakšno količino ipd.
- Orodno vrstico: Gre za komponento, na kateri lahko uporabnik išče in filtrira podatke, menja prikazano raven, približuje in oddaljuje zemljevid ipd.
- Okno z informacijami: gre za okno z dodatnimi informacijami o določenem delu zemljevida, ki nam omogoča poglobljeno razumevanje, kaj določen del prikazuje.

Geografski informacijski sistemi podatke prikažejo na zemljevidih s procesom preslikave. To uporabniku zelo poenostavi in pohitri delo, saj podatke, ki jih uporabnik vnese oz. uvozi, npr. iz podatkovne baze, datoteke ipd., samodejno preslika na zemljevid. Uporabnik mora vnesti podatke, jih po potrebi urediti, geografski informacijski sistem pa njihov prikaz opravi sam. S preslikavo podatkov uporabniki te boljše razumejo, saj jim jih ni treba prebirati iz tabele, ampak z najoptimalnejšega prikaza prostorskih podatkov – zemljevidov. Prikazovati je možno tudi razne poizvedbe nad podatki in ne samo celotne tabele, tj. filtriranje le tistih podatkov, ki ustrezajo določenemu pogoju. Tako lahko uporabnik še bolj poglobljeno razišče, kar ga zanima. Vrste preslikave podatkov ločimo glede na vrste zemljevidov, zato poznamo več vrst preslikave. Primer preslikave je preslikava na toplotni zemljevid (Gigante, 2019).

Geografski informacijski sistem je zelo uporaben tudi zaradi svoje interaktivnosti. Ta uporabniku omogoča sprotno brskanje po informacijah, tj. prikaz informacij na več ravneh, izbiro lokacije in prikaz dodatnih informacij ob kliku ipd. S tem vizualizacije niso več statične, ampak dinamične. To ne pripomore le k večji skalabilnosti vizualizacij, ampak tudi k boljšemu in bolj poglobljenemu razumevanju podatkov in povezav med njimi (National Geographic, brez datuma).

1.4 Vrste zemljevidov

Poznamo več vrst zemljevidov, vsak pa služi svojemu namenu. Glede na medvladni odbor za geodezijo in kartiranje (angl. Intergovernmental Committee on Surveying and Mapping), poznamo v osnovi 2 vrsti zemljevidov (ICSM, brez datuma):

- topografske (angl. topographic) in splošne referenčne (angl. general reference) zemljevide – z njimi prikazujemo informacije o pokrajini;
- tematske (angl. thematic) zemljevide, s katerimi prikazujemo značilnosti lokacij na zemlji.

Hierarhična organizacija elementov na zemljevidih se pri tematskih zemljevidih razlikuje od topografskih in splošnih referenčnih zemljevidov. Če pri slednjih v večini primerov elemente prikazujemo na enaki ravni oz. z enako pomembnostjo, npr. kraje, meje, ceste, pa imajo pri tematskih zemljevidih večji poudarek atributi, ki jih za neko lokacijo prikazujemo, npr. populacija, kot pa sam zemljevid in njegovi elementi. Če npr. prikazujemo populacijo v posameznih državah, nas ostali elementi zemljevida, kot so ceste ali teren, ne zanimajo, zato jih ni treba prikazati oz. lahko damo naš atribut bolj v ospredje, da je na vrhu hierarhije in takoj opazen (Buckley, 2012).

Za lažje razumevanje lahko zgoraj omenjeni vrsti zemljevidov razširimo na 5 vrst, ki jih Barker (2015) po ICSM (brez datuma) opiše tako:

- Splošni referenčni: Gre za zemljevid, na katerem so prikazane naravne značilnosti, npr. reke in jezera, ter druge umetne značilnosti, ki se nahajajo na zemljevidu, npr. ceste in imena mest. Gre za splošen zemljevid, ki ga lahko uporabljamo za orientacijo.
- Topografski: Ta vrsta zemljevida prikazuje podrobnejše naravne značilnosti, med drugim tudi izočrte in nadmorsko višino.
- Tematski: Za razliko od splošnih referenčnih in topografskih zemljevidov, nam tematski zemljevidi ne pomagajo z orientacijo in predstavljajo naravnih značilnosti. Ampak nam predstavljajo informacijo na neko temo, ki pripada lokaciji. To sta lahko, npr. prikaz vremena ali gostote populacije. Tematske zemljevide med drugim uporabljajo tudi organizacije, za lažje sprejemanje odločitev.
- Navigacijske karte (angl. navigational charts): Ta vrsta zemljevidov se uporablja za navigiranje v pomorskem in letalskem prometu. Prikazujejo različne značilnosti, kot so ovire, npr. skale v morju, in pomagajo preprečevati nesreče.
- Katastrske karte in načrti (angl. cadastral maps and plans): Prikazujejo stavbe, meje med parcelami ipd.

Geografski informacijski sistemi za prikaz podatkov uporabljajo tematske zemljevide (Centers for Disease Control and Prevention, brez datuma). Obstaja več vrst tematskih zemljevidov. To uporabnikom geografskih informacijskih sistemov zagotavlja večjo izbiro in večjo verjetnost, da bodo podatke lahko prikazali na optimalen način (Tennekes, 2018). V naslednjih podpoglavjih je opisanih nekaj glede na literaturo in svetovni splet najbolj uporabljenih tematskih zemljevidov.

1.4.1 Zemljevidi s kategorijami

Zemljevidi s kategorijami (angl. category maps) so zelo uporabljena vrsta zemljevidov, kjer je zemljevid razdeljen na več delov, vsak del pa predstavlja svojo kategorijo. Vrednosti so na zemljevidu omejene z mejami območij in prikazane diskretno. Uporabimo ga, ko želimo predstaviti, katero območje pripada kateri kategoriji. Za to vrsto zemljevida je značilno, da ima vsaka kategorija svojo barvo (Sergieieva, 2021).

1.4.2 Horoplet zemljevidi

Horoplet zemljevidi (angl. choropleth maps) se od zemljevidov s kategorijami razlikujejo v tem, da je na zemljevidu uporabljena zgolj ena barva, ki je na različnih delih prikazana z drugačnim odtenkom (GISGeography, brez datuma a). Vrednosti so na zemljevidu prav tako omejene z mejami območij (države, regije, občine ipd.) in prikazane diskretno (Trame & Keßler, 2011). Ljudje ga velikokrat zamenjajo z zemljevidom s kategorijami, se pa od njega poleg zgoraj omenjenih lastnosti razlikuje tudi v tem, da posamezno območje ne predstavlja kategorije, ampak količino podatkov na tistem območju (Tennekes, 2018).

Zaradi prikazovanja količine za večja območja na zemljevidu ni prikazanih točnih lokacij. Zemljevid uporabimo, ko nas ne zanimajo točne lokacije, ampak želimo prikazati vrednosti za določena diskretna območja (Guldåker, 2020). Zaradi prikazovanja podatkov za določena območja je priporočljivo, da so podatki enakomerno porazdeljeni po posameznem območju. V nasprotnem primeru nas prikaz lahko zavaja, saj bi lahko en del območja imel nadpovprečno, ostali deli pa podpovprečno gostoto podatkov (Centers for Disease Control and Prevention, brez datuma).

Prednost uporabe horoplet zemljevida je tudi ta, da gre za zelo priljubljen zemljevid, zato ga veliko ljudi razume brez dodatne razlage. Zemljevid je tudi že sam po sebi sestavljen tako, da je hitro razumljiv. Ob uporabi moramo biti pozorni na to lastnost zemljevida, da je namenjen podatkom, ki se nanašajo na neko generalno območje in ne na specifične lokacije, zato ga je priporočljivo uporabljati pri podatkih, ki predstavljajo vrednost diskretnega območja (Maptive, 2020).

1.4.3 Toplotni zemljevidi

Tako kot pri horoplet zemljevidih, tudi pri toplotnih zemljevidih (angl. heat maps) prikazujemo razliko v količini oz. gostoti med posameznimi območji (GISGeography, brez datuma a). Toplotni zemljevidi se od horoplet zemljevidov razlikujejo v tem, da imamo pri horoplet zemljevidih vrednosti omejene z mejami območij (države, regije, občine ipd.), toplotni zemljevidi pa geografskih meja ne poznajo. Vrednosti toplotnih zemljevidov se prosto gibljejo prek geografskih območij, glede na vrednosti, ki jih ta območja zasedajo (Trame & Keßler, 2011).

Količina oz. gostota je na zemljevidu prikazana barvno, kjer se barve gibljejo od toplih (večja gostota) proti mrzlim (manjša gostota). Uporaba barve za razločitev toplih in mrzlih vrednosti ni natančno določena, v največ primerih pa sta uporabljena modra za mrzlo in rdeča za toplo barvo (Sergieieva, 2021).

Toplotni zemljevidi ne prikazujejo točnih lokacij, kjer se vrednosti nahajajo. Vseeno pa zaradi svoje neomejenosti z mejami prikazujejo bolj točne vrednosti za določeno območje, kot to prikazujejo horoplet zemljevidi. Iz njih lahko učinkovito primerjamo koncentracijo vrednosti med različnimi območji (Guldåker, 2020).

1.4.4 Zemljevidi skupkov točk

Zemljevidi skupkov točk (angl. cluster maps) so podobna vrsta zemljevida, kot so zemljevidi s točkami, npr. zemljevid gostote točk, ki je opisan v poglavju Zemljevidi gostote točk. Ker imamo lahko pri zemljevidih s točkami veliko točk, ki so lahko prikazane precej skupaj in jih je zato težko razbrati, te točke združujemo v skupne točke. Tako dobimo zemljevid z manj točkami, kjer ima vsaka točka prikazano tudi informacijo o tem, koliko podatkov

predstavlja, npr. koliko točk je bilo združenih vanjo. Zemljevid skupkov točk je odlična rešitev za izboljšanje berljivosti v primeru velikega števila točk na zemljevidu (Sergieieva, 2021).

1.4.5 Zemljevidi z mehurčki

Zemljevidi z mehurčki (angl. bubble maps) prikazujejo razliko v količini podatkov na različnih območjih. Na zemljevidu imamo prikazane mehurčke, katerih velikost se spreminja s količino podatkov na tistem območju. Če je podatkov več, je prikazan mehurček večji, in obratno. Količino podatkov za neki mehurček lahko razberemo iz legende. Je zelo lahko berljiv zemljevid, s katerega lahko hitro razberemo, kje je podatkov več in kje manj (Sergieieva, 2021).

1.4.6 Zemljevidi gostote točk

Pri zemljevidih gostote točk (angl. dot density maps) so podatki prikazani s točkami na zemljevidu. Z zemljevida lahko s tem, da preverimo gostoto točk, hitro razberemo, kje je gostota podatkov večja in kje manjša. V primerjavi s toplotnim in horoplet zemljevidom, zemljevid gostote točk prikazuje podatke bolj natančno, saj vidimo, kje na zemljevidu se posamezne točke nahajajo. Za razliko od zemljevida z graduiranimi simboli in zemljevida s proporcialnimi simboli, ki sta opisana v poglavjih Zemljevidi z graduiranimi simboli in Zemljevidi s proporcialnimi simboli, in gostoto podatkov predstavljata z velikostjo točk, so tukaj vse točke prikazane z enako velikostjo (Maptive, 2020). Uporaba zemljevida s točkami je zelo uporabna, ko želimo prikazati točne lokacije na zemlji (Dougherty & Ilyankou, 2021). Točka na zemljevidu ni omejena na eno, temveč lahko predstavlja več vrednosti, mora pa vsaka točka predstavljati enako število vrednosti (Caliper, brez datuma).

1.4.7 Zemljevidi z graduiranimi simboli

Zemljevid z graduiranimi simboli (angl. graduated symbol map) prav tako kot zemljevid gostote točk podatke prikaže kot točke. Razlika je v tem, da pri zemljevidu z graduiranimi simboli podatke razvrstimo v kategorije. Različne kategorije predstavljajo različno gostoto podatkov za neko območje. Različne kategorije se ne razlikujejo v barvi, ampak v velikosti. Če je gostota podatkov večja, je tudi točka na zemljevidu večja, v nasprotnem primeru pa manjša. Velikosti točk so vnaprej določene in predstavljene v legendi (GISGeography, brez datuma b).

Dobra lastnost zemljevidov z graduiranimi simboli je, da lahko istočasno poleg gostote podatkov prikažemo tudi porazdelitev podatkov v podkategorije za to območje. Primer je uporaba tortnega diagrama za prikaz točk. V tem primeru velikost točke še vedno pomeni količino podatkov, iz same točke pa lahko razberemo, kako so ti podatki porazdeljeni med posamezne podkategorije (Maptive, 2020).

Pri razdelitvi podatkov v kategorije moramo paziti, da slednjih ni preveč. Uporabniki imajo pri velikem številu kategorij težavo z ugotavljanjem, katera točka pripada kateri kategoriji. Težava se pojavi zaradi majhne razlike v velikostih različnih kategorij. V primeru, da imamo preveč kategorij oz. podatkov ne želimo razvrstiti v kategorije, lahko uporabimo zemljevid s proporcialnimi simboli, ki je opisan v poglavju Zemljevidi s proporcialnimi simboli (Maptive, 2020).

1.4.8 Zemljevidi s proporcialnimi simboli

Prav tako kot pri zemljevidih z graduiranimi simboli, tudi pri zemljevidih s proporcialnimi simboli (angl. proportional symbol maps) podatke prikazujemo s točkami, ki se glede na gostoto podatkov razlikujejo v velikosti (Tennekes, 2018). Razlika je v tem, da pri zemljevidih s proporcialnimi simboli nimamo določenih kategorij. Znana je informacija o tem, kakšno vrednost zasedata najmanjša in največja točka na zemljevidu, ostale točke, ki so predstavljene z vmesnimi velikostmi, pa so izdelane proporcialno glede na podani točki. Točka lahko zavzame eno od velikosti, ki se nahaja med velikostma najmanjše in največje točke, glede na gostoto podatkov, ki jo točka predstavlja (GISGeography, brez datuma b).

2 KAKOVOST PODATKOV

Doseganje kakovosti podatkov je pred njihovo uporabo zelo pomembno. Za pridobitev kakovostnih in pravilnih informacij morajo biti tudi podatki kakovostni (Baur in drugi, 2015). Rezultat uporabe nekakovostnih podatkov so napačne vizualizacije, zaradi katerih tudi odločitve uporabnikov niso optimalne (Sinar, 2015). Cilj doseganja kakovosti podatkov je, da lahko končni uporabniki iz teh podatkov pridobijo kakovostne informacije, na podlagi katerih lahko sprejemajo odločitve.

Ko govorimo o fizičnih izdelkih v svetu, se lahko o njihovi kakovosti pogovarjamo precej preprosto, saj imajo ti dostopne in vidne fizične značilnosti. Ko pa govorimo o kakovosti podatkov, teh oprijemljivih značilnosti nimamo na voljo. Na voljo imamo le neoprijemljive značilnosti. Zato kakovosti podatkov ne moremo aplicirati na splošno na vse podatke, ampak se ta razlikuje glede na primer uporabe. Kakovostni podatki so podatki, ki ustrezajo določenim merilom kakovosti za določen primer uporabe (Veregin, 1999).

Triglav (v Starček in Kovač, 2019, str. 19) opredeli kakovost prostorskih podatkov tako: »Prostorski podatki so raznovrstni tako po položajni in časovni kakovosti kot po kakovosti pomenske opredeljenosti pojmov. Opredelitev kakovosti prostorskih podatkov je odvisna od področja obravnave, namena, zahtev in pričakovanj uporabnikov ter drugih subjektivnih dejavnikov. Na splošno izraža kakovost prostorskih podatkov celotnost lastnosti zbirke podatkov glede na njeno sposobnost, da ustreza izraženemu ali vsebovanemu nizu zahtev. Je razlika med podatki in stvarnim svetom, ki ga podatki ponazarjajo. Večja je ta razlika,

slabša je kakovost podatkov, s tem sta manjši tudi uporabna in siceršnja vrednost teh podatkov.«

Kakovost podatkov ocenjujemo s pomočjo dimenzij kakovosti podatkov. Obstaja več različnih dimenzij, vsaka dimenzija pa preverja, ali podatki zadostujejo standardom kakovosti, ki so bili zanje določeni. Podjetje mora samo ovrednotiti, katere dimenzije bo uporabilo, kakšne uteži bo, glede na kontekst uporabe podatkov, kateri dimenziji določilo, ter določiti, katere vrednosti podatkov so sprejemljive in katere ne. Izbira dimenzij je odvisna tudi od poslovnih potreb podjetja in panoge, s katero se podjetje ukvarja (Askham in drugi, 2013). Kakovost podatkov s pomočjo dimenzij lahko izvedemo v nadaljevanju opisanih procesih profiliranja ali pred fazo preoblikovanja pri procesu pridobivanja-preoblikovanja-nalaganja (angl. Extract-Transform-Load, v nadaljevanju ETL).

Preden podatke naložimo v naše podatkovno skladišče, je priporočljivo, da nad njimi izvedemo profiliranje. Namen profiliranja je, da se spoznamo s strukturo in vsebino podatkov ter odkrijemo nepravilnosti, ki jih podatki vsebujejo. Tako se s podatki že takoj spoznamo, odločimo, ali so ti podatki primerni za nas in dobimo predstavo o tem, koliko dela bo potrebnega, da podatke spravimo v kakovostno obliko (Elena, 2011).

Po profiliranju lahko začnemo s procesom ETL. Sestavljajo ga pridobivanje, preoblikovanje in nalaganje podatkov. V tem procesu podatke pridobimo, jih preoblikujemo (v fazi preoblikovanja podatke spreminjamo in izboljšujemo njihovo kakovost), tako da ustrezajo merilom kakovosti za naš namen uporabe, in jih naložimo v naše podatkovno skladišče za nadaljnjo uporabo (Ong, Siew & Wong, 2011).

2.1 Dimenzije kakovosti podatkov

V različnih virih avtorji definirajo različne modele kakovosti podatkov, ki so si po večini sorodni, se pa med seboj razlikujejo v kateri od dimenzij. Izbira modela se med podjetji razlikuje, saj je odvisna od tega, za kakšen namen podatke potrebujemo in kakšni so ti podatki (Askham in drugi, 2013).

Poznamo 6 dimenzij, ki so generične in se jih lahko adaptira v veliko primerih (Askham in drugi, 2013):

- Popolnost (angl. completeness): Delež podatkov, ki jih imamo shranjene glede na vse podatke v resničnem svetu.
- Konsistentnost (angl. consistency): Ali je podatek konsistenten v različnih virih, npr. vrednost ali format.
- Edinstvenost (angl. uniqueness) – vsaka entiteta v resničnem svetu se mora v podatkovnem viru pojaviti zgolj enkrat.
- Veljavnost (angl. validity): Ali so podatki skladni s pravili, ki so zanje določene, npr. tip in struktura.

- Točnost (angl. accuracy): V kolikšni meri podatki pravilno predstavljajo entiteto v resničnem svetu.
- Pravočasnost (angl. timeliness): V kolikšni meri podatki predstavljajo resnično stanje glede na zahtevani čas.

2.2 Profiliranje podatkov

Profiliranje podatkov je proces, katerega namen je boljše razumevanje vsebine in strukture podatkov, zato ga je priporočljivo izvesti že pred uporabo oz. nalaganjem podatkov v naše podatkovno skladišče (Elena, 2011). S profiliranjem lahko že na samem začetku odkrijemo nepravilnosti, ki jih podatki vsebujejo in si s tem ustvarimo boljšo predstavo o tem, koliko dela in kakšno delo bo potrebno, da podatke preoblikujemo do te mere, da bodo za nas uporabni (Chaudhuri, Dayal & Narasayya, 2011).

S profiliranjem analiziramo podatke in iz njih izluščimo različne statistične ter metapodatke. Primeri metapodatkov, ki jih s profiliranjem dobimo, so podatkovni tipi stolpcev oz. atributov, količina praznih vrednosti v posameznem stolpcu, podatek o najpogostejših vrednostih v stolpcu, podatek o tem, koliko različnih vrednosti se v stolpcu pojavi, podatek o tem, kateri so primarni in kateri tuji ključi v tabelah ipd. Profiliranje podatkov se uporablja na različne načine in v različnih primerih. Znanje, ki ga pridobimo s profiliranjem podatkov, igra pomembno vlogo pri hranjenju združljivih in strukturiranih podatkov, ki ustrezajo merilom naše uporabe (Naumann, 2014).

Primeri uporabe so (Naumann, 2014):

- Optimizacija poizvedb: Profiliranje za namen optimizacije poizvedb izvajajo sistemi za upravljanje s podatkovnimi bazami (angl. database management systems), kjer s pomočjo različnih statističnih podatkov, ki jih pridobijo iz podatkovne baze, podajo informacije, ki jih lahko uporabimo za optimizacijo poizvedb.
- Čiščenje podatkov: Profiliranje podatkov za namen čiščenja je zelo priljubljen primer uporabe. Tu odkrijemo pomanjkljivosti naše podatkovne zbirke. Primeri pomanjkljivosti oz. napak so prazne vrednosti, različni podatkovni tipi vrednosti v stolpcu in različen format vrednosti v stolpcu. Če imamo v organizaciji definirana pravila oz. merila, kakšni podatki zadoščajo uporabi, lahko s takšnim profiliranjem izvemo tudi, kolikšen delež podatkov ustreza našim merilom.
- Integracija podatkov: Preden podatke naložimo v naše podatkovno skladišče, jih moramo bolje spoznati. S profiliranjem boljše razumemo njihovo vsebino, kako so posamezne tabele med seboj povezane, za kakšno količino podatkov gre, katere atribute imamo na voljo, kakšne podatkovne tipe hrani kateri atribut, v kakšnem formatu so shranjene vrednosti posameznih stolpcev ipd.
- Upravljanje znanstvenih podatkov: Podatkovno profiliranje svojo vrednost najbolj pokaže pri surovih podatkih, ki pred dodajanjem v podatkovno bazo niso bili urejani. Pri

podatkih, ki so bili zaradi narave pridobivanja v podatkovno bazo dodani v surovi obliki, npr. spletno luščenje podatkov, je profiliranje zelo pomembno. S tem podatke lažje spravimo v obliko, ki bo naši uporabi ustrezala.
- Analiza podatkov: Pred vsako analizo podatkov je zelo priporočljivo izvesti profiliranje podatkov. S tem izvajalec analize ali drugih procesov nad podatki te bolje razume in njim primerno tudi prilagodi informacijski sistem, kjer jih bo uporabljal.

2.3 Priprava podatkov

Proces ETL je sestavljen iz 3 faz: pridobivanja, preoblikovanja in nalaganja podatkov (Ong, Siew & Wong, 2011). V nadaljevanju so opisane vse 3 faze, najbolj podrobno pa je zaradi namena magistrskega dela opisano njihovo preoblikovanje.

2.3.1 Pridobivanje

Pridobivanje podatkov je faza, pri kateri izberemo podatke, ki so potrebni za našo analizo. Podatke lahko pridobimo iz internih in zunanjih virov (Ong, Siew & Wong, 2011).

2.3.2 Preoblikovanje

Preoblikovanje podatkov je edina faza, v kateri pridobljene podatke spreminjamo. Tu podatke preoblikujemo tako, da so ti uporabni za nadaljnjo analizo. Pri tem upoštevamo pravila, definirana v podjetju (Ong, Siew & Wong, 2011).

Založnik (2018) preoblikovanje podatkov deli na več operacij:

- **Izbira atributov**: Pri pridobitvi podatkov, ti lahko vsebujejo tudi atribute, ki jih pri analizi ne bomo potrebovali. Ohranjanje teh atributov je nesmiselno, saj nam v podatkovnem skladišču zasedajo dodaten prostor, kar vodi v dodatne stroške. Hkrati večja količina podatkov vodi v počasnejše poizvedbe in s tem upočasnjuje analizo. Zato lahko s to operacijo izberemo samo tiste atribute, ki jih bomo v nadaljevanju potrebovali.
- **Pretvorba podatkovnih tipov**: V koraku pridobivanja podatkov sem omenil, da te lahko pridobimo iz različnih virov. Zaradi tega so lahko podatki v različnih formatih, kar pomeni, da niso kompatibilni. Zato moramo podatke formatirati tako, da so med seboj skladni. Primer je atribut, ki hrani število. Ta bi lahko bil v podatkovni bazi shranjen kot število (angl. integer) ali besedilo (angl. string). Če bi želeli v tem primeru nad vnosi v bazi izvajati matematične operacije, to na besedilu ne bi bilo možno.
  Druga težava, ki se pojavi zaradi združevanja podatkov iz različnih virov, je, da so lahko ti viri shranjeni z različnim kodiranjem, npr. s formatom transformacije Unicode, 8-bitnim (angl. Unicode Transformation Format – 8-bit - UTF-8), kar lahko vodi v neupoštevanje oz. napačno prikazovanje podatkov. Zato moramo v tem koraku preveriti

tudi vrsto kodiranja posameznih virov in jih poenotiti ter biti pozorni na ohranjanje podatkov v pravilnem stanju.

- **Čiščenje podatkov in zagotavljanje kakovosti**: Zaradi obsega je ta operacija opisana v poglavju Čiščenje podatkov.
- **Povezovanje podatkov**: Podatke moramo po pridobitvi združiti v celote, ki bodo smiselne in priročne za našo analizo. Pri tem koraku govorimo o združevanju različnih podatkovnih baz in drugih virov podatkov ter same tabele znotraj podatkovnih baz (v primeru relacijskih podatkovnih baz). Pri tem moramo biti pozorni, da podatke združujemo glede na naš primer kasnejše uporabe.
- **Izpeljanke**: Včasih atributi, ki jih imamo na voljo, niso dovolj za potrebe naše analize, zato ustvarimo nove. Te kreiramo iz atributov, ki so na voljo. Primer novega atributa je »število_neaktivnih_uporabnikov«, ki predstavlja razliko med vrednostma atributov »število_vseh_uporabnikov« in »število_aktivnih_uporabnikov«.
- **Agregacija**: Za neke podatke obstaja več načinov, v kakšni obliki so ti lahko shranjeni. Za obliko shranjevanja se odloči uporabnik oz. podjetje, ki bo te podatke uporabljalo, razlog za določitev neke oblike pa je poslovni problem. Glede na poslovni problem podatke preoblikujemo tako, da jih združujemo v smiselne celote, ki nam bodo prav prišle v nadaljevanju. Primer takšnega združevanja je združevanje vnosov po datumu. Recimo, da potrebujemo podatke o količini porabljene vode za posamezen mesec v letu 2022, v podatkovni bazi pa imamo podatke za vse dni v letu. V tem primeru bi podatke združili glede na atribut meseca in jih tako zreducirali na potrebnih 12 vnosov. S tem se znebimo tudi odvečne porabe prostora in pohitrimo proces uporabe podatkov.
- **Revizija informacij in skladnost sistema**: Pri tej operaciji gre za primerjavo preoblikovanih podatkov s surovimi podatki, ki smo jih pridobili na začetku. Primer primerjave je število tabel ali vrstic v tabelah podatkovne baze. S to primerjavo preverimo, da pri preoblikovanju podatkov ni prišlo do kakšne neujete napake.
- **Upravljanje manjkajočih vrednosti**: Manjkajoče vrednosti so lahko shranjene v različnih oblikah, npr. NULL, »« in 0. S to operacijo manjkajoče vrednosti poenotimo, drugače imamo lahko pri analizi in vizualizaciji težave. Pri tej operaciji moramo paziti, da manjkajoče vrednosti razlikujemo od tistih, ki zasedajo prazno oz. ničto vrednost, npr. »« in 0. Zato je pomembno, da podatke pred tem dobro razumemo in šele potem nad njimi izvajamo preoblikovanje.

2.3.3 Nalaganje

Nalaganje podatkov je zadnja faza. V tej fazi preoblikovane podatke naložimo na želeno lokacijo, kjer bomo podatke hranili (Ong, Siew & Wong, 2011).

2.4 Čiščenje podatkov

Čiščenje podatkov je proces, ki ga izvedemo po tem, ko smo analizirali kakovost podatkov. S čiščenjem nekakovostne podatke pretvorimo v kakovostne podatke, ki bodo ustrezni za našo analizo (Ganti & Sarma, 2013).

Čiščenje podatkov je sestavljeno iz 5 korakov (Sheoran & Parmar, 2022, str. 54):

- Analiza podatkov: Najprej moramo podatke analizirati in s tem odkriti njihovo kakovost. V tem koraku odkrijemo, za kakšne morebitne napake gre, da jih bomo lahko v nadaljevanju naslovili.
- Določitev poteka preoblikovanja podatkov: Za podatke iz različnih virov je treba določiti, kako bomo te podatke preoblikovali, da bodo združeni podatki shranjeni z enakimi pravili, npr. format, imena atributov ipd. in da bodo podatki iz posameznega vira na koncu kakovostni.
- Testiranje podatkov: V tem koraku s pomočjo dimenzij ocenjevanja kakovosti podatkov preverimo njihovo kakovost pred začetkom preoblikovanja.
- Preoblikovanje: Preoblikovanje oz. čiščenje podatkov.
- Shranjevanje podatkov: Nekakovostne podatke v podatkovnem skladišču nadomestimo z očiščenimi.

Poleg operacij, opisanih v poglavju Preoblikovanje, Rapid Insight navaja tudi naslednje operacije čiščenja (Rapid Insight, brez datuma):

- **Filtriranje**: Odstranjevanje vnosov (vrstic), ki jih ne potrebujemo. Vnose v tabeli filtriramo glede na vrednosti atributov. Primer: Če imamo v tabeli shranjene vse elemente periodnega sistema, za našo analizo pa potrebujemo samo kovine, lahko tabeli filtriramo tako, da na koncu vsebuje samo vnose, ki za atribut »tip« hranijo vrednost »kovina«.
- **Odstranjevanje duplikatov**: Iz tabele odstranimo podvojene vnose, tako da so na koncu vsi vnosi unikatni. Unikatnost vnosov preverjamo glede na primarni ključ, ki se mora v tabeli pojaviti samo enkrat.
- **Transformacija**: Kreiranje novih vrednosti iz stolpcev. Primer je združevanje več stolpcev v skupen stolpec, npr. združevanje stolpcev »ime« in »priimek« v stolpec »celo ime«. V tem primeru smo združili 2 stolpca in se vrednost ni spremenila, lahko pa s transformacijo ustvarjamo tudi nove vrednosti.
- **Standardizacija**: Če v nekem stolpcu vnosi enako vrednost predstavljajo drugače, npr. moški, M, m, je te vrednosti treba pretvoriti tako, da bodo shranjene na enak način.

3 KAKOVOST PROSTORSKIH PODATKOV

Med prostorske podatke uvrščamo vse podatke, ki v vsaj enem od svojih atributov hranijo lokacijo na zemlji. Pred uporabo prostorskih podatkov je zelo pomembno, da te pred uporabo

ustrezno očistimo. Nekakovostni podatki vodijo v nepravilne rezultate, zaradi katerih so tudi odločitve na podlagi teh rezultatov neoptimalne. S čiščenjem prostorskih podatkov odstranimo ali popravimo podatke, ki bi lahko zaradi svojih pomanjkljivosti ali nepravilnosti negativno vplivali na rezultate raziskave. Pred procesom čiščenja moramo odkriti, kateri podatki so problematični. To so podatki, ki imajo manjkajoče, podvojene, nedosledne ali nepotrebne vrednosti. S čiščenjem izboljšamo kakovost teh podatkov in s tem kakovost celotne podatkovne baze (Parmar & Sheoran, 2021). Poleg manjkajočih, podvojenih, nedoslednih ali nepotrebnih vrednosti s čiščenjem naslovimo tudi napačno zapisane podatke, npr. napačno črkovane vrednosti (Sheoran & Parmar, 2022, str. 52).

Satyanarayana in Guptha (2021) pri procesu čiščenja prostorskih podatkov opozarjajo na napake, ki jih moramo odpraviti: nepopolni podatki, prazne vrednosti, duplikati in nekonsistentni podatki. Ker je proces ročnega čiščenja pri veliki količini podatkov zamuden, si lahko pomagamo s temu namenjenimi programi.

Razlog za slabo kakovost prostorskih podatkov je lahko tudi ta, da podatke združujemo iz različnih virov. Podatki iz različnih virov lahko že sami po sebi vsebujejo napake, poleg tega so ti lahko v različnih virih shranjeni z različnimi pravili, npr. format in struktura. Zato je pred uporabo teh podatkov, te nujno potrebno očistiti (Eldrandaly, Abdel-Basset & Shawky, 2019).

Eden od največjih povzročiteljev težkega upravljanja prostorskih podatkov je, da podatke lahko pridobivamo iz različnih virov. Zaradi tega podatki niso enotni, ampak so shranjeni v različnih formatih. Zato je zelo pomembno, da pred uporabo teh podatkov v geografskem informacijskem sistemu, podatke očistimo in pretvorimo v skupen format. Težava se lahko pojavi tudi pri uporabi enega podatkovnega vira, saj se nam lahko zgodi, da posamezni atributi niso shranjeni v enakem formatu. Razlog za to je lahko neupoštevanje pravil formatov pri vnašanju oz. zbiranju podatkov ali pa zbiranje podatkov ob različnih časih (Chiang, Wu, Anand, Akade & Knoblock, 2014).

Koraki, pri katerih lahko pride do napak v prostorskih podatkih, so (Koshley & Halder, 2015):

- zbiranje podatkov,
- vnašanje podatkov,
- shranjevanje podatkov,
- obdelava podatkov,
- prikaz podatkov.

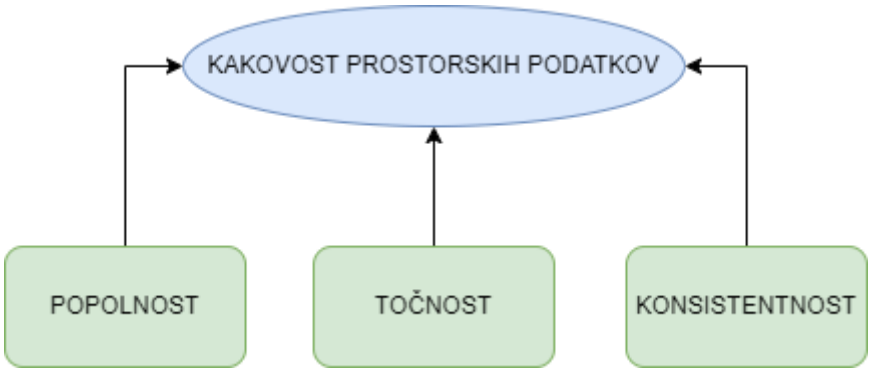
3.1 Elementi kakovosti prostorskih podatkov

Ko govorimo o elementih kakovosti prostorskih podatkov, se ti nekoliko razlikujejo od modelov, ki se uporabljajo za ocenjevanje kakovosti podatkov na splošno oz. so ti prilagojeni

za delo s prostorskimi podatki. Parmar in Sheoran (2021, 2022) predstavita model, ki ni bistveno drugačen od splošnih modelov kakovosti podatkov, ampak ima sorodne elemente, le da so ti prilagojeni delu s prostorskimi podatki. Predstavljen model je prikazan na sliki 1 in kakovost prostorskih podatkov ocenjuje na podlagi 3 elementov (Sheoran & Parmar, 2021, str. 55):

- Popolnost: Koliko entitet imamo shranjenih in koliko izpuščenih v primerjavi z entitetami v resničnem svetu. Meri se lahko z različnimi enotami, npr. delež vsebovanih entitet ali število vsebovanih in manjkajočih entitet.
- Točnost: Časovna in prostorska točnost vrednosti, glede na dejanske vrednosti v resničnem svetu.
- Konsistentnost: Preverja, ali so podatki shranjeni skladno s pravili naše podatkovne strukture, npr. tip vrednosti in njen format.

Slika 1: Elementi kakovosti prostorskih podatkov



Prirejeno po Sheoran & Parmar (2022).

Na voljo pa so tudi modeli, ki elemente kakovosti prostorskih podatkov bolj konkretizirajo. Bindzárová Gergel'ová in drugi (2020) ugotavljajo, da veliko znanstvenih virov uporablja standard, ki ga je razvila mednarodna organizacija za standardizacijo (angl. International Organization for Standardization, v nadaljevanju ISO). To je standard ISO 19157:2013. Ferster, Nelson, Roberston in Feick (2018) prav tako ugotavljajo, da je navedeni standard, ki zajema veliko elementov, dobro izhodišče za ocenjevanje kakovosti prostorskih podatkov.

Namen standarda ISO 19157 je predstaviti oz. definirati kakovost prostorskih podatkov s pomočjo različnih elementov. Rezultate ocenjevanja kakovosti prostorskih podatkov lahko kasneje uporabijo različni deležniki, ki jih ta informacija zanima (Fonte in drugi, 2017).

Standard ISO 19157 kakovost prostorskih podatkov deli na 2 dela: notranjo kakovost (angl. internal quality) in zunanjo kakovost (angl. external quality). Notranja kakovost predstavlja kakovost z vidika tistega, ki podatke ustvarja oz. ureja, zunanja kakovost pa z vidika končnega uporabnika teh podatkov (Devillers & Jeansoulin, 2006). Zato moramo vedeti, da

se lahko zgodi, da imamo notranjo kakovost podatkov visoko, kar pa še ne pomeni, da je zato visoka tudi zunanja kakovost; ni nujno, da so podatki, ki imajo visoko notranjo kakovost, primerni za reševanje težave končnega uporabnika. Pri ugotavljanju kakovosti podatkov, moramo biti pozorni na notranjo in zunanjo kakovost (Fonte in drugi, 2017).

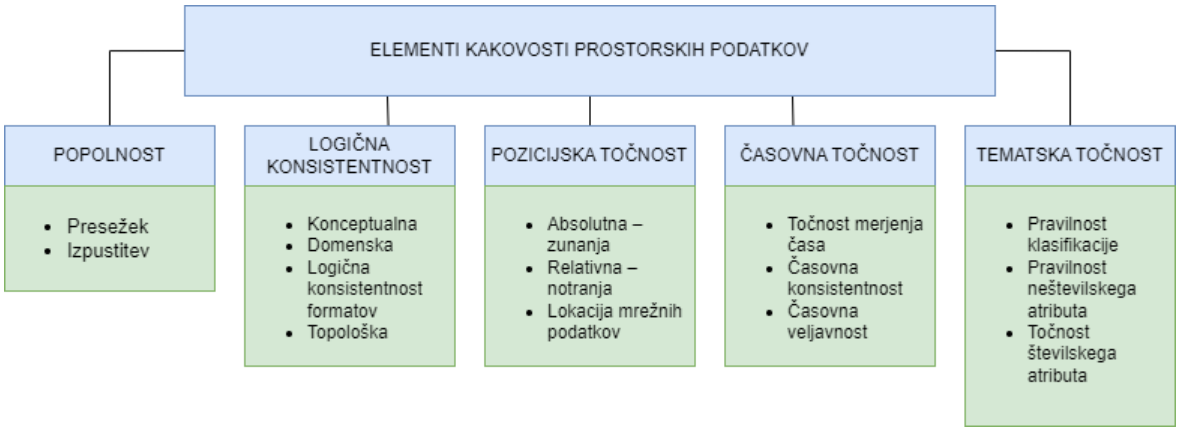
Fonte in drugi (2017), International Organization for Standardization (2013) in Namibia Statistics Agency (2016) elemente standarda ISO 19157 opišejo tako:

- Pozicijska točnost (angl. positional accuracy): Gre za vrednost, ki nam pove, kako točno lokacijo imajo entitete v naši podatkovni bazi glede na vrednosti v resničnem svetu. Pozicijsko točnost lahko ugotovimo s pomočjo dodatne podatkovne baze, katere podatki predstavljajo enako časovno obdobje kot naši podatki in za naše entitete hranijo pravilne vrednosti. Tako lahko entitete v naši podatkovni bazi primerjamo z referenčno podatkovno bazo in izračunamo, kakšna so odstopanja od pravih lokacij. Absolutna pozicijska točnost (angl. absolute positional accuracy) predstavlja bližino med koordinatami v naši bazi in pravilnimi koordinatami na zemlji. Relativna pozicijska točnost (angl. relative positional accuracy) predstavlja bližino relativnih pozicij med entitetami v podatkovni bazi, glede na pravilne relativne pozicije. Zadnji podelement je absolutni pozicijski točnosti soroden v tem, da preverja, kako blizu so lokacije podatkov v naši podatkovni bazi pravilnim podatkom na zemlji, le da to preverja za mrežne podatke (angl. gridded data position).
- Tematska točnost (angl. thematic accuracy): Tematska točnost ne preverja točnosti atributov, ki predstavljajo lokacijo, ampak atribute, ki to lokacijo opisujejo. Primer bi lahko bil atribut, ki opisuje vrsto stavbe, npr. banka, pošta, šola ipd. Tematsko točnost lahko tako kot pri pozicijski točnosti preverimo s primerjavo referenčne podatkovne baze. Paziti pa moramo na podatke, ki niso zgolj binarni oz. zasedajo eno od n-kategorij. Lahko se zgodi, da en podatek v resničnosti zaseda 2 vrednosti ali pa da se je vrednost skozi čas spremenila, npr. sprememba lastnika stavbe. Zato je priporočljivo, da imamo pri preverjanju tematske točnosti nekaj vsebinskega znanja, da lahko lažje določimo, ali gre za pravo vrednost ali ne. Pravilnost neštevilskega atributa (angl. non-quantitative attribute corectness) preverja ali je vrednost neštevilskega atributa pravilna. Točnost številskega atributa (angl. quantitative attribute accuracy) preverja, kako blizu je številska vrednost v bazi pravilni vrednosti. Pravilnost klasifikacije (angl. classification correctness) preverja, ali so vrednosti atributov, ki jim pripada kategorična vrednost, pravilno klasificirani.
- Popolnost (angl. completeness): Popolnost nam pove, kolikšen presežek (angl. commission) oz. kolikšno izpustitev (angl. omission) entitet, njihovih atributov in povezav med njimi, predstavlja naša podatkovna baza. Lahko se zgodi, da je bilo merjenje podatkov pristransko bodisi zaradi potrebe po podatkih iz samo določenih območij bodisi zaradi potrebe po samo določenih tipih podatkov in se drugih podatkov ni merilo oz. vnašalo ali pa se nanje ni dalo tolikšnega poudarka.

- Časovna točnost (angl. temporal accuracy) oz. časovna kakovost (angl. temporal quality): Sem spadajo datum zbiranja, datum objave, pogostost posodabljanja, datum zadnje posodobitve, časovna veljavnost (angl. temporal validity ali curentness) in drugi metapodatki, ki se nanašajo na čas podatkov. Primer uporabnosti časovne točnosti je, da vidimo kdaj so bili podatki zbrani in s tem vemo, ali so glede na čas zbiranja primerni za naš primer uporabe. Časovna veljavnost preverja, ali so podatki glede na njihov čas zajema ustrezni za nas. Preveriti moramo tudi, ali so bili podatki izmerjeni v ustreznem času (ali vrednosti podatkov res odražajo vrednosti iz navedenega časa), ki je naveden v metapodatkih podatkovnega vira. S tem preverjanjem dobimo delež podatkov, ki ne ustrezajo temu pogoju. Časovna konsistentnost (angl. temporal consistency) preverja, da so datumi, ki opisujejo neko zaporedje dogodkov, v pravem vrstnem redu, npr. datum konca nekega dogodka mora biti vedno kasneje od datuma začetka tega dogodka. Točnost merjenja časa (angl. accuracy of a time measurement) nam pove, kako točni so datumi v bazi, glede na dejanske datume, ko so se neki dogodki zgodili, npr. datumi konca izgradnje stavb v tabeli stavbe.
- Logična konsistentnost (angl. logical consistency): Konceptualna logična konsistentnost (angl. conceptual logical consitency) preverja upoštevanje pravil konceptualne podatkovne sheme, npr. imena razredov, imena atributov, nabor vrednosti, ki jih lahko neki atribut zaseda ipd. Domenska logična konsistentnost (angl. domain logical consistency) preverja zasedanje vrednosti iz nabora kategorij, ki je za neki atribut določen. Preverjanje domenske logične konsistentnosti lahko preverimo ločeno, ker pa konceptualna podatkovna shema vsebuje nabore podatkov, ki so za posamezne atribute na voljo, lahko to preverjanje izvedemo tudi v sklopu konceptualne logične konsistentnosti. Topološka logična konsistentnost (angl. topological logical consistency) preverja, da podatkovna baza ne vsebuje topoloških napak, npr. nepovezanih cest in odprtih poligonov. Logična konsistentnost formatov (angl. format logical consistency) pa preverja, da vrednosti v podatkovni bazi zasedajo pravilen format glede na pravila v podatkovni shemi.
- Uporabnost (angl. usability): Uporabnost podatkov predstavlja zunanjo kakovost podatkov. Ko govorimo o uporabnosti, velikokrat slišimo tudi izraz primernost za uporabo (angl. fitness for use), kar pomeni, da je odvisna od uporabnikovih želja. Elementi notranje kakovosti in uporabnost skupaj predstavljajo kakovost podatkov za določen primer uporabe.

Fonte in drugi (2017) pravijo, da za vse elemente notranje kakovosti potrebujemo sorodno podatkovno bazo, ki ima podobne specifikacije kot naša in ustreza časovnemu obdobju, v katero spadajo naši podatki. Tako lahko podatke primerjamo s pravilnimi. Standard ISO 19157:2013 je prikazan na sliki 2.

Slika 2: Elementi kakovosti prostorskih podatkov po standardu ISO 19157:2013



Prirejeno po Bindzárová Gergel'ová in drugi (2020).

3.2 Identifikacija napak in čiščenje prostorskih podatkov

V literaturi in spletnih virih so definirane različne napake, ki so značilne za prostorske podatke. Opazimo, da so nekatere sorodne oz. se lahko aplicirajo tudi na splošne podatke, npr. duplikati, druge pa so specifične za prostorske podatke. V nadaljevanju sem napake in morebitne načine za njihovo reševanje smiselno združil in opisal.

3.2.1 Manjkajoče koordinate

Če za vnose nimamo podanih koordinat, to predstavlja veliko težavo, saj teh entitet ne moremo prikazati na zemljevidu. Ali naša podatkovna baza vsebuje takšne entitete, lahko preverimo na preprost način, z iskanjem entitet, ki pri atributih koordinat ne hranijo nobene vrednosti (Spencer & Wilkes, 2019, str. 11).

Geokodiranje je proces, ki na podlagi danega naslova oz. drugih podatkov o lokaciji kot rezultat vrne koordinate te lokacije. Pri čiščenju podatkov ga lahko uporabimo pri vnosih, ki nimajo shranjenih koordinat ali pa so te napačne, da te izračunamo iz drugih podatkov, ki jih imamo za te vnose na voljo. Primer: Če imamo pri vnosu na voljo samo naslov, ne pa tudi koordinat, ki so potrebne za našo analizo, lahko z geokodiranjem izračunamo koordinate. Pri geokodiranju pa moramo biti pazljivi, saj nam ta proces lahko vrne tudi napačne koordinate, npr. zaradi podobnih oz. enakih imen entitet ali njihovih naslovov na zemlji. Zato moramo pridobljene koordinate ustrezno preveriti (Spatial Data Science, brez datuma).

3.2.2 Manjkajoči naslovi

Poleg manjkajočih koordinat se nam lahko zgodi podobna težava, pri kateri podatkovni vir ne vsebuje naslovov. V primeru, da naslove potrebujemo in imamo na voljo koordinate,

lahko naslove pridobimo s procesom nasprotnega geokodiranja. Nasprotno geokodiranje deluje tako, da za dane koordinate vrne naslov, ki je tem koordinatam najbližje (KAISPE, brez datuma).

3.2.3 Odstranjevanje nepotrebnih stolpcev in vrstic

V poglavju Kakovost podatkov sem opisal odstranjevanje stolpcev in vrstic, ki jih ne potrebujemo. Ta proces lahko apliciramo tudi na čiščenje prostorskih podatkov. Tudi spletna stran Mango v videu čiščenja prostorskih podatkov omeni ta 2 pristopa (Brown, brez datuma):

- odstranjevanje stolpcev, ki jih ne potrebujemo;
- odstranjevanje vrstic, ki jih ne potrebujemo (glede na vrednost katerega od atributov, recimo lokacije, če potrebujemo samo vnose za Slovenijo, lahko zbrišemo vnose ostalih držav).

Sheoran in Parmar (2022) opišeta delovanje ogrodja, ki je bil razvit z namenom čiščenja prostorskih podatkov in pri čiščenju med drugim tudi:

- odstrani vnose, ki so prazni (v prikazanem primeru predstavljenega ogrodja so vrednosti vseh atributov za izbrisan vnos prazne);
- odstrani atribute, ki jih ne potrebujemo ali pa so vrednosti za ta atribut prazne.

3.2.4 Premajhna natančnost koordinat

Premajhna natančnost je odvisna od števila decimalk, ki so za koordinati podane. Natančnosti ne moremo posplošiti na vse primere uporabe. Vedeti moramo, kakšna odstopanja so za naš primer še sprejemljiva (Spencer & Wilkes, 2019). Primer: Če želimo prikazati stavbe, so odstopanja do 5 metrov sprejemljiva, če pa želimo prikazati lokacije dreves, je takšno odstopanje preveliko.

3.2.5 Napačna lokacija

Podatki o lokaciji so lahko napačni zaradi različnih razlogov. Univerza v Kaliforniji kot primer našteje naslednje (UC Davis DataLab, brez datuma):

- programi za odpiranje datotek odrežejo oz. zaokrožijo decimalna mesta,
- zamenjani zemljepisna širina in dolžina (človeška napaka pri vnašanju),
- napačen rezultat geokodiranja.

Velikokrat se nam lahko zgodi, da so koordinate na nepričakovani lokaciji. Preveriti moramo, da se entiteta nahaja na pravi lokaciji. Primer: Če ima entiteta v enem od atributov

shranjeno vrednost, da se mora nahajati v Sloveniji, koordinate pa se nahajajo izven Slovenije, gre lahko za 1 od 2 težav. Lahko gre za napačne koordinate ali pa je napačen atribut, ki hrani ime države. Preveriti moramo, ali se koordinate ujemajo z atributi in v primeru neujemanj odkriti, za katero težavo gre (Spencer & Wilkes, 2019, str. 13).

Lahko se nam zgodi, da hrani podatkovna baza napačno vnesene koordinate. Primer: Če vse entitete pri zemljepisni širini zasedajo vrednost –79, 1 entiteta pa vrednost –97, se je najverjetneje zgodila napaka pri vnosu podatka in moramo števki pri napačni entiteti obrniti (Spencer & Wilkes, 2019, str. 11).

Na Spatial Data Science (brez datuma) so navedeni tudi naslednji kazatelji napačne lokacije:

- napačen predznak koordinat,
- prekopirana zemljepisna širina v zemljepisno dolžino (ali obratno),
- zemljepisna širina in/ali dolžina z vrednostjo 0 (lahko so se manjkajoče vrednosti NULL ob nalaganju v program pretvorile v 0).

Ker je format lokacij, ki so shranjene v podatkovni bazi, za človeka težko berljiv, pravilnost lokacij težko ugotovimo z branjem podatkov. Zato je najlažji način za odkritje napačnih lokacij ta, da jih prikažemo na zemljevidu in tam preverimo, če je katera od točk prikazana na napačni lokaciji (UC Davis DataLab, brez datuma). Prikaz z vizualizacijo je človeku prijazen način in tako lahko ugotovimo največ napak. Zato ni dovolj, da podatke preverimo samo s poizvedbami in pregledom v podatkovni bazi (Spatial Data Science, brez datuma). S prikazom na zemljevidu lahko takoj opazimo, ali se katera od entitet nahaja na nesmiselni lokaciji, npr. sredi morja ali v napačni državi. S tem pristopom pa vseeno ne ulovimo vseh napak. Primere, ko se entitete nahajajo na smiselnih, ampak nepravilnih lokacijah, npr. v napačni občini, ki pa vseeno predstavlja del podatkov, s tem pristopom ne opazimo. V slednjem primeru gre lahko za napačne koordinate ali pa napačno vrednost občine (Spencer & Wilkes, 2019).

Prikaz lahko podkrepimo tako, da zraven točk izpišemo tudi vrednost katerega od atributov, ki nam lahko pomaga odkriti napačno lokacijo. Primer: Če za vnose v tabeli hranimo tudi atribut »država«, lahko vrednost tega izpišemo in hitro opazimo, če se točka nahaja v napačni državi (UC Davis DataLab, brez datuma).

Chapman (2005) način preverjanja s prikazom razdeli na več tehnik:

- Preverjanje lokacije na zemljevidu s katerim drugim atributom vnosa, npr. državo. Če se lokacija točke na zemljevidu ne nahaja v državi, ki je navedena za ta vnos, je lokacija napačna. Pri tem načinu si lahko preverjanje olajšamo tako, da dodaten atribut izpišemo poleg točk na zemljevidu.
- Preverjanje ali je lokacija vnosa skladna z lokacijo, kjer so se podatki zbirali, npr. če podatke zbiramo v Sloveniji, morajo biti pri vizualizaciji vsi podatki prikazani v Sloveniji.

– Preverjanje, ali je točka na zemljevidu prikazana na pravi oz. smiselni lokaciji, npr. točka, ki predstavlja lokacijo na kopnem, se ne sme prikazati sredi morja.
– Preverjanje izstopajočih točk, ki se glede na lokacijo bistveno razlikujejo od preostalih.

Pravilnost lokacij lahko sicer preverimo tudi s poizvedbami v podatkovni bazi. Vzemimo primer, kjer za vse vnose pričakujemo, da se ti nahajajo v Sloveniji. S prikazom lahko takoj opazimo, ali se katera od točk nahaja izven Slovenije. V tem primeru bi uporabili funkcijo, če je ta vgrajena v program ali programski jezik, ki na podlagi podanih koordinat vrne ime države. Ta imena držav bi nato primerjali z imeni, ki jih že imamo v bazi in tako preverili neujemanja. Tako bi pridobili 2 vrsti napačnih lokacij (Spatial Data Science, brez datuma):

– lokacije, ki se ne nahajajo v nobeni državi (so v morju),
– lokacije, ki se nahajajo v napačni državi.

3.2.6 Napaka v ostalih atributih

Napačne vrednosti atributov prostorskih podatkov lahko ugotavljamo na enak način kot napačne vrednosti lokacije. Točke v več iteracijah prikazujemo na zemljevidu z atributi, ki jih potrebujemo. Tako lahko hitro opazimo, ali kakšna vrednost atributa manjka ali pa hrani nepričakovano vrednost (UC Davis DataLab, brez datuma).

3.2.7 Neveljavne koordinate

Ena od težav je, da katera od koordinat hrani vrednost izven dovoljenega obsega. Če zemljepisna širina in/ali zemljepisna dolžina hranita vrednosti izven njunega obsega, so njune vrednosti napačne (Spencer & Wilkes, 2019, str. 11). Zemljepisna širina lahko zaseda samo vrednosti med –90 in 90, zemljepisna dolžina pa vrednosti med –180 in 180 (GISGeography, brez datuma c).

Univerza v Kaliforniji (UC Davis DataLab, brez datuma) na svoji spletni strani navede napačen format koordinat. Zemljepisna širina in zemljepisna dolžina sta vrednosti, ki skupaj predstavljata lokacijo na zemljevidu. Težava, ki se pojavi pri formatu teh 2 koordinat, je, da računalnik razume samo format decimalnih stopinj. To pomeni, da moramo v primeru, ko so koordinate v bazi shranjene v formatu stopinje-minute-sekunde (angl. degrees-minutes-seconds, v nadaljevanju DMS), te pretvoriti v format decimalnih stopinj. Primer obeh formatov za lokacijo Ekonomske fakultete Univerze v Ljubljani je prikazan v tabeli 1.

Tabela 1: Lokacija Ekonomske fakultete Univerze v Ljubljani v različnih formatih

	Zemljepisna širina	**Zemljepisna dolžina**
Decimalne stopinje	46.07449314555481	14.516534831410016

Tabela 1: Lokacija Ekonomske fakultete Univerze v Ljubljani v različnih formatih (nad.)

	Zemljepisna širina	**Zemljepisna dolžina**
DMS	46° 4' 28.17532"	14° 30' 59.52539"

Vir: lastno delo.

Decimalne stopinje iz formata DMS pridobimo z enačbo (1).

$$decimalne\ stopinje = stopinje + \frac{minute}{60} + \frac{sekunde}{3600} \quad (1)$$

Pred uporabo podatkov moramo biti pozorni tudi na referenčni koordinatni sistem. Poznamo več sistemov, lokacije na zemlji pa so za različne sisteme drugačne. Če bi koordinate enega sistema želeli prikazati na drugem sistemu, ne da jih ustrezno pretvorimo, te ne bodo prikazane na pravi lokaciji. Podatke moramo ustrezno pretvoriti v referenčni sistem, ki ga bomo uporabljali za prikaz točk na zemljevidu. Primer referenčnega sistema je referenčni sistem evropske raziskovalne skupine za nafto (angl. European Petroleum Survey Group, v nadaljevanju EPSG) (UC Davis DataLab, brez datuma). Primera zemljepisne širine in dolžine Ekonomske fakultete Univerze v Ljubljani v 2 različnih koordinatnih sistemih sta prikazana v tabeli 2.

Tabela 2: Lokacija Ekonomske fakultete Univerze v Ljubljani v različnih referenčnih koordinatnih sistemih

	Zemljepisna širina	**Zemljepisna dolžina**
EPSG 3395 »World Mercator«	1615973.2655153824	5761506.500787223
EPSG 4326 »WGS 84«	46.07449314555481	14.516534831410016

Vir: lastno delo.

3.2.8 Duplikati

Težavo z duplikati sem opisal že v poglavju Kakovost podatkov. To težavo lahko apliciramo tudi na prostorske podatke.

Podvojeni vnosi v tabeli podatkovne baze so težava, ki se največkrat pojavi zaradi združevanja podatkov iz različnih virov. Takšne vnose moramo poiskati in odstraniti vse duplikate, da za 1 entiteto v resničnem svetu ostane samo 1 vnos (Spatial Data Science, brez datuma).

Težava, ki se lahko pojavi, so vnosi z enakimi koordinatami pri unikatnih lokacijah. Če imata 2 ali več vnosov v podatkovni bazi enake koordinate, gre lahko za napako. Če so ti vnosi na različnih lokacijah, to predstavlja težavo v podatkovni bazi. Na drugi strani pa lahko ne gre za napako in imata 2 različni entiteti enako lokacijo na zemlji, npr. 2 različni podjetji, ki imata sedež v isti stavbi (Spencer & Wilkes, 2019, str. 12).

Lahko se nam pojavi tudi težava s podvojenimi primarnimi ključi. Entitete v podatkovni bazi morajo imeti unikatno vrednost, ki jih loči od ostalih entitet. Ta vrednost je največkrat neki identifikator (ID), lahko pa je tudi kakšen drug atribut, npr. ime, naslov ipd. Če imata 2 entiteti enako vrednost, ki bi morala biti unikatna, moramo preveriti, za kakšno težavo gre. Lahko je vnos podvojen, lahko pa imata entiteti enako vrednost pri nekem atributu ali pa se nahajata na isti lokaciji (Spencer & Wilkes, 2019). Tudi Sheoran in Parmar (2022) omenjata, da moramo pri čiščenju odstraniti podvojene vrednosti glede na primarni ključ (če imata 2 vnosa enak primarni ključ, to pomeni, da predstavljata isto entiteto v resničnem svetu).

Iskanje duplikatov lahko rešujemo s postopkom razreševanja prostorskih entitet. To je proces, s katerim rešujemo 2 težavi (Kang, Sehgal & Getoor, 2007):

- Podvajanje v primeru uporabe 1 podatkovnega vira: vnose, ki predstavljajo isto entiteto na zemlji, združimo v 1 vnos in se s tem znebimo podvojenih vrednosti.
- Podvajanje v primeru integracije več podatkovnih virov: vnose iz različnih virov ustrezno povežemo, da v končni skupni bazi predstavljajo isto entiteto.

V obeh primerih želimo preprečiti podvajanje, zato želimo vse vnose v neki bazi združiti tako, da na koncu vsak od vnosov predstavlja svojo entiteto na zemlji. Recimo, da imamo 2 podatkovni bazi s prostorskimi podatki (angl. geospatial datasets): bazo A in bazo B. Pri iskanju duplikatov v posamezni bazi moramo poiskati vse pare v bazi, ki predstavljajo isto entiteto na zemlji. Takšne pare označimo z $\{l_i, l_j\}$, kjer sta l_i in l_j 2 vnosa v bazi, ki predstavljata isto entiteto na zemlji. l_i je vnos, ki ga sestavljajo njegovi atributi. Primer: Vnos v neki bazi bi lahko bil sestavljen iz koordinat, dodatnih podatkov, ki pripadajo tej lokaciji, npr. število prebivalcev, imena lokacije ipd. Ko iščemo duplikate znotraj posamezne baze, iščemo pare, ki ustrezajo pogoju, prikazanem z enačbo (2) (Kang, Sehgal & Getoor, 2007).

$$(l_i, l_j \in A \; ali \; l_i, l_j \in B) \tag{2}$$

Do težave lahko pride tudi pri integraciji več različnih podatkovnih virov. V tem primeru mora ustrezati pogoj, prikazan z enačbo (3) (Kang, Sehgal & Getoor, 2007).

$$((l_i \in A \; in \; l_j \in B) \; ali \; (l_i \in B \; in \; l_j \in A)) \tag{3}$$

Slednji pogoj predstavlja situacijo, kjer se 2 vnosa, ki predstavljata isto lokacijo na zemlji, nahajata v 2 različnih bazah. Pri čiščenju moramo biti pozorni ne samo na duplikate v 1 bazi, ampak tudi na to, da ne ustvarimo duplikatov z integriranjem več baz (Kang, Sehgal & Getoor, 2007).

Poznamo več pristopov za iskanje duplikatov oz. enakih vnosov, v osnovi pa se pristopi delijo na ročne in avtomatizirane. Avtomatizirane rešitve same poiščejo morebitne sorodne vnose bodisi v 1 podatkovni bazi bodisi v več različnih, pri ročnem pregledu pa moramo to delo opraviti sami. Avtomatizirane rešitve sorodne vnose iščejo na različne načine, kot sta npr. iskanje lokacij, ki so si na zemljevidu zelo blizu (pod nekim pragom), in vnosov, ki

imajo enake oz. podobne vrednosti ostalih atributov. Na kakšen način bomo vnose ločili, je odvisno od tega, kakšen problem s podatki rešujemo. V primeru iskanja duplikatov s pristopom iskanja lokacij, ki so si blizu, določimo neki prag, npr. 10 km, in poiščemo pare, ki so med seboj oddaljeni 10 km ali manj. Na koncu pa moramo sami ovrednotiti, ali vnosa v bazi res predstavljata isto entiteto (Kang, Sehgal & Getoor, 2007).

4 KAKOVOST VIZUALIZACIJ

Vizualizacija podatkov je vsak grafični prikaz, s katerim sporočamo podatke, ne glede na to, v kateri disciplini je uporabljen (Few, 2009). Je zelo uporaben način sporočanja podatkov, saj ljudje veliko lažje razumemo grafični prikaz kot pa navadne številke in besede (Cukier, 2010). Z vizualizacijami lahko velike količine neberljivih podatkov prikažemo na človeku prijazen in berljiv način (Ware, 2019), biti pa moramo pozorni, da vizualizacije končnim uporabnikom ne sporočajo napačnih informacij in da niso dvoumne (Szafir, 2018).

Od kakovosti vizualizacij je odvisno, kako so podatki predstavljeni in ali ti končnim uporabnikom sporočajo pravilne informacije. Slaba vizualizacija lahko sporoča napačne informacije in je težje berljiva, kar vodi v to, da si jo končni uporabniki narobe interpretirajo. S kakovostno vizualizacijo postanejo tudi informacije bolj jasne, uporabnikove odločitve pa bolj kakovostne (Wilke, 2019).

Poznamo več vrst vizualizacij. Pomembno je, da uporabimo tisto, ki je primerna za naš primer uporabe. Nekatere od najbolj uporabljenih so (Sadiku, Shadare, Musa, Akujuobi & Perry, 2016, str. 12):

- Črtni diagram: Primer uporabe je prikaz spreminjanja vrednosti v različnih časovnih obdobjih.
- Stolpčni diagram: Uporabimo ga, ko želimo primerjati količino neke vrednosti med različnimi kategorijami.
- Tortni diagram: Prikazuje, kolikšen delež zavzema posamezen del celote.
- Razsevni diagram: Gre za prikaz podatkov, ki imajo 2 dimenziji. Pove nam, kako so podatki med seboj povezani.

Obstaja veliko smernic z upoštevanjem katerih lahko izboljšamo kakovost vizualizacij, kot so uporaba pravilne oblike in formata, izogibanje uporabi nepotrebnih elementov, uporaba besed, števil in slik skupaj, prikaz podrobnosti do ravni, ki je še razumljiva, predstavitev podatkov z zgodbo ter profesionalna izdelava vizualizacij, kjer poskrbimo tudi za detajle oz. se pri izdelavi potrudimo (Tufte, 2001). V naslednjih poglavjih so opisane splošne smernice, v poglavju Kakovost vizualizacij z zemljevidi pa smernice za vizualizacijo prostorskih podatkov.

4.1 Predhoden razmislek

Izredno pomembno je, da pred začetkom kreiranja vizualizacije določimo, katere informacije želimo z našo vizualizacijo predstaviti oz. kakšno sporočilo želimo deliti s končnimi uporabniki. Ne smemo kar takoj začeti z izdelavo, saj lahko brez predhodnega razmisleka izberemo programsko opremo ali pa vrste vizualizacij, ki nas bodo omejevale ali pa celo vplivale na kakovost sporočenih informacij. Priporočljivo je, da predhodno določimo, kaj uporabnika zanima in kaj ter na kakšen način mu želimo to sporočiti, npr. s primerjavo vrednosti v različnih obdobjih. Ko imamo navedene stvari določene, pa lahko na njihovi podlagi začnemo razmišljati o samih vrstah vizualizacij (Midway, 2020).

Priporočljivo je, da pred začetkom kreiranja vizualizacije razumemo, kakšen je njen namen. Samo tako lahko zagotovimo, da bomo res izdelali vizualizacijo, ki bo reševala težavo končnega uporabnika. Poskrbeti moramo, da naša vizualizacija res odgovarja na vprašanja, ki si jih je končni uporabnik zastavil (Ali, Gupta, Nayak & Lenka, 2016).

Pred začetkom izdelave vizualizacij naši možgani pogosto sami povežejo določeno vrsto podatkov s sorodno geometrijo, npr. primerjava vrednosti, porazdelitev, deleži itd. Vendar pa se ne smemo omejiti samo na 1 vrsto geometrije, ampak imeti pri izbiri v mislih vse. Različne geometrije podatke prikazujejo na različne načine in lahko se nam zgodi, da moramo prikazati podatke na več načinov. Primer možne uporabe različnih geometrij je prikaz deležev različnih kategorij. Tukaj je velikokrat uporabljen tortni diagram, vendar pa lahko uporabimo tudi druge geometrije, kot je npr. stolpčni diagram (angl. stacked bar plot) (Midway, 2020, str. 2).

Pomembna je izbira dobrega programa za izdelavo vizualizacij. Preprosti programi, ki niso dovolj izpopolnjeni, nas lahko omejujejo in nam z njimi ne uspe izdelati tako dobrih vizualizacij. Obstaja veliko brezplačnih programov, ki so temu namenjeni, vendar je za njihovo uporabo potrebno določeno znanje za uporabo programa kot takšno, pa tudi zato, da znamo s programom izdelati kakovostne vizualizacije (Midway, 2020, str. 2).

4.2 Barve

Uporaba barv je v današnjem času pri vizualizacijah zelo pomembna, saj predstavlja način sporočanja informacij. Primer uporabe barv je pri zemljevidih, kjer uporabnik lahko takoj vidi, kje je morje (modra barva) in kje zemlja (zelena barva) (Midway, 2020, str. 4).

V osnovi barve predstavljamo s 3 različnimi barvnimi paletami (Midway, 2020, str. 4):

- Kvalitativna (angl. qualitative): Uporaba popolnoma različnih barv za prikaz različnih kategorij.
- Zaporedna (angl. sequential): Uporaba ene ali več barv, kjer se glede na količino podatkov spreminja njihov odtenek.

– Razhajajoča (angl. divergent): Uporaba 2 zaporednih barvnih shem, kjer vsaka od shem pripada 1 ekstremu. Zelo pogosto uporabljeni barvi sta rdeča in modra, kjer vsaka barva predstavlja 1 kategorijo. Temnejša, kot je katera od barv, večja je količina podatkov, ki tej barvi pripada.

Ko barve predstavljamo s kvalitativnim načinom ali pa ko vizualizacijo sestavlja več različnih geometrij, ne smemo posameznih barv preveč poudarjati, da te ne zasenčijo ostalih elementov in si uporabnik narobe interpretira, da je kateri izmed elementov bolj pomemben od ostalih. Barva je del vizualizacije, ki lahko zelo vpliva na razumevanje uporabnika, saj je to ena od prvih stvari, ki je opažena. Zato je priporočljivo, da barve prikazujemo precej transparentno (Midway, 2020). Barvne palete so prikazane na sliki 3. Podnaslovi niso prevedeni v slovenščino, saj ti predstavljajo imena posameznih barvnih palet.

Slika 3: Barvne palete

Prirejeno po Krzywinski, Birol, Jones & Marra (2012).

Ena od smernic je, da pri vizualizacijah uporabimo takšne barve, ki jih razumejo tudi barvno slepi ljudje. Tako so naše vizualizacije berljive za širše občinstvo, prav tako pa si lahko tudi barvno slepi ljudje pravilno interpretirajo vizualizacije (Midway, 2020, str. 4). Obstajajo tudi orodja, s pomočjo katerih lahko vidimo, kako so naše vizualizacije vidne barvno slepim ljudem (Duke University Libraries, 2022). Na podlagi rezultatov teh orodij lahko barve v naših vizualizacijah ustrezno popravimo.

Ko prikazujemo različne kategorije, jih uporabniki najlažje ločijo, če uporabimo popolnoma drugačne barve. Te se med seboj bolj razlikujejo kot pa različni odtenki iste barve. Na drugi strani pa moramo pri prikazu nekega razpona uporabiti isto barvo in glede na vrednosti izbrati ustrezen odtenek. Če pri slednjem primeru uporabimo različne barve, uporabniki ne bodo nujno vedeli, katera barva predstavlja najvišjo in katera najnižjo vrednost, ali pa jih bodo prebrali napačno (Duke University Libraries, 2022). Različne odtenke iste barve (zaporedna barvna paleta) lažje razvrstimo od najsvetlejše do najtemnejše kot pa različne barve (kvalitativna barvna paleta).

Pri vizualizacijah se velikokrat uporablja kombinacija 4 barv: modre, zelene, rumene in rdeče (angl. rainbow colormap). Vseeno pa veliko študij ne priporoča uporabe te barvne kombinacije pri zveznih prikazih, saj naj bi ta v veliko primerih uporabnike zavajala in vodila do napačnih zaključkov. Razlog za to je ravno uporaba 4 različnih barv. Uporaba različnih barv je sicer dobra, ko prikazujemo različne kategorije, ni pa priporočljiva pri zveznih prikazih in prikazih, kjer vrednosti razvrščamo. Ta težava se nam lahko pojavi tudi pri uporabi toplotnih zemljevidov. Ena od težav je, da so si nekatere od barv bolj podobne kot druge. Recimo, da imamo na toplotnem zemljevidu po vrsti razvrščene rdečo, rumeno, zeleno in modro barvo. Rumena barva se na prvi pogled zdi bližje oranžni barvi kot pa zeleni, čeprav je od obeh enako oddaljena. Primer je tudi modra barva, kjer naši možgani vse odtenke modre povežejo med seboj, kot da so si ti zelo blizu, čeprav je najsvetlejši odtenek modre lahko bližje zeleni kot pa najtemnejšemu odtenku modre. Uporaba različnih barv ustvari nekakšne navidezne meje, ki uporabnika lahko zavedejo, saj posamezne barve poveže med seboj (Szafir, 2018, str. 29).

4.3 Razmerje med podatkovnim in celotnim črnilom

Tufte (2001) govori o razmerju med podatkovnim in celotnim črnilom (angl. data-ink ratio), kjer to razmerje predstavlja razmerje med črnilom, ki uporabniku podaja uporabno informacijo, ter vsem črnilom na vizualizaciji. Doseči želimo čim večje razmerje, se znebiti črnila, ki ne pripomore k razumevanju vizualizacije oz. uporabniku ne podaja neke nove informacije (angl. chartjunk). Vizualizacija mora biti preprosta in vsebovati samo elemente, ki nam pomagajo reševati našo težavo.

Midway (2020, str. 5) se strinja s tem, da je dobro doseči čim višje razmerje med podatkovnim in celotnim črnilom, vendar pa moramo na vizualizacijo vseeno dodati dovolj oznak, da so vizualizacije razumljive oz. končni uporabniki razumejo, kaj predstavljajo. Poskusiti moramo z oznakami narediti vizualizacijo čim bolj razumljivo samo po sebi (angl. self-explanatory), da uporabnik ob pogledu na vizualizacijo hitro ugotovi, kaj ta predstavlja. Pri nekaterih vizualizacijah nam to lahko uspe, pri zahtevnejših, pa je potrebna tudi dodatna razlaga.

Vizualizacije, ki so na prvi pogled zelo privlačne in informativne, ne nujno pripomorejo k sporočanju prave informacije (Szafir, 2018). Paziti moramo, da v vizualizacijo ne dodamo preveč informacij. Prevelika količina informacij končnega uporabnika lahko zmede, s čimer pa dosežemo ravno obratni učinek od želenega (Kulyk, Kosara, Urquiza & Wassink, 2007).

4.4 Transparentnost

Pri vizualizacijah si želimo, da so vizualizacije transparentne, da uporabnika ne zavajajo in da sporočajo pravo informacijo. Zato je dobro, da končnemu uporabniku omogočimo vpogled v procese, ki smo jih nad podatki izvajali, da lahko preveri, ali so bili procesi

kakovostno izvedeni in tako sam preveri ali vizualizacije res odražajo dejansko stanje (Kirk, 2016).

4.5 Animacije

Animacije se pri vizualizacijah podatkov največkrat uporabljajo, ko želimo prikazati spremembe vrednosti skozi čas. Gre za učinek, ki naredi vizualizacijo privlačnejšo. Vseeno pa moramo biti ob njihovi uporabi pazljivi, da se na vizualizaciji ne odvija preveč stvari hkrati – ljudje lahko dojemamo samo določeno količino premikajočih oz. spreminjajočih se objektov. Prezahtevne oz. preobsežne animacije lahko uporabniku onemogočijo, da si zna v celoti predstavljati sporočilo vizualizacije. Razlog za to je ta, ker se osredotočamo samo na določen del animacij, ostale pa spregledamo (Szafir, 2018, str. 29).

Szafir (2018, str. 30) kot rešitev zgornje težave opiše 3 možne alternative, ki ne zahtevajo uporabe animacij. To so (Szafir, 2018, str. 30; Midway, 2020):

- Prikaz več vizualizacij za različne čase (angl. juxtaposition): Če delamo primerjavo med različnimi podatki, lahko to storimo tako, da enako vizualizacijo prikažemo večkrat – za vsake podatke enkrat. Ta način se je pokazal kot uspešen in razumljiv. Primer je prikaz neke vrednosti v različnih obdobjih. Tukaj lahko vizualizacijo dupliciramo tako, da imamo za vsako obdobje 1 vizualizacijo, npr. 4 vizualizacije za 4 letne čase. Vsaka vizualizacija ima iste specifikacije (ista vrsta, iste enote na oseh, isti razmiki na oseh ipd.), vsaka pa prikazuje podatke za svoj letni čas. Ker so vizualizacije enake, lahko uporabnik zelo hitro in učinkovito primerja vrednosti med seboj.
- Prikaz vrednosti ob različnih časih na isti vizualizaciji (angl. superposition): Pri tem načinu uporabljamo samo 1 vizualizacijo. Tu lahko vsako od kategorij predstavlja 1 barva, ki se ji glede na čas spreminja odtenek. Vsaki točki na vizualizaciji glede na kategorijo pripada barva. Če želimo za to kategorijo prikazati vrednost v 3 različnih obdobjih, prikažemo 3 točke, ki si sledijo kronološko od svetle proti temni. Uporaba tega načina je smiselna, ko želimo prikazati manjše številko točk. V primerih, ko želimo prikazati veliko število točk, uporaba tega načina ni najbolj ustrezna, saj se točke lahko med seboj prekrivajo in vizualizacija postane nepregledna.
- Neposreden prikaz sprememb na 1 vizualizaciji (angl. explicit encoding): Pri tem načinu za vsako od kategorij prikažemo 1 črto, kjer začetek črte prikazuje kronološko najstarejšo vrednost, njene vmesne točke vmesne vrednosti, njen konec pa najnovejšo vrednost.

4.6 Interaktivnost

Ena od funkcionalnosti, ki jo orodja za vizualizacije omogočajo, je uporaba interaktivnosti v vizualizacijah. Z interaktivnostjo se lahko uporabnik premika po želenih ravneh, vizualizacijo gleda z želenim zumiranjem ipd., in se tako osredotoči na informacije, ki ga res zanimajo (Few & Edge, 2007).

4.7 Koordinatne osi

Ena od praks, ki se jo je dobro držati, je, da pri številskih vrednostih koordinatne osi začenjamo z vrednostjo nič. Če začnemo koordinatne osi z vrednostmi, višjimi od nič, lahko vizualizacija sporoča napačno informacijo, saj so v slednjem primeru razlike med vrednostmi precej bolj očitne oz. izgledajo večje, kot so v resnici. Tega se je dobro držati predvsem pri stolpčnih diagramih (Duke University Libraries, 2022), vendar lahko v določenih primerih za osi uporabimo tudi poljuben razpon vrednosti, npr. ko želimo izpostaviti samo določen nabor. Vseeno pa moramo biti pozorni, da s tem ne vplivamo na resničnost informacij (Kelleher & Wagener, 2011).

Vizualizacija sporoča napačno informacijo tudi, ko pri koordinatnih oseh s številskimi vrednostmi, katero od vrednosti izpustimo. Zato moramo vse vrednosti na številski osi vključiti, čeprav katera od njih ne vsebuje podatkov. Prav tako moramo paziti, da vrednosti rastejo z enakomernim razmikom, npr. če želimo predstaviti število neke pojavitve za vsak dan v mesecu marcu, moramo na koordinatni osi, ki vsebuje datume, vključiti vseh 31 dni (razen v posebnih primerih) (Duke University Libraries, 2022).

4.8 Izogibanje tridimenzionalnim vizualizacijam

Več študij omenja, da se je pri vizualizacijah treba izogibati 3D-učinka. Ta povzroči manjše razumevanje in težjo primerjavo med različnimi vrednostmi (Duke University Libraries, 2022).

Szafir (2018, str. 31) našteje več težav, ki se nam lahko pojavijo pri prikazu 3D-vizualizacij z 2D-načinom, npr. na papirju. Primer je okluzija (angl. occlusion), kjer se nekateri od 3D-elementov med seboj prekrivajo, npr. če je višji stolpec prikazan pred nižjim stolpcem, potem nižjega ne vidimo. Ker gre za 2D-prikaz, vizualizacije ne moremo obračati, da bi stolpce, ki so skriti, videli. Težava je tudi v tem, da je prikaz 3D-vizualizacij v 2D težko razumljiv, ker s tem izgubimo nekatere informacije, ki nam omogočajo njihovo razumevanje v prostoru. Ena od težav 3D-prikaza je nagib vizualizacije. Vzemimo primer tortnega diagrama. Če ga zarotiramo horizontalno, so nekateri deli od nas bolj oddaljeni, nekateri pa so nam bližje. Posledično so deli, ki so bolj oddaljeni, manjši, bližji deli pa večji – iz česar lahko pridemo do napačnega sklepa, da so tudi deleži manjši in večji.

Zato je priporočljivo, da 3D-prikaza ne uporabimo, če to ni res nujno. Namesto tega pa lahko uporabimo alternative, kot so različne velikosti elementov ali različne barve. Če se vseeno zaradi potrebe po uporabi 3D-vizualizacije zanje odločimo, jih lahko vseeno podpremo z dodatnimi 2D-vizualizacijami, ki razrešijo kakršne koli dvoumnosti (Szafir, 2018, str. 32).

4.9 Posvetovanje z drugimi ljudmi

Z upoštevanjem smernic za kakovostne vizualizacije veliko pripomoremo k temu, da so te bolj razumljive in primerne. Vseeno pa je priporočljivo, da se pred objavo oz. predstavitvijo vizualizacij, posvetujemo z drugimi ljudmi, saj z njihove strani dobimo objektivno mnenje o vizualizacijah, tj. razumevanje, berljivost, kaj bi spremenili ipd. Navsezadnje delamo vizualizacije za ljudi in s tem pristopom lahko potrdimo, ali so te res dovolj dobre ter jih na podlagi povratnih informacij ustrezno popravimo (Midway, 2020, str. 5).

5 KAKOVOST VIZUALIZACIJ Z ZEMLJEVIDI

Tako kot pri splošnih vizualizacijah imamo tudi pri vizualizacijah z zemljevidi na voljo smernice, ki nam pomagajo izdelati dobre in informativne vizualizacije (Dougherty & Ilyankou, 2021). V nadaljevanju so te smernice opisane. Zaradi sledenju enotnega sloga so uporabljene slike istega avtorja.

5.1 Izbira pravega zemljevida

Pred začetkom izdelave vizualizacije se moramo odločiti, katero bomo glede na primer uporabe uporabili. Za prikaz prostorskih podatkov bomo, npr. v večini primerov uporabili zemljevide, za prikaz preostalih podatkov pa različne grafe s koordinatnimi osmi (Chaudhri, 2021).

Pred uporabo zemljevida se moramo vprašati, ali je ta res najbolj primeren za naš primer uporabe. V nekaterih primerih uporaba zemljevida ni potrebna, čeprav hranimo prostorske podatke in je uporaba druge vizualizacije celo boljša. Primerjavo neke vrednosti med državami lahko prikažemo tudi s stolpčnim diagramom. Ker izdelava vizualizacije z zemljevidom lahko zahteva kar nekaj truda in časa, se je pred njeno uporabo treba vprašati, ali ta res prispeva dodano vrednost (Dougherty & Ilyankou, 2021).

Ob uporabi zemljevida pa moramo paziti tudi na to, katerega od zemljevidov bomo uporabili. Izbira zemljevida je odvisna od tega, kaj želimo z vizualizacijo prikazati in sporočiti ter od tipa prostorskih podatkov. Poznamo različne tipe prostorskih podatkov (Dougherty & Ilyankou, 2021):

- Točke (angl. points): Gre za predstavitev točne lokacije na zemlji, kjer vsako lokacijo predstavlja par koordinat.
- Črte (angl. polylines): To so povezane točke. Z njimi lahko predstavljamo, npr. ceste ali infrastrukturne povezave.
- Poligoni (angl. polygons): Sestavljeni so iz sklenjenih črt, npr. stavbe ali države.

Vprašati se moramo, katerega od zgornjih tipov naši podatki zasedajo in tudi na podlagi tega izbrati ustrezen zemljevid. Najbolj priljubljena tipa pri prikazu podatkov sta točke in

poligoni, kjer lahko npr. pri prvem zemljevidu prikažemo 1 točko za vsako lokacijo, pri drugem pa uporabimo horoplet zemljevid. Vendar pa ne smemo podatkov vedno kar tako umestiti k tema 2 zemljevidoma, ampak moramo upoštevati vse (Dougherty & Ilyankou, 2021).

Prav tako ne smemo zaradi tipa podatkov omejiti opcij, na kakšen način jih bomo prikazali. Točke lahko pretvorimo tudi v podatke, ki jih lahko prikažemo kot poligone. Vzemimo primer, da v podatkovni bazi hranimo točke, ki predstavljajo bolnišnice v Združenih državah Amerike. Vsaka točka je predstavljena s podatki o naslovu, mestu in državi. Če želimo prikazati, kje se bolnišnice nahajajo, je najbolj smiseln prikaz s točkami. Če pa želimo primerjati število bolnišnic med različnimi državami, je za to bolj primeren horoplet zemljevid. Za prikaz s horoplet zemljevidom moramo podatke ustrezno preurediti oz. dopolniti. V tem primeru bi za vsako od držav prešteli število bolnišnic in potem ta podatek uporabili pri prikazu (Dougherty & Ilyankou, 2021). Vrste zemljevidov in njihovi primeri uporabe so opisani v poglavju Vrste zemljevidov.

5.2 Izbira ustrezne ravni pri horoplet zemljevidih

Ko prikazujemo podatke s horoplet zemljevidi, se moramo pred tem odločiti, kako podrobno bomo podatke prikazovali. Lahko primerjamo države, občine, mesta ipd. To je odvisno od tega, kaj želimo prikazati, vseeno pa moramo biti pozorni, saj pri uporabi večjih poligonov ti ne prikazujejo, kakšno je stanje znotraj njih. Dougherty in Ilyankou (2021) opišeta primer, kjer so na enem zemljevidu prikazane cene domov v posameznih državah, na drugem pa v posameznih okrajih Združenih držav Amerike. Če želimo prikazati primerjavo med državami, je bolj primeren prvi zemljevid, ki odraža resnično stanje. Če pa nas zanimajo podatki bolj podrobno, je vsekakor bolje izbrati drugi zemljevid. To pride še posebej v poštev, če so vrednosti med posameznimi okraji različne. V tem primeru imajo nekatere države cene precej višje v obalnem delu kot pa v notranjosti. Ta informacija na prvem zemljevidu ni vidna. Pred prikazom se moramo odločiti, na kateri ravni bomo podatke prikazali. Če nismo popolnoma prepričani, je bolje vzeti podrobnejšo raven, saj s tem izpustimo manj informacij. Paziti pa moramo, da podatkov ne prikažemo preveč podrobno in s tem zmanjšamo berljivost zemljevida.

5.3 Ne prikazujmo preveč informacij naenkrat

V poglavju Razmerje med podatkovnim in celotnim črnilom sem omenil, da lahko preveč informacij pri vizualizacijah zmanjšuje berljivost. To velja tudi pri zemljevidih. Pri zemljevidih se je priporočljivo držati tega, da na zemljevidu ne prikazujemo različnih spremenljivk z različnimi elementi, npr. barva poligona za 1 spremenljivko in velikost simbola za drugo spremenljivko. S tem lahko prikazujemo preveč grafičnih elementov in informacij ter s tem otežujemo branje. Zato lahko za primerjavo različnih spremenljivk

uporabimo kakšen drug pristop, kot sta uporaba kakšne druge vizualizacije, npr. razsevni grafikon, ali pa uporabimo več vizualizacij (Dougherty & Ilyankou, 2021).

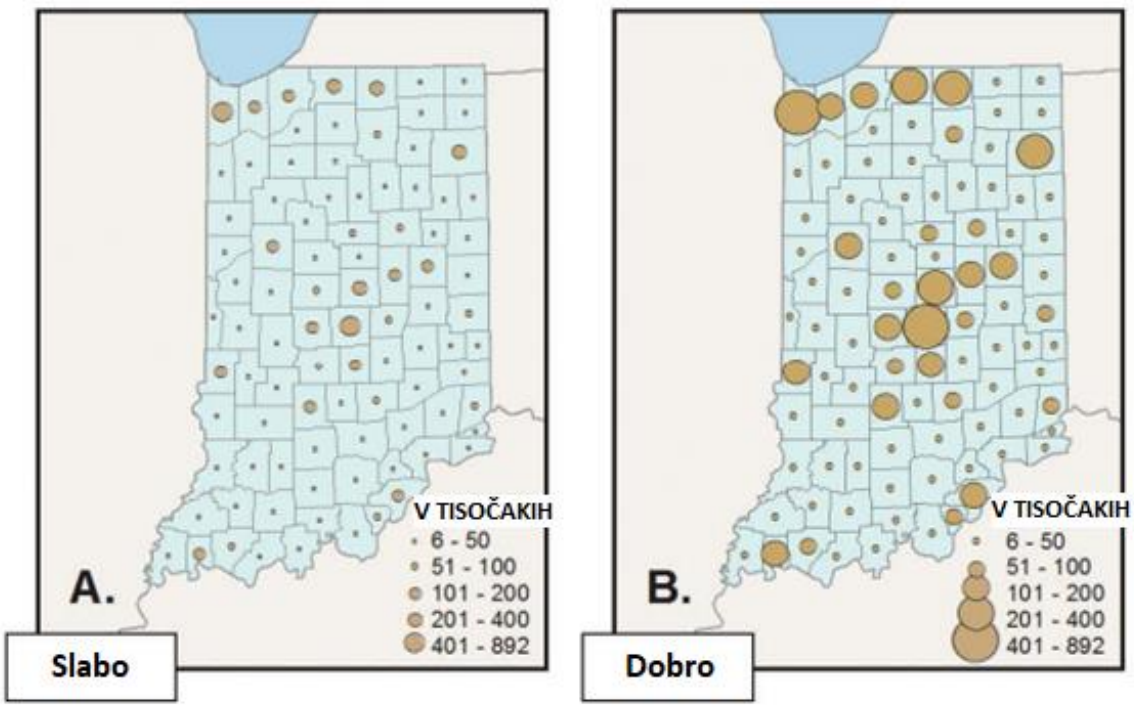
5.4 Normalizacija podatkov

Pred prikazom podatkov se moramo vprašati, ali jih je treba normalizirati. Včasih je ta korak potreben, če želimo uspešno primerjati vrednosti na različnih lokacijah. Recimo, da želimo s horoplet zemljevidom primerjati število okuženih ljudi s covidom-19 med različnimi državami. Če bomo med državami primerjali število okuženih ljudi, nam ta zemljevid ne bo povedal veliko. Države imajo različno število prebivalcev in je za večje države logično, da imajo večje število okuženih od manjših držav. Zato je smiselna predhodna normalizacija, ki jo za ta primer lahko dosežemo tako, da pri vsaki državi število okuženih ljudi delimo s številom prebivalcev in na zemljevidu prikazujemo ta delež (Dougherty & Ilyankou, 2021).

5.5 Izbira in velikost simbolov

Pri izbiri simbolov, ki jih bomo prikazovali na zemljevidu, moramo paziti na to, da so primerni in razumljivi ter dovolj veliki, da uporabniki s čim manjšim naporom razločijo, kolikšno vrednost predstavljajo. Na sliki 4 lahko na levem zemljevidu vidimo slabo prikazane simbole, saj so premajhni in zato težko berljivi, na desnem zemljevidu pa lahko vidimo bolje prikazane simbole, ki so dovolj veliki in zato bolj berljivi (Buckley, 2012).

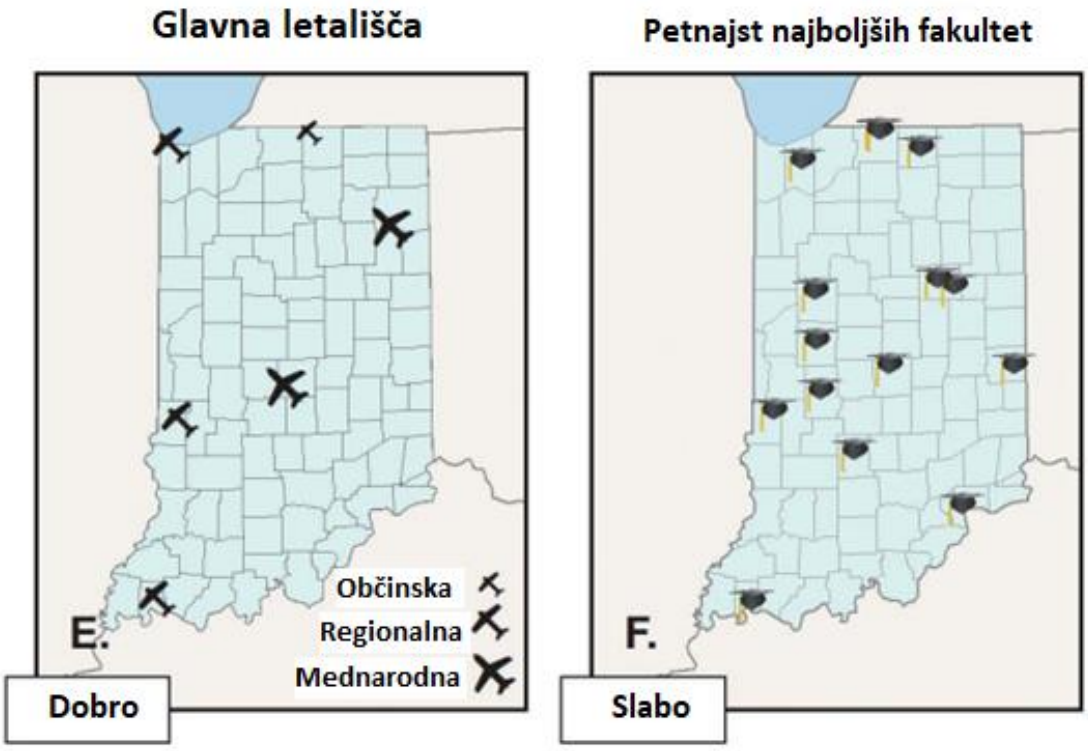
Slika 4: Primer premajhnih in dovolj velikih simbolov



Prirejeno po Buckley (2012).

Na sliki 5 lahko vidimo, da so na levem zemljevidu uporabljeni simboli letal, ki so precej preprosti in uporabnikom omogočajo, da jih takoj razumejo. Na desnem zemljevidu pa imamo simbole, ki so bolj kompleksni in zato uporabniki ne dojamejo tako hitro, kaj predstavljajo, povrh vsega pa so tudi premajhni. Zato je pri uporabi kompleksnejših simbolov potrebno, da so dovolj veliki, da uporabniki lahko razumejo, kaj predstavljajo (Buckley, 2012).

Slika 5: Primer preprostejših in kompleksnejših simbolov



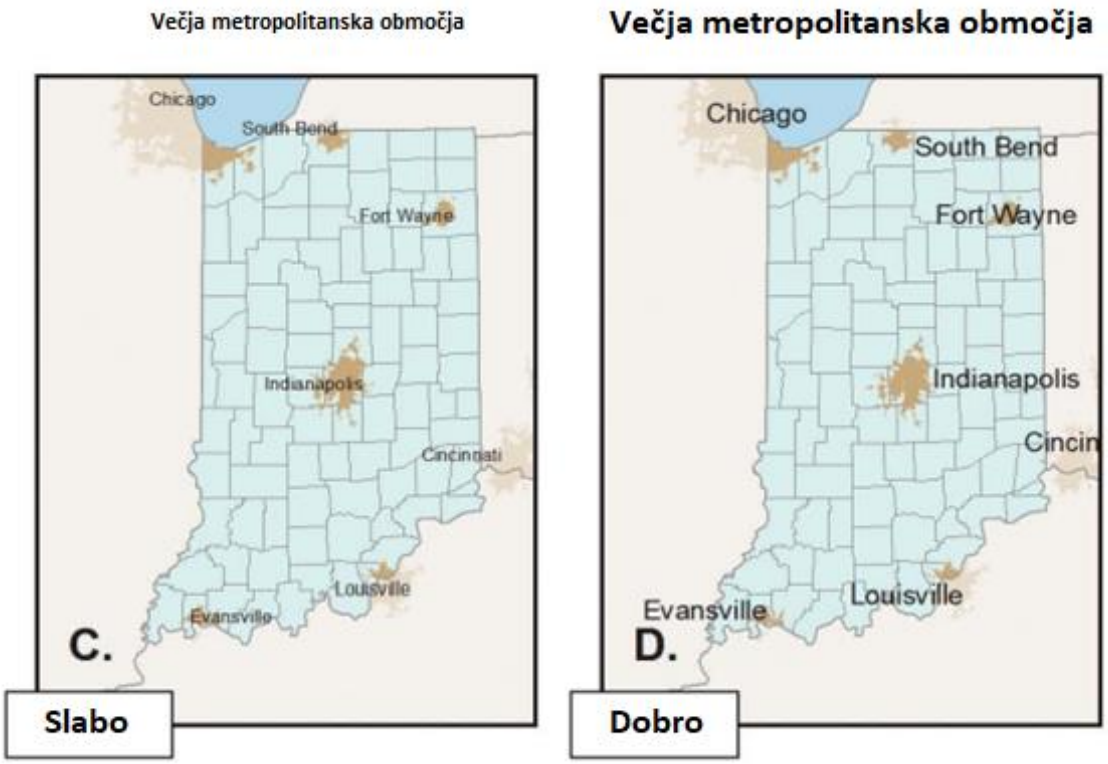
Prirejeno po Buckley (2012).

5.6 Uporaba in velikost besedila

Na zemljevide moramo včasih zaradi boljše razumljivosti dodati tudi različna besedila, npr. imena držav. Manj kot bralci poznajo območja, ki jih prikazujemo, bolj pomembno je, da na zemljevidu navedemo, za katera območja gre (Datawrapper, 2021).

Prav tako kot na velikost simbolov moramo biti pozorni tudi na velikost besedila. Na sliki 6 lahko vidimo, da je besedilo na levem zemljevidu zelo majhno, zaradi česar je tudi težko berljivo. Vidimo lahko, da je premajhno tako besedilo na zemljevidu, ki opisuje posamezne elemente, kot tudi naslov vizualizacije. Besedilo mora biti dovolj veliko, da ga lahko uporabniki preberejo brez napora in tako hitreje razberejo sporočilo vizualizacije (Buckley, 2012).

Slika 6: Primer premajhnega in dovolj velikega besedila

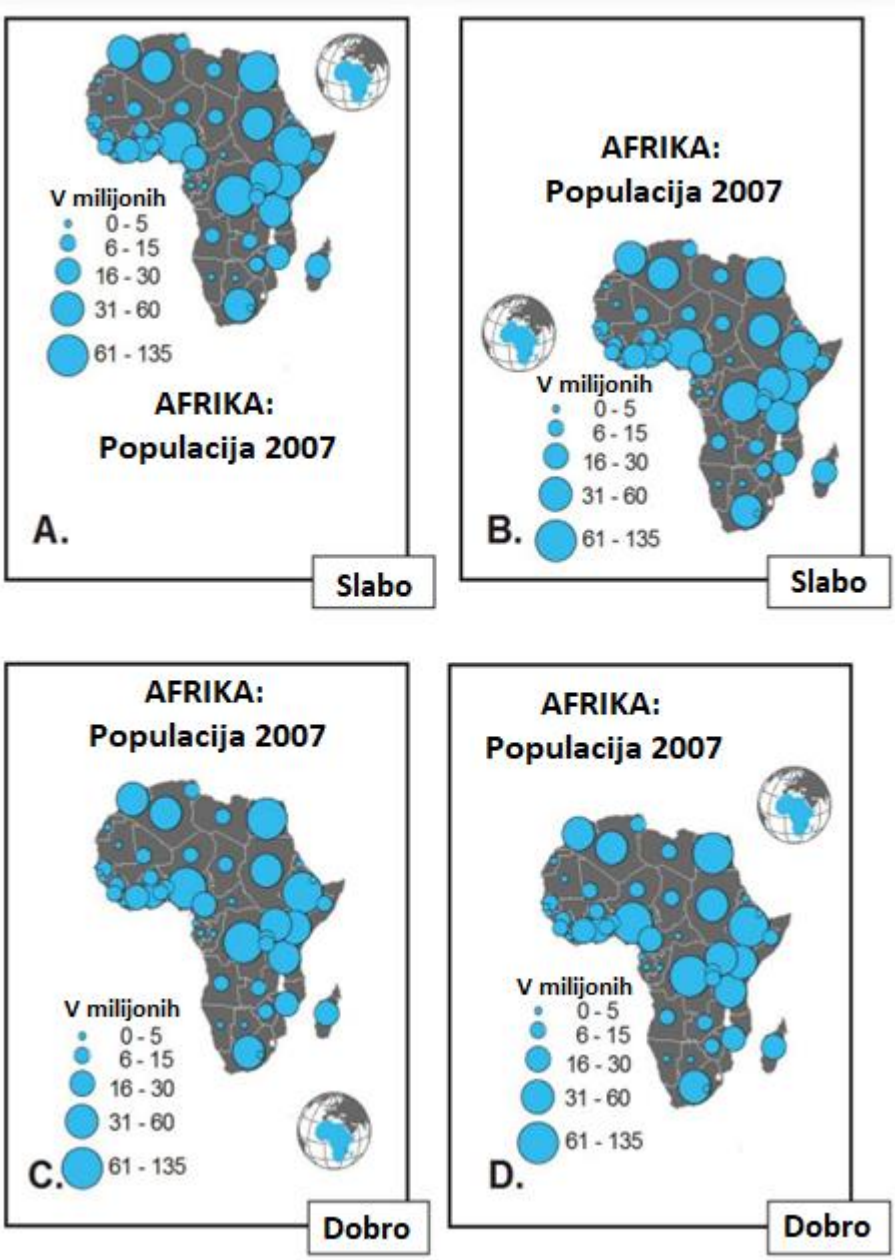


Prirejeno po Buckley (2012).

5.7 Postavitev elementov in ravnotežje

Pri vizualizaciji imamo prikazanih več elementov, kot so zemljevid, legenda in naslov. Pri organizaciji teh elementov moramo paziti na to, da je prikaz lepo uravnotežen. Ne želimo imeti vseh elementov na eni strani ali pa stisnjene na kupu. Z dobro organizacijo ustvarimo privlačen prikaz, ki je lažje berljiv in uporabnika usmerja po različnih elementih. Primer slabih in dobrih organizacij elementov je prikazan na sliki 7. Pri prikazu A so elementi preveč potisnjeni na vrh, pri prikazu B pa preveč na dno, kar ne odraža dobrega ravnotežja. Pri prikazih C in D pa je postavitev izboljšana. Zemljevid je prikazan približno na sredini, kar ga naredi bolj izpostavljenega in pomembnega. Za primer dobre postavitve elementov uporabimo prikaz D: ker smo ljudje navajeni brati od zgoraj navzdol, bi pri prikazu D najprej prebrali naslov, nato globus, sledila pa bi še zemljevid in legenda. Prikaz D poleg dobrega ravnotežja poskrbi tudi za to, da si uporabnik ogleda elemente v ustreznem vrstnem redu (Buckley, 2012).

Slika 7: Primeri slabega in dobrega ravnotežja med elementi



Prirejeno po Buckley (2012).

5.8 Izbira ustreznih barv

V poglavju Barve sem omenil, da lahko barvne palete razdelimo na kvalitativne, zaporedne in razhajajoče ter jih tudi opisal. Ta razdelitev velja tudi pri zemljevidih. Med barvnimi paletami izbiramo glede na tip podatkov in na to, kaj želimo z zemljevidom sporočiti. Zaporedno barvno paleto uporabimo, ko želimo vrednosti prikazati z enako barvo, razhajajočo barvno paleto pa ob prikazu 2 nasprotujočih si ekstremov. V obeh primerih so višje vrednosti prikazane s temnejšo barvo, nižje pa s svetlejšo. Pri zemljevidu z razhajajočo

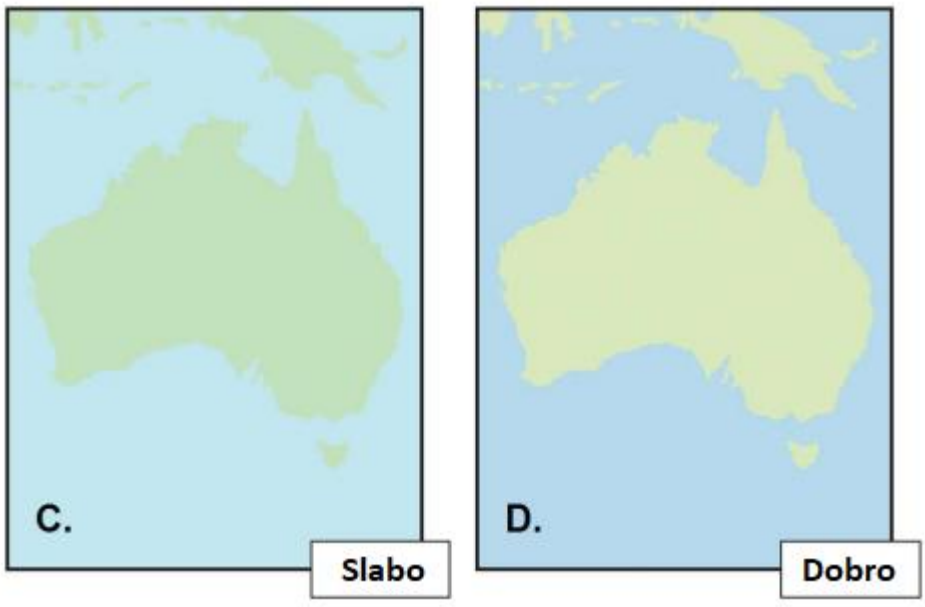
barvno paleto je uporabljena tudi bela barva, ki pripada sredinskim vrednostim med obema ekstremoma. Pri izbiri ustrezne barvne palete in barv si lahko pomagamo z različnimi temu namenjenimi orodji (Dougherty & Ilyankou, 2021). Biti moramo pozorni, da res uporabljamo ustrezno barvno paleto, saj lahko v nasprotnem primeru ustvarimo nejasne vizualizacije (Hohnova & Vondrakova, 2017). Kot že omenjeno, je ta težava podrobneje opisana pri splošnih vizualizacijah v poglavju Barve.

Treba se je držati tudi tega, da interval narašča enakomerno, npr. 0, 25, 50 namesto 0, 15, 50, drugače lahko bralca zmedemo. Pri prikazu kvalitativne barvne palete moramo biti pozorni na to, da ne prikazujemo preveč kategorij, in s tem barv. Zaradi prevelikega števila barv bo bralec moral večkrat preverjati legendo in bo z zemljevida težje razbiral, za katere kategorije gre (Datawrapper, 2021).

5.9 Kontrast uporabljenih barv

Ko prikazujemo podatke, ni priporočljiv prikaz z barvami, ki imajo nizek medsebojni kontrast. To lahko vodi v to, da težko ločimo med posameznimi elementi na vizualizaciji. Pri barvah z nizkim kontrastom dobimo občutek, da spadajo skupaj. Da se temu izognemo, moramo pri izbiri barv upoštevati tako svetlostni kontrast kot tudi barve same. Z uporabo visokega kontrasta ustvarimo izrazite elemente, ki jih ni težko ločiti med seboj. Primer lahko vidimo na sliki 8. Na levi strani imamo zemljevid z nizkim svetlobnim kontrastom, zato težko ločimo med kopnim in morjem. Na desni strani pa je prikazan levi zemljevid z višjim svetlobnim kontrastom in se zato tudi bolj razločno vidi. Višji kontrast je bil dosežen tako, da se je svetlost morja (modra barva) znižala (Buckley, 2012).

Slika 8: Primer nizkega in visokega kontrasta med barvami



Prirejeno po Buckley (2012).

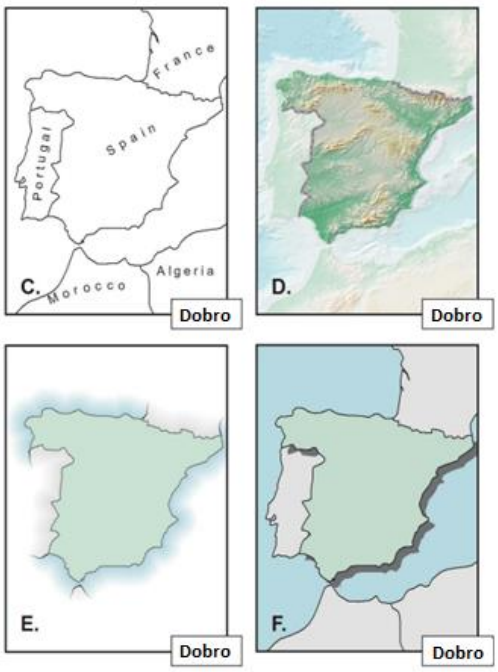
5.10 Izbira ustreznega barvnega intervala

Pri prikazu s horoplet zemljevidi moramo izbrati interval, s katerim bomo vrednosti razvrstili v skupine. Večji kot je interval, v manj skupin in s tem odtenkov so vrednosti razdeljene ter obratno. Če je interval prevelik, lahko preveč vrednosti pripada enakemu odtenku, če pa je premajhen, je zaradi majhnih sprememb v odtenkih lahko med njimi težko ločiti. Ob uporabi napačnega intervala se nam lahko zgodi tudi, da preveč skupin na zemljevidu pripada temnejšim odtenkom ali pa ravno obratno, in je zato težko primerjati vrednosti med območji. Najboljše je izbrati interval, pri katerem so skupine porazdeljene čez vse odtenke in lahko na zemljevidu brez težav primerjamo vrednosti med različnimi območji. Dober pristop je, da podatke prikažemo z različnimi intervali in se na podlagi tega odločimo, kateri je za naše podatke najbolj ustrezen (Dougherty & Ilyankou, 2021). Če vidimo, da je med različnimi območji težko ločiti, ker so si vrednosti in s tem odtenki preveč podobni, lahko za boljše razumevanje uporabimo katero drugo vrsto vizualizacije, npr. stolpčni diagram (Datawrapper, 2021).

5.11 Razločevanje med ospredjem in ozadjem

Če ocenimo, da deli zemljevida, s katerimi prikazujemo podatke, niso dovolj razločno ločeni od ozadja ali pa jih želimo bolj izpostaviti, lahko to spremenimo z različnimi načini. Primeri načinov so prikaz sence, dodajanje besedila in zakrivanje ali odstranitev ozadja. Z omenjenimi načini izpostavimo prikaz podatkov v ospredje in jim tako damo večji poudarek. Primeri poudarjanja so prikazani na sliki 9.

Slika 9: Primeri ločevanja ospredja od ozadja



Prirejeno po Buckley (2012).

6 IZDELAVA VIZUALIZACIJ PODATKOV Z ZEMLJEVIDI NA PRIMERU

6.1 Predstavitev problema

Za ta problem sem se odločil, ker mi omogoča dobro predstavitev zagotavljanja kakovosti vizualizacij podatkov z zemljevidi. Najprej sem poiskal podatke. Poiskal sem podatke, ki so dovolj nekakovostni, da sem na njih lahko izvedel čiščenje in z njimi predstavil čim več težav, ki sem jih opisal v poglavjih o kakovosti podatkov. Po pridobitvi podatkov sem določil problem. Zamislil sem si tak problem, da sem z njim lahko predstavil različne zemljevide ter čim več težav, opisanih v poglavjih kakovosti vizualizacij.

V Italiji želijo svojo turistično ponudbo dvigniti na višjo raven in eden od pristopov je tudi dvig ravni kulinarike. Zato so se odločili, da bodo regijam, v katerih ta segment najbolj peša, namenili sredstva, ki jih morajo porabiti v ta namen. Za pregled nad splošnim stanjem in lažjim razporejanjem sredstev želijo vizualizacije z zemljevidi, ki bodo stanje v Italiji informativno prikazovale in jim s tem olajšale delo. Prav tako želijo za namen oglaševanja iz podatkov razbrati, katere so najbolje ocenjene restavracije v posameznih regijah.

Informacije, ki jih želijo prikazane z zemljevidi, so:

- povprečna ocena restavracij v posamezni regiji,
- katera lastnost restavracij je v posamezni regiji najslabše ocenjena,
- prikaz lokacij in naslovov treh najbolje ocenjenih restavracij v posamezni regiji.

6.2 Power BI Desktop

Za reševanje problema sem uporabil aplikacijo Power BI Desktop. To je aplikacija, s katero se lahko povežemo s podatki, jih preoblikujemo in vizualiziramo. Podprti so tudi povezovanje več različnih podatkovnih virov, povezovanje več različnih vizualizacij in izdelava poročil (Microsoft, 2023).

Za uporabo aplikacije Power BI Desktop sem se odločil, ker sem jo že uporabljal in mi je njeno poznavanje omogočalo takojšnjo uporabo. Zaradi zgoraj omenjenih funkcionalnosti in velikega nabora različnih zemljevidov mi je dajala tudi dovolj možnosti za raziskovanje problematike. Podpira tudi povezovanje različnih vizualizacij, kar mi je omogočalo razširitev zemljevidov z dodatnimi prikazi.

6.3 Pridobivanje podatkov

Podatke sem pridobil s platforme Kaggle. Gre za platformo, ki poleg ostalih storitev ponuja tudi velik nabor podatkovnih virov. Podatkovni vir je rezultat spletnega strganja javno

dostopne spletne strani Tripadvisor, ki je bilo izvedeno v začetku maja 2021. Tripadvisor je platforma, ki med drugim hrani tudi podatke o restavracijah in uporabnikom omogoča njihovo ocenjevanje ter komentiranje.

Izbrani podatkovni vir (Leone, brez datuma) vsebuje podatke o restavracijah, ki se nahajajo v Evropi in so objavljene na platformi Tripadvisor. Podatkovni vir je v shranjen z vrednostmi, ločenimi z vejico (angl. comma-separated values, v nadaljevanju CSV) in je velik 679.68 megabajtov. V njem je shranjenih 1.083.349 restavracij s pripadajočimi atributi.

6.4 Profiliranje podatkov in ocenjevanje kakovosti

6.4.1 Identifikacija atributov

Podatkovni vir hrani 42 atributov. Opisani so v tabeli 3.

Tabela 3: Opis atributov

Ime	**Tip**	**Opis**
restaurant_link	Besedilo	Unikatni del enoličnega krajevnika vira (angl. Uniform Resource Locator – URL) na spletni strani Tripadvisor
restaurant_name	Besedilo	Ime restavracije
original_location	Besedilo	Seznam, ki hrani vrednosti tipa string za celino, državo, regijo, provinco in mesto
country	Besedilo	Država
region	Besedilo	Regija
province	Besedilo	Provinca
city	Besedilo	Mesto
address	Besedilo	Naslov
latitude	Decimalno število	Zemljepisna širina
longitude	Decimalno število	Zemljepisna dolžina
claimed	Besedilo	Binarna vrednost, ki nam pove, ali je restavracijo na spletni strani Tripadvisor prevzel lastnik
awards	Besedilo	Nagrade, ki jih je restavracija prejela, ločene z vejico
popularity_detailed	Besedilo	Tako kot atribut popularity_generic, le bolj podrobno
popularity_generic	Besedilo	Besedilo, ki opisuje priljubljenost restavracije glede na ostale v okolici
top_tags	Besedilo	Lastnosti, ki opisujejo tip restavracije

Tabela 3: Opis atributov (nad.)

Ime	**Tip**	**Opis**
price_level	Besedilo	Cena hrane, prikazana z različnim številom znakov v evrih
price_range	Besedilo	Razpon cen, zapisan z najnižjo in najvišjo ceno
meals	Besedilo	Vrste obrokov, ločene z vejico
cuisines	Besedilo	Vrste kuhinje
special_diets	Besedilo	Posebne diete, ki so v ponudbi
features	Besedilo	Lastnosti, ki opisujejo posebne lastnosti restavracije
vegetarian_friendly	Logična vrednost	Binarna vrednost, ki nam pove, ali restavracija ponuja tudi vegetarijanske obroke
vegan_options	Logična vrednost	Binarna vrednost, ki nam pove, ali restavracija ponuja tudi veganske obroke
gluten_free	Logična vrednost	Binarna vrednost, ki nam pove, ali restavracija ponuja tudi obroke brez glutena
original_open_hours	Besedilo	Odpiralni časi
open_days_per_week	Decimalno število	Število dni v tednu, ko je restavracija odprta
open_hours_per_week	Decimalno število	Število ur v tednu, ko je restavracija odprta
working_shifts_per_week	Decimalno število	Število delovnih izmen v tednu
avg_rating	Decimalno število	Povprečna ocena
total_reviews_count	Decimalno število	Število vseh ocen
default_language	Besedilo	Privzeti jezik
reviews_count_in_default_language	Decimalno število	Število ocen v privzetem jeziku
excellent	Decimalno število	Število odličnih ocen v privzetem jeziku
very_good	Decimalno število	Število zelo dobrih ocen v privzetem jeziku
average	Decimalno število	Število povprečnih ocen v privzetem jeziku
poor	Decimalno število	Število slabih ocen v privzetem jeziku
terrible	Decimalno število	Število zelo slabih ocen v privzetem jeziku
food	Decimalno število	Povprečna ocena hrane
service	Decimalno število	Povprečna ocena postrežbe
value	Decimalno število	Povprečna ocena vrednosti
atmosphere	Decimalno število	Povprečna ocena vzdušja

Tabela 3: Opis atributov (nad.)

Ime	Tip	Opis
keywords	Besedilo	Besede, ki se na spletni strani največkrat pojavijo

Prirejeno po Leone (brez datuma).

V tabeli 3 lahko vidimo, da je v podatkovnem viru na voljo velik nabor atributov. Za reševanje mojega problema pa nisem potreboval vseh, zato sem ocenjevanje kakovosti podatkov izvedel le na tistih, ki so bili za moj primer relevantni. Preostale atribute sem v nadaljevanju pri čiščenju podatkov tudi odstranil.

Prav tako nisem potreboval vseh vnosov oz. restavracij, ampak le tiste, ki se nahajajo v Italiji. V poglavjih Čiščenje podatkov in Odstranjevanje nepotrebnih stolpcev in vrstic je odstranjevanje vrstic sicer uvrščeno v proces čiščenja podatkov, ker pa so bile za moj problem pri procesu ocenjevanja kakovosti relevantne samo restavracije iz Italije, sem preostale odstranil že prej. S tem sem podatkovni vir izčrpno zmanjšal in preprečil lažne rezultate ocenjevanja kakovosti podatkov, saj ga nisem izvajal na podatkih, ki za moj primer niso bili relevantni.

Glede na zahtevane vizualizacije sem identificiral 5 lokacijskih in 6 opisnih atributov. Lokacijski atributi, ki sem jih potreboval, so:

- country,
- region,
- address,
- latitude,
- longitude.

Opisni atributi, ki sem jih potreboval, so:

- restaurant_name,
- avg_rating,
- food,
- service,
- value,
- atmosphere.

6.4.2 Začetno odstranjevanje vrstic

Ker je problem zastavljen tako, da zanj potrebujemo samo restavracije v Italiji, bi ocenjevanje podatkov s tudi preostalimi podatki odražalo napačno stanje. Zato sem restavracije iz ostalih držav odstranil že na tem koraku. S tem sem si ustvaril začetni

podatkovni vir, ki je sicer še vedno vseboval nekakovostne podatke, vendar je bil primeren za reševanje mojega problema. Vrstice, ki so vsebovale pomanjkljive podatke oz. jih nisem znal uvrstiti, v katero državo spadajo, sem ohranil in počistil v procesu čiščenja podatkov. Odstranil sem le restavracije, za katere sem bil prepričan, da niso iz Italije.

To sem storil tako, da sem vrstice, ki pri atributu country ne predstavljajo Italije, odstranil. Moral pa sem biti pozoren, saj bi se mi lahko zgodila težava, opisana v poglavju identifikacije in čiščenja prostorskih podatkov, ki pravi, da je lahko kateri od opisnih atributov napačen. V tem primeru bi se mi lahko zgodilo, da bi katera od sicer tujih restavracij lažno hranila vrednosti Italije ali pa se katera od restavracij iz Italije izdajala za tujo. Zato sem moral biti pri tem procesu pazljiv.

Podatke sem naložil v aplikacijo Power BI Desktop in tam preveril, ali kateri od vnosov hrani napačno vrednost države. Pri čiščenju prostorskih podatkov sem opisal, da lahko to preverjamo bodisi s prikazom na zemljevidu bodisi s poizvedbami na podatkih. Težava s prikazom na zemljevidu je, da bi zanj potreboval pravilne koordinate oz. naslove, pri mojih podatkih pa tudi nisem vedel, ali to so (to sem ugotavljal v nadaljevanju pri zagotavljanju kakovosti in pri čiščenju podatkov). Napačne koordinate ne bi prikazovale pravilne lokacije na zemljevidu, pri prikazu naslovov pa bi se lahko zgodila že omenjena težava z napačnimi rezultati geokodiranja.

Zato sem restavracije iz Italije iskal s poizvedbami. Več atributov hrani podatek o tem, v kateri državi se restavracija nahaja. Preveril sem, ali se atribut country ujema s preostalimi atributi, in tako potrdil, da je restavracija res iz Italije. Atributi, ki sem jih uporabil za preverjanje, so:

- original_location: Za primerjavo vrednosti države.
- address: Za primerjavo vrednosti države.
- restaurant_link: Za ročno preverjanje vnosov na spletu.

Pred nalaganjem podatkov v Power BI Desktop sem v integriranem orodju Power Query odstranil nepotrebne atribute in ohranil le zgoraj omenjene. Nato sem podatke naložil v Power BI Desktop. Na začetku sem iz podatkov razbral nekatere informacije, ki so mi pomagale pri ugotavljanju, katere vrstice moram odstraniti. Te informacije so:

- Število vseh vnosov: **1.083.349**.
- Število vnosov, ki pri atributu country hranijo vrednost »Italy«: **224.703**:
  - vsebujejo besedilo »Italy« pri atributih original_location in address: **224.703**,
  - ne vsebujejo besedila »Italy« pri atributih original_location in address: **0**.
- Število vnosov, ki pri atributu country hranijo vrednost druge države: **858.633**:
  - vsebujejo besedilo »Italy« pri atributu original_location: **0**,
  - vsebujejo besedilo »Italy« pri atributu address: **1**.
- Število vnosov, ki pri atributu country hranijo neveljavno vrednost: **13**.

V zgornjih informacijah lahko vidimo, da so podatki precej v redu. Problematičnih je 14 vnosov. 13 vnosov, ki pri atributu country hranijo neveljavno vrednost, sem pustil za kasnejše čiščenje. Težavo z 1 vnosom pa sem moral rešiti že na tem koraku, saj ta pri atributu country vsebuje veljavno vrednost, ki ni »Italy«, pri atributu address pa vsebuje vrednost »Italy«. Ta vnos je prikazan na sliki 10. Že iz slike lahko vidimo, da gre najverjetneje za restavracijo v Franciji in za veljaven naslov.

Slika 10: Restavracija z morebitnimi napačnimi vrednostmi

Vir: lastno delo.

Da sem se prepričal, sem to restavracijo s pomočjo atributa restaurant_link poiskal tudi na spletu. Po pregledu spletne strani Tripadvisor sem odkril, da gre za restavracijo iz Francije, katere naslov vsebuje besedo »Italy« (Tripadvisor, brez datuma).

Glede na zgornje ugotovitve, sem v orodju Power Query iz podatkovnega vira odstranil vse restavracije, ki pri atributu country vsebujejo veljavno vrednost tuje države in razveljavil korak, pri katerem sem odstranil odvečne stolpce. Prefiltrirane podatke sem ponovno naložil v Power BI Desktop in potrdil, da se novo število vnosov ujema z zgornjimi ugotovitvami. Novo število vnosov je bilo 224.716.

6.4.3 Ocenjevanje kakovosti

Pri ocenjevanju kakovosti sem se posluževal elementov kakovosti, ki so specifični za prostorske podatke in jih opisal pri kakovosti prostorskih podatkov. Kot sem že omenil, elementi notranje kakovosti prostorskih podatkov zahtevajo referenčno bazo, ki hrani resnične podatke in s katero lahko podatke primerjamo. Ker teh nisem imel na voljo, sem ocenjevanje nekoliko prilagodil za moj primer uporabe. Ocenjevanje sem izvedel samo na 11 atributih, ki so potrebni za analizo.

Ocenjevanje sem izvedel samo na elementih, ki so za moj primer relevantni. Ker moj primer uporabe ne zahteva nobenega datuma in tudi podatkovni vir ne vsebuje nobenega stolpca, ki bi vseboval datume, časovne konsistentnosti in točnosti merjenja časa nisem preverjal. Prav tako sem izpustil topološko konsistentnost, saj sem pri pregledu virov naletel samo na primere napak pri črtah in poligonih. Ker so v mojem primeru prostorski podatki predstavljeni kot točke, primerov napak konsistentnosti nisem mogel aplicirati. Ker nisem imel na voljo referenčne podatkovne baze, s katero bi primerjal svoje vrednosti, sem izpustil tudi preverjanje tematske točnosti.

6.4.3.1 Pozicijska točnost

Pri pozicijski točnosti bi za izračun potreboval podatkovni vir z istimi restavracijami, ki zanje hrani pravilne podatke. Koordinate v mojem viru bi nato primerjal s pravilnimi in tako odkril, kakšno napako hranijo koordinate v mojem podatkovnem viru.

Ena od zahtev je tudi prikaz lokacij nekaterih restavracij na zemljevidu. Zaradi te zahteve morajo biti koordinate precej točne. Dobro bi bilo, da bi bila napaka manjša od 10 metrov, saj pri tej oddaljenosti lokacija že lahko predstavlja drugo stavbo. Ker podatkovnega vira, ki bi hranil pravilne podatke, nisem imel na voljo, pozicijske točnosti nisem mogel izračunati. Ker pa je v zahtevi navedeno, da moram prikazati le 3 restavracije na regijo, sem lokacije točk preveril na koncu kar ročno in se s tem prepričal, da se nahajajo na pravilnih lokacijah. Ročno preverjanje sem izvedel z uporabo platforme Tripadvisor.

6.4.3.2 Konceptualna logična konsistentnost

Ker za moj primer uporabe ni določeno, kako morajo biti razredi in atributi v podatkovnem viru poimenovani, so ti lahko poimenovani poljubno. Vseeno pa sem preveril, ali so imena logična ali zavajajoča.

Podatkovni vir je le 1 datoteka, ki vsebuje vse podatke in ni razdeljen na razrede. Ta datoteka je 1 razred, ki predstavlja restavracije. Ime datoteke je »tripadvisor_european_restaurants«. To ime je bilo ustrezno, preden sem iz nje odstranil vse restavracije, ki se ne nahajajo v Italiji. Zdaj pa je ime zavajajoče, saj po novem vsebuje le restavracije iz Italije. Bolj ustrezno ime bi bilo »tripadvisor_italian_restaurants«.

Imena izbranih atributov so country, region, address, latitude, longitude, restaurant_name, avg_rating, food, service, value in atmosphere. Glede na opise, ki sem jih opisal pri identifikaciji atributov, so njihova imena smiselna. Kar bi spremenil so imena atributov food, service, value in atmosphere, saj sami po sebi niso opisni. Prav tako kot atribut avg_rating tudi ti 4 atributi opisujejo povprečno oceno, zato bi jih lahko poimenovali sorodno, in sicer food_avg_rating, service_avg_rating, value_avg_rating in atmosphere_avg_rating.

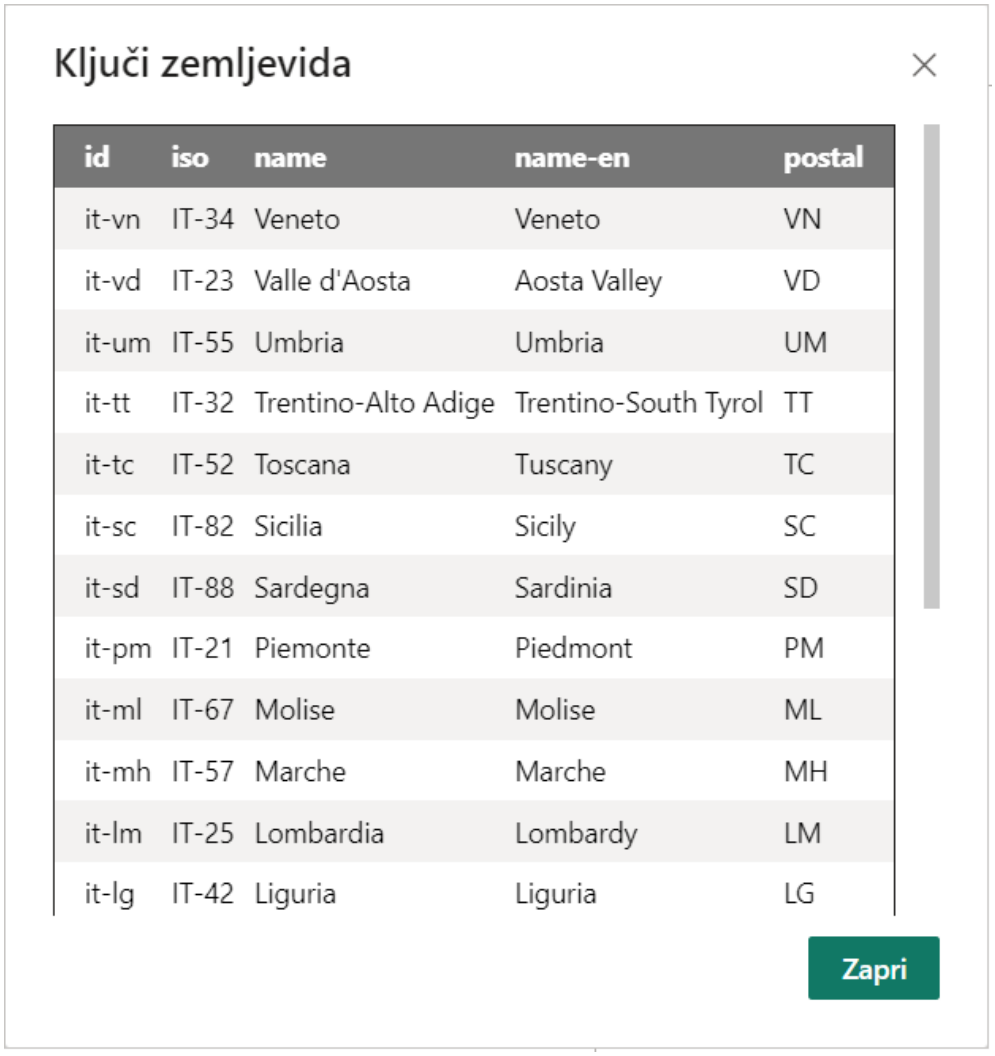
6.4.3.3 Domenska logična konsistentnost

Pri preverjanju domenske logične konsistentnosti sem za posamezne atribute definiral naslednje veljavne vrednosti:

- country: Vsi vnosi morajo hraniti vrednost »Italy«.
- latitude: Preveril sem, da je vrednost znotraj dovoljenega razpona –90 in 90, ki sem ga omenil pri čiščenju neveljavnih koordinat. Vrednost v dovoljenem razponu je sicer še vedno lahko izven Italije, vseeno pa sem preveril, ali je ta veljavna.

– longitude: Preveril sem, da je vrednost znotraj dovoljenega razpona –180 in 180, ki sem ga omenil pri čiščenju neveljavnih koordinat. Vrednost v dovoljenem razponu je sicer še vedno lahko izven Italije, vseeno pa sem preveril, ali je ta veljavna.
– avg_rating: Vrednost mora biti med 1 in 5.
– food: Vrednost mora biti med 1 in 5.
– service: Vrednost mora biti med 1 in 5.
– value: Vrednost mora biti med 1 in 5.
– atmosphere: Vrednost mora biti med 1 in 5.
– region: Preveril sem, kakšna poimenovanja italijanskih regij pričakuje privzeti horoplet zemljevid v Power BI Desktopu. Gre za angleška poimenovanja, ki so na sliki 11 prikazana v stolpcu name-en in so Veneto, Aosta Valley, Umbria, Trentino-South Tyrol, Tuscany, Sicily, Sardinia, Piedmont, Molise, Marche, Lombardy, Liguria, Lazio, Friuli-Venezia Giulia, Emilia-Romagna, Campania, Calabria, Basilicata, Apulia in Abruzzo. Preveril sem, ali vsak od vnosov zaseda eno od teh vrednosti.

Slika 11: Imena italijanskih regij pri privzetem horoplet zemljevidu v Power BI Desktopu



id	iso	name	name-en	postal
it-vn	IT-34	Veneto	Veneto	VN
it-vd	IT-23	Valle d'Aosta	Aosta Valley	VD
it-um	IT-55	Umbria	Umbria	UM
it-tt	IT-32	Trentino-Alto Adige	Trentino-South Tyrol	TT
it-tc	IT-52	Toscana	Tuscany	TC
it-sc	IT-82	Sicilia	Sicily	SC
it-sd	IT-88	Sardegna	Sardinia	SD
it-pm	IT-21	Piemonte	Piedmont	PM
it-ml	IT-67	Molise	Molise	ML
it-mh	IT-57	Marche	Marche	MH
it-lm	IT-25	Lombardia	Lombardy	LM
it-lg	IT-42	Liguria	Liguria	LG

Vir: lastno delo.

Atributov address in restaurant_name zaradi velikega nabora vrednosti, ki mi niso znane, nisem preverjal. Po preverjanju sem prišel do rezultatov, ki so prikazani v tabeli 4.

Tabela 4: Ocenjevanje domenske logične konsistentnosti

Ime atributa	**Število veljavnih vrednosti**	**Število neveljavnih vrednosti**
country	224.703	13
region	200.419	24.297
latitude	2	224.714
longitude	26	224.690
avg_rating	0	224.716
food	0	224.716
service	0	224.716
value	0	224.716
atmosphere	0	224.716

Vir: lastno delo.

Pri atributih latitude, longitude, avg_rating, food, service, value in atmosphere sem opazil, da je le malo vrednosti veljavnih in da so vrednosti v večini primerov prevelike. Opazil sem tudi, da te vrednosti nimajo decimalne vejice, čeprav bi morale imeti decimalni tip. Po odprtju podatkovnega vira v programu Microsoft Excel sem opazil, da so decimalne vejice zapisane s piko. Gre za znano težavo, ki se zgodi zaradi območnih nastavitev za uvoz v Power BI Desktop. Od območnih nastavitev za uvoz je med drugim odvisno, kako Power BI Desktop ob uvozu datoteke obravnava števila (How to Power BI, 2021). Ker sem imel izbrano nastavitev »slovenščina (Slovenija)«, so bile pike obravnavane namesto decimalnih ločil kot pike za ločevanje tisočic od nižjih enot števila. Napačne vrednosti sem naslovil v nadaljevanju pri čiščenju podatkov.

6.4.3.4 Logična konsistentnost formatov

Glede na tabelo 3 zasedajo vsi atributi, ki jih potrebujem, pravilen tip. Sem pa že pri preverjanju domenske logične konsistentnosti ugotovil, da atributi latitude, longitude, avg_rating, food, service, value in atmosphere ne vsebujejo vrednosti, skladne z definiranimi tipi. V vseh primerih so vrednosti cela števila namesto decimalnih števil. V to težavo sem se, kot sem že prej omenil, poglobil v nadaljevanju pri čiščenju podatkov.

6.4.3.5 Popolnost

Popolnosti vnosov nisem mogel preveriti, saj na voljo nisem imel referenčne podatkovne baze. Potreboval bi podatkovno bazo, da bi lahko ugotovil kolikšen presežek in izpustitev restavracij ima moj podatkovni vir.

Pri atributih pa sem lahko za vnose, ki jih moj podatkovni vir vsebuje, za posamezen atribut preveril, koliko vrednosti hrani in koliko jih manjka. Po preverjanju sem prišel do rezultatov, prikazanih v tabeli 5.

Tabela 5: Ocenjevanje popolnosti

Ime atributa	**Število izpolnjenih vrednosti**	**Število praznih vrednosti**
country	224.709	7
region	224.709	7
address	224.694	22
latitude	222.555	2.161
longitude	222.555	2.161
avg_rating	208.013	16.703
restaurant_name	224.710	6
food	135.434	89.282
service	135.811	88.905
value	135.658	89.058
atmosphere	73.030	151.686

Vir: lastno delo.

V tabeli 5 lahko vidimo, da imajo atributi latitude, longitude, avg_rating, food, service, value in atmosphere veliko praznih vrednosti. Pri atributih, ki predstavljajo povprečno oceno, moramo vedeti to, da lahko te restavracije niso prejele še nobene ocene in je vrednost zato tam prazna, pri vrednostih latitude in longitude, pa sem v nadaljevanju pri samih vizualizacijah preveril, ali jih potrebujem, ali zadostuje atribut address. Za vizualizacijo z zemljevidi se lahko uporabi tudi naslov. Ker je število manjkajočih vrednosti pri naslovu samo 22, se lahko te restavracije poišče na spletni strani Tripadvisor in naslove dopolni ročno ali pa se jih pridobi pri čiščenju podatkov opisanim nasprotnim geokodiranjem s pomočjo koordinat.

6.4.3.6 Časovna veljavnost

Na spletni strani, s katere sem naložil podatkovni vir, je napisano, da so bili podatki v začetku maja 2021 strgani s spletne strani Tripadvisor. Resničnost navedenega časa zajema podatkov žal ne morem preveriti, ker nimam referenčne podatkovne baze. Če je navedeni čas resničen, pa za moje zahteve ustreza, saj se ni zgodil tako dolgo nazaj in odraža dovolj ažurno stanje.

Vseeno pa se je lahko od takrat katera od restavracij zaprla. Zato bi moral preveriti, ali je bila katera od restavracij, ki jih moj podatkovni vir vsebuje, v tem času s spletne strani Tripadvisor izbrisana. Ker nisem imel na voljo referenčne podatkovne baze in je restavracij veliko, bi to težko preveril. Odločil sem se, da je za identifikacijo povprečnih ocen restavracij in najslabše ocenjenih lastnosti v regijah podatkovno stanje iz 2021 v redu. Pri identifikaciji najboljših restavracij na zemljevidu pa sem pri vizualizacijah za te restavracije na spletni strani Tripadvisor ročno preveril, ali so še vedno odprte.

6.5 Čiščenje podatkov

6.5.1 Odstranjevanje duplikatov

V orodju Power Query sem poiskal vse duplikate, ki pri atributu restaurant_link nimajo edinstvene vrednosti. Rezultat je bila tabela, ki je prikazana na sliki 12. Na njej sem z rdečo barvo označil vnose, ki hranijo neveljavne podatke, z modro pa vnose, ki hranijo veljavne podatke in so podvojeni. Z rdečo barvo označeni vnosi nimajo veljavnih vrednosti niti pri atributu restaurant_link niti pri ostalih atributih. Zato sem čiščenje teh vnosov pustil za kasneje, na tem koraku pa se osredotočil samo na vnose, označene z modro barvo.

Slika 12: Rezultat iskanja duplikatov

	restaurant_link	restaurant_name	original_location	country	region
1					
2	London				
3	EC1V 1NR	London England"	51.527794	-0.089627	Unclaimed
4	London				
5	EC1V 1NR	London EC1V 1NR England"	51.511173	-0.08737	Unclaimed
6	g187881-d10792250	Asd Circolo Ippico La Fattoria	["Europe", "Italy", "Sardinia", "Province of Cagliari", "Cagliari"]	Italy	Sardinia
7	g187881-d10792250	Asd Circolo Ippico La Fattoria	["Europe", "Italy", "Sardinia", "Province of Cagliari", "Cagliari"]	Italy	Sardinia
8	g635632-d21132220	Spaghetteria Tavola Calda	["Europe", "Italy", "Sardinia", "Province of Olbia-Tempio", "Trinita d'Ag...	Italy	Sardinia
9	g635632-d21132220	Spaghetteria Tavola Calda	["Europe", "Italy", "Sardinia", "Province of Olbia-Tempio", "Trinita d'Ag...	Italy	Sardinia
10					

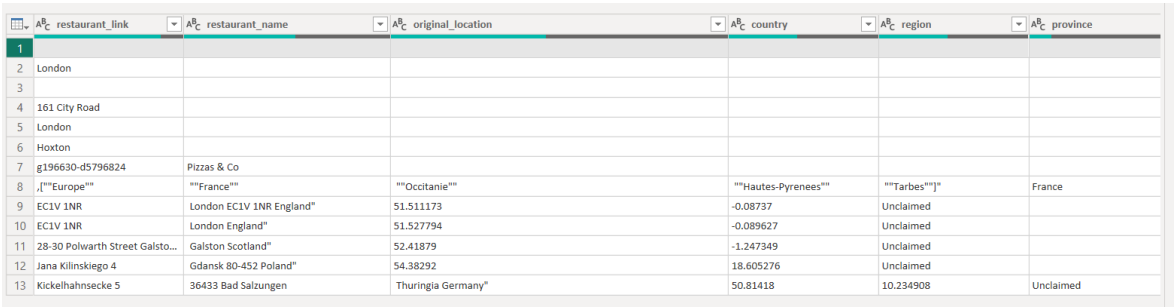
Vir: lastno delo.

Da bi se prepričal, ali so z modro barvo označeni vnosi res duplikati, sem preverjanje iz atributa restaurant_link razširil na vse atribute. Kot rezultat sem dobil tabelo s 4 vnosi, to so vnosi, ki so na sliki 12 označeni z modro barvo. To pomeni, da se vnosa na vrsticah 6 in 7, ter vnosa na vrsticah 8 in 9 ujemata v vseh atributih oz. sta popolnoma enaka. Zato sem enega v vsakem paru odstranil. V nasprotnem primeru bi ostala 2 vnosa podvojena, kar bi odražalo napačno stanje tudi na kasnejših vizualizacijah, saj bi določene lokacije imele več predstavnikov, kot je teh v resničnosti.

6.5.2 Preverjanje neveljavnih vrstic

Že pri ocenjevanju kakovosti sem odkril, da podatkovni vir vsebuje tudi neveljavne vrednosti. Nekatere od njih so prikazane tudi na sliki 12 in so obkrožene z rdečo barvo. Zato sem se odločil, da te vnose podrobneje pregledam in jih na podlagi ugotovitev popravim ali odstranim. Ker so na sliki 12 prikazani le duplikati, sem na podlagi sumljivih vrednosti pri posameznih atributih ponovno izvedel filtriranje in prišel do rezultatov, ki so prikazani na sliki 13.

Slika 13: Neveljavne vrstice



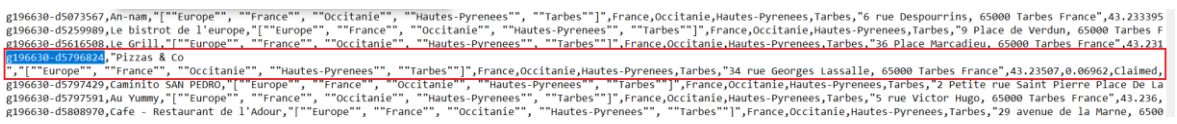
	restaurant_link	restaurant_name	original_location	country	region	province
1						
2	London					
3						
4	161 City Road					
5	London					
6	Hoxton					
7	g196630-d5796824	Pizzas & Co				
8	,[""Europe""	""France""	""Occitanie""	""Hautes-Pyrenees""	""Tarbes""]"	France
9	EC1V 1NR	London EC1V 1NR England"	51.511173	-0.08737	Unclaimed	
10	EC1V 1NR	London England"	51.527794	-0.089627	Unclaimed	
11	28-30 Polwarth Street Galsto...	Galston Scotland"	52.41879	-1.247349	Unclaimed	
12	Jana Kilinskiego 4	Gdansk 80-452 Poland"	54.38292	18.605276	Unclaimed	
13	Kickelhahnsecke 5	36433 Bad Salzungen	Thuringia Germany"	50.81418	10.234908	Unclaimed

Vir: lastno delo.

Preverjanje identificiranih vrstic sem začel tako, da sem originalni podatkovni vir odprl z aplikacijo Notepad in preveril, ali so te vrstice bile neveljavne že tam in je do napak mogoče prišlo pri nalaganju v Power BI Desktop. Za posamezne vrstice sem prišel do naslednjih ugotovitev:

- Vrstici 2 in 10 sta neveljavni zaradi preloma vrstice in pripadata vrstici oz. restavraciji, ki ima vrednost restaurant_link enako g186338-d22953153. Po združitvi vrstic sem ugotovil, da gre za restavracijo iz Anglije, zato sem jo odstranil.
- Vrstice 4, 5, 6 in 9 so neveljavne zaradi preloma vrstice in pripadajo restavraciji, ki ima vrednost restaurant_link enako g186338-d22992346. Po združitvi vrstic sem tudi pri tej restavraciji odkril, da gre za restavracijo iz Anglije, zato sem jo odstranil iz podatkovnega vira.
- Prav tako gre za težavo s prelomom vrstic pri restavracijah v vrsticah 11, 12 in 13. Tu gre za 3 različne restavracije, ki se nahajajo v Angliji, na Poljskem in v Nemčiji. Zato sem tudi te 3 restavracije odstranil iz podatkovnega vira.
- Vrstici 1 in 3 sta popolnoma prazni in pri nobenem od atributov ne vsebujeta vrednosti. Tudi ti 2 vrstici sem uspel izslediti v prvotnem podatkovnem viru in sta posledica preloma vrstic pri 2 ostalih zgoraj omenjenih restavracijah, zato sem ju odstranil.
- Vrstici 7 in 8 predstavljata 1 restavracijo, ki je bila že v prvotnem podatkovnem viru za atributom restaurant_name prelomljena v novo vrstico. To lahko vidimo na sliki 14. Po popravku, s katerim sem 2 vrstici združil v eno, in ponovnem nalaganju vira v Power BI Desktop, sem ugotovil, da gre za restavracijo iz Francije in zato vrstico odstranil.

Slika 14: Primer prelomljene vrstice



```
g196630-d5073567,An-nam,"[""Europe"", ""France"", ""Occitanie"", ""Hautes-Pyrenees"", ""Tarbes""]",France,Occitanie,Hautes-Pyrenees,Tarbes,"6 rue Despourrins, 65000 Tarbes France",43.233395
g196630-d5259989,Le bistrot de l'europe,"[""Europe"", ""France"", ""Occitanie"", ""Hautes-Pyrenees"", ""Tarbes""]",France,Occitanie,Hautes-Pyrenees,Tarbes,"9 Place de Verdun, 65000 Tarbes F
g196630-d5616508,Le Grill,"[""Europe"", ""France"", ""Occitanie"", ""Hautes-Pyrenees"", ""Tarbes""]",France,Occitanie,Hautes-Pyrenees,Tarbes,"36 Place Marcadieu, 65000 Tarbes France",43.231
g196630-d5796824,"Pizzas & Co
","[""Europe"", ""France"", ""Occitanie"", ""Hautes-Pyrenees"", ""Tarbes""]",France,Occitanie,Hautes-Pyrenees,Tarbes,"34 rue Georges Lassalle, 65000 Tarbes France",43.23507,0.06962,Claimed,
g196630-d5797429,Caminito SAN PEDRO,"[""Europe"", ""France"", ""Occitanie"", ""Hautes-Pyrenees"", ""Tarbes""]",France,Occitanie,Hautes-Pyrenees,Tarbes,"2 Petite rue Saint Pierre Place De La
g196630-d5797591,Au Yummy,"[""Europe"", ""France"", ""Occitanie"", ""Hautes-Pyrenees"", ""Tarbes""]",France,Occitanie,Hautes-Pyrenees,Tarbes,"5 rue Victor Hugo, 65000 Tarbes France",43.236,
g196630-d5808970,Cafe - Restaurant de l'Adour,"[""Europe"", ""France"", ""Occitanie"", ""Hautes-Pyrenees"", ""Tarbes""]",France,Occitanie,Hautes-Pyrenees,Tarbes,"29 avenue de la Marne, 6500
```

Vir: lastno delo.

Ugotovil sem, da se vseh 13 vnosov bodisi ne nahaja v Italiji bodisi so produkt prelomljenih in praznih vrstic. Zato sem vse te vnose odstranil.

6.5.3 Odstranjevanje odvečnih atributov

Pri preoblikovanju podatkov sem opisal, da je iz podatkovnega vira treba odstraniti atribute, ki jih ne potrebujemo. Na začetku praktičnega dela pa je definirano, da za raziskavo potrebujem le 11 od 42 atributov. Zato sem iz podatkovnega vira odstranil vse atribute, ki jih pri raziskavi ne potrebujem. Pri odstranjevanju sem uporabil orodje Power Query, po odstranjevanju pa je moj podatkovni vir vseboval naslednje atribute:

- country,
- region,
- address,
- latitude,
- longitude,
- restaurant_name,
- avg_rating,
- food,
- service,
- value,
- atmosphere.

6.5.4 Pretvorba podatkovnih tipov

Že pri ocenjevanju kakovosti sem ugotovil, da vrednosti atributov, ki bi morali biti decimalnega tipa, ne vsebujejo decimalnih vejic. Po pregledu podatkovnih tipov, ki jih je Power BI Desktop določil posameznim atributom, sem opazil, da vseh 7 številskih atributov vsebuje napačen tip, tj. celo število namesto decimalno število. Ti atributi so latitude, longitude, avg_rating, food, service, value in atmosphere. Po spreminjanju podatkovnega tipa teh atributov v decimalni tip, decimalne vejice še vedno ni bilo prisotne, saj se je podatek o tem izgubil že ob nalaganju podatkov.

Težavo sem identificiral že pri ugotavljanju domenske logične konsistentnosti. Območno nastavitev v Power BI Desktopu sem nastavil na »angleščina (Združeno kraljestvo)« in podatke še enkrat naložil v aplikacijo. Po nalaganju je vseh 7 številskih atributov pravilno vsebovalo decimalne vejice.

6.5.5 Standardizacija

Pri domenski logični konsistentnosti sem ugotovil, da nekatere od regij ne vsebujejo vrednosti, ki jih pričakuje privzeti horoplet zemljevid v Power BI Desktopu. To so zapisi, ki hranijo vrednost:

- Emilia Romagna namesto Emilia-Romagna,
- Puglia namesto Apulia,
- Valle d'Aosta namesto Aosta Valley,
- Friuli Venezia Giulia namesto Friuli-Venezia Giulia,
- Trentino-Alto Adige namesto Trentino-South Tyrol.

Zato sem na tem koraku standardiziral vrednosti pri regiji Emilia-Romagna, kjer le nekatere od vrednosti vsebujejo napačno vrednost Emilia Romagna, preostale zgoraj naštete regije pa v celoti popravil. Na sliki 15 je na levi strani prikazan zemljevid s prvotnimi vrednostmi imen regij, na desni strani pa zemljevid s prilagojenimi imeni regij. Regije, ki jih zemljevid ni znal preslikati, so pobarvane s sivo barvo, preostale regije pa z modro barvo. Vidimo lahko, da levi zemljevid nekaterih regij ni znal preslikati, desni zemljevid s prilagojenimi imeni pa vse regije uspešno preslika.

Slika 15: Prikaz regij pred (levo) in po (desno) prilagoditvi imen

Vir: lastno delo.

6.5.6 Nasprotno geokodiranje

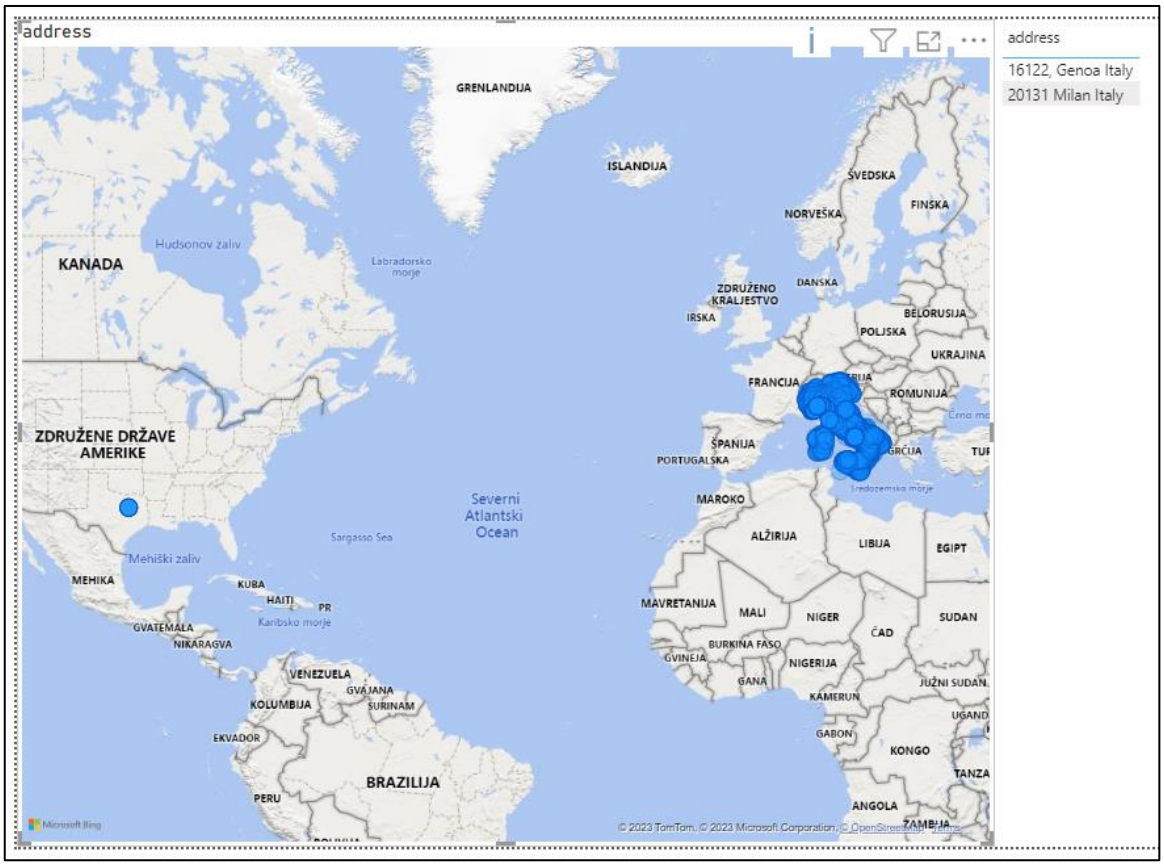
Naslove oz. koordinate sem potreboval za prikaz najboljše ocenjenih restavracij. Ker manjši delež vnosov ne hrani koordinat ali naslovov, je obstajala velika verjetnost, da bodo na koncu te restavracije vsebovale navedene vrednosti. V primeru, da katera od restavracij ne bi vsebovala kakšne od teh vrednosti, pa bi jih lahko poiskal na koncu, saj teh restavracij ni

veliko. Vseeno pa sem za demonstracijo manjkajoče vrednosti pridobil z nasprotnim geokodiranjem.

Ker je število manjkajočih koordinat 2.161, število manjkajočih naslovov pa le 22, sem izvedel nasprotno geokodiranje. Pri restavracijah, ki ne vsebujejo naslovov, in restavracijah, katerih naslov se ne preslika pravilno na zemljevidu, sem pridobil naslov s pomočjo koordinat. Prav tako sem naslove potreboval tudi za vizualizacijo, s katero sem v nadaljevanju prikazal lokacije najbolje ocenjenih restavracij.

Na začetku sem s pomočjo prikaza na zemljevidu identificiral restavracije, ki vsebujejo naslov, vendar se ta preslika na napačno lokacijo. Takšni restavraciji sta 2 in sta prikazani na sliki 16 na zemljevidu v Združenih državah Amerike ter sta izpisani v tabeli na desni strani. Točki sta zelo blizu, zato se na zemljevidu vidita kot 1 točka. Kot lahko vidimo v tabeli na desni strani, sta naslova precej nenatančna, saj ne vsebujeta podatka o ulici, ampak le o mestu samem. Ob prikazu teh 2 točk s koordinatami sem ugotovil, da se oba para koordinat pravilno preslikata v Italijo. Zaradi pravilnih koordinat sem za ti 2 restavraciji v nadaljevanju lahko izvedel nasprotno geokodiranje.

Slika 16: Restavracija z napačno preslikanima naslovoma



Vir: lastno delo.

Za tem sem identificiral restavracije, ki ne vsebujejo naslova. Prikazane so na sliki 17.

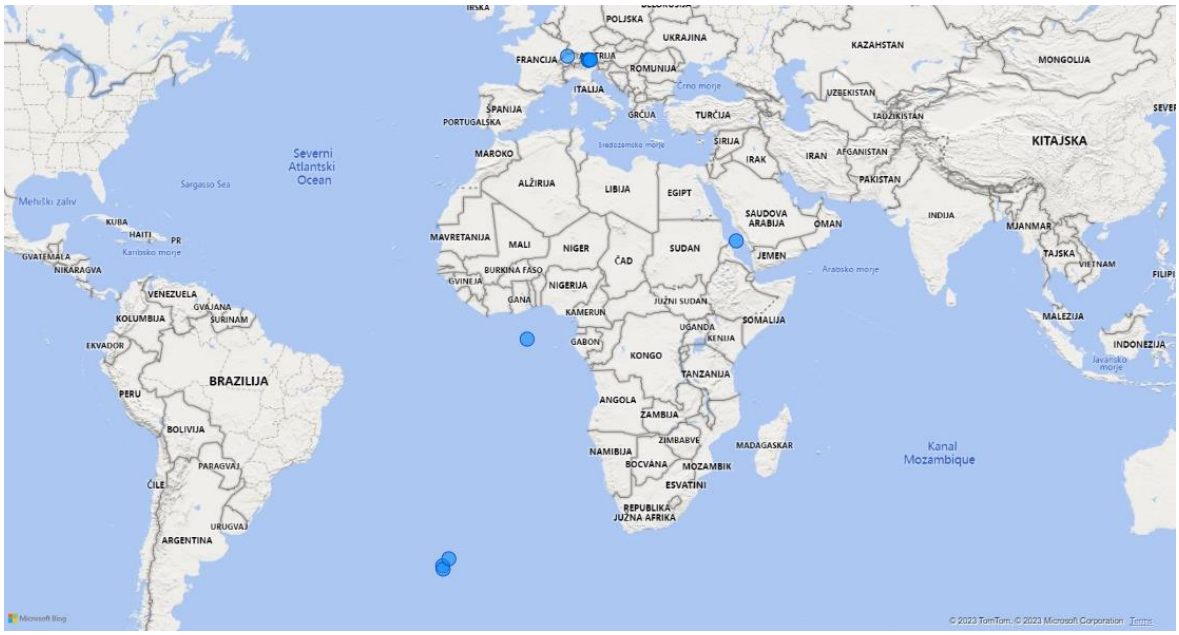
Slika 17: Restavracije brez naslova

restaurant_name	country	region	address	latitude	longitude	avg_rating	food	service	value	atmosphere
Chalet Bosco Difesa	Italy	Basilicata		-40,01399	-15,98272	4	4	4	4	4
Hofschanke Niedristhof	Italy	Trentino-South Tyrol		46,793415	12,016966	4,5	4,5	4	4,5	4,5
Die Muhle	Italy	Trentino-South Tyrol		46,85105	11,621282	4,5	4,5	4	4	4,5
Albergo Oberraut Ristorante Sas	Italy	Trentino-South Tyrol		46,81239	11,96225	4,5	4,5	4,5	4	4,5
Osteria TerraMasci nuova gestione	Italy	Apulia		18,35394	39,809677	4	4,5	4,5	4	
Ristorante dell'Agriturismo Tivoli	Italy	Sicily		-38,50566	-14,91184	3,5				
Bar da Tony	Italy	Umbria		0	0					
La Taverna del Buongustaio	Italy	Calabria		-39,55034	-16,07224	4				
Bar Marmorata	Italy	Lazio		47,295788	7,70888	4				
Rifugio Urbano	Italy	Lombardy		0	0	5				
Buschenschank Trinnerhof - Osteria Contac	Italy	Trentino-South Tyrol		46,7412	11,661479	5				
Agriturismo Montebeltrano	Italy	Calabria		3919718	16294022	4,5				

Vir: lastno delo.

Po preverjanju, kako se koordinate s slike 17 prikažejo na zemljevidu, sem ugotovil, da se jih kar precej preslika narobe. Lahko jih vidimo na sliki 18, kjer so 4 restavracije pravilno prikazane v Italiji, 1 v Švici in 6 sredi morja. V morju lahko vidimo samo 5 točk, saj imata restavraciji, ki sta na sliki 17 označeni z rdečo barvo, enake koordinate. Z oranžno barvo označena restavracija na zemljevidu ni prikazana, saj ima neveljavne koordinate.

Slika 18: Koordinate restavracij brez naslova



Vir: lastno delo.

Lokacije 4 restavracij, ki so prikazane v Italiji, sem preveril na spletni strani Tripadvisor in ugotovil, da so res prikazane na pravih lokacijah. Zato bom nad njimi lahko izvedel nasprotno geokodiranje. Te restavracije so na sliki 17 označene z modro barvo.

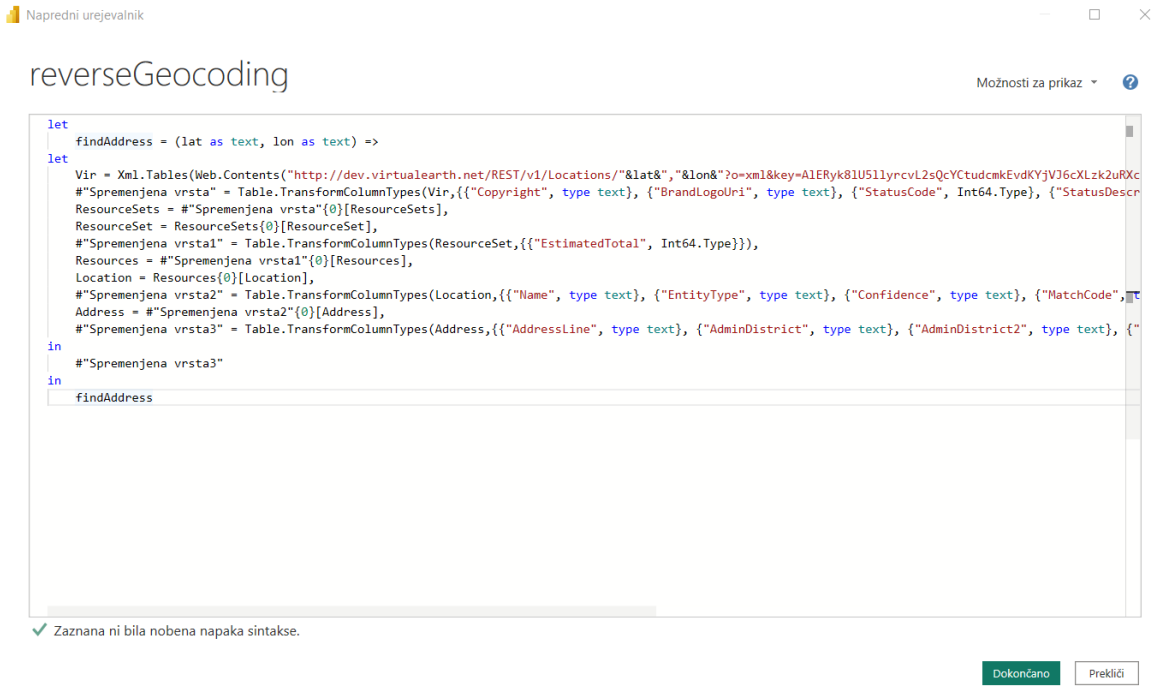
Nadaljeval sem z identifikacijo težav, ki jih ima preostalih 8 parov koordinat. Te težave sem identificiral in popravil v nadaljevanju v poglavjih od 6.5.7 do 6.5.10. Pri 2 restavracijah, ki imata koordinate enake 0, sem naslove ročno pridobil s spletne strani Tripadvisor, tako da nasprotnega geokodiranje nad njima nisem izvedel. Pri restavraciji, ki hrani lokacijo v Švici

in je na sliki 17 označena s črno barvo, pa sem ugotovil, da je napačna lokacija shranjena tudi na spletni strani Tripadvisor. Zato sem omenjeno restavracijo izpustil iz raziskave.

Po popravku napačnih in neveljavnih koordinat sem nadaljeval z nasprotnim geokodiranjem. Izvedel sem ga s pomočjo navodil, ki sem jih pridobil na svetovnem spletu (KAISPE, brez datuma; Hari's BI, 2018). Spodaj opisan postopek je izveden s pomočjo teh navodil.

Na začetku sem na spletni strani Bing Maps Dev Center ustvaril uporabniški račun in kreiral api ključ, ki sem ga kasneje uporabljal pri klicanju storitve za nasprotno geokodiranje (Microsoft, 2022). Nato sem s pomočjo navodil ustvaril funkcijo po meri, ki s pomočjo prenosa predstavitvenega stanja (angl. Representational state transfer – REST) pridobi naslov za dane koordinate. Prikazana je na sliki 19.

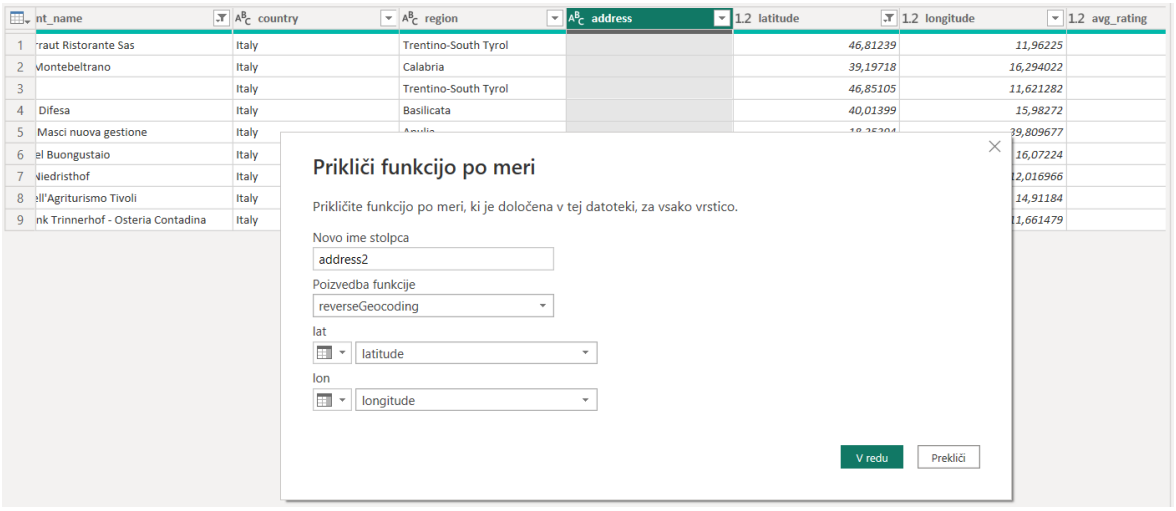
Slika 19: Funkcija za nasprotno geokodiranje



Vir: lastno delo.

Funkcija kot parametra sprejme zemljepisno širino in dolžino, vrne pa naslov. Za pridobivanje novih naslovov za dane restavracije sem ustvaril nov začasni stolpec address2. Funkcija je kot parametra prejela vrednosti obstoječih stolpcev latitude in longitude, ustvarjen naslov pa zapisala v nov stolpec. Nastavitve, ki sem jih funkciji moral določiti, so prikazane na sliki 20.

Slika 20: Nastavljanje klicev funkcije



Vir: lastno delo.

Na sliki 21 lahko vidimo naslove, ki so rezultat nasprotnega geokodiranja. Pri vseh 11 restavracijah sem pridobil pravilne naslove, ki so na zemljevidu prikazani na enakih lokacijah kot pripadajoče koordinate. Pridobljene naslove sem zapisal v atribut address.

Slika 21: Rezultati nasprotnega geokodiranja

	restaurant_name	address	latitude	longitude	address2.FormattedAddress
1	Buschenschank Trinnerhof - Osteria Contadina		46.7412	11.661479	Michael Pacherstraße 76, 39040 Naz-Sciaves Bolzano, Italy
2	Hofschanke Niedristhof		46.793415	12.016966	Fraktion Aschbach 2, 39030 Perca Bolzano, Italy
3	Albergo Oberraut Ristorante Sas		46.81239	11.96225	Amaten 2/A, 39031 Bruneck Bolzano, Italy
4	Die Muhle		46.85105	11.621282	Jochtalstraße 4, 39037 Rio di Pusteria Bolzano, Italy
5	Ristorante dell'Agriturismo Tivoli		38.50566	14.91184	Via Quattara 17, 98055 Lipari Messina, Italy
6	Tra Le Righe	20131 Milan Italy	45.485275	9.229938	Via Teodosio 36, 20131 Milan Milan, Italy
7	Ittiturismo all'Amo Sestri ponente	16122, Genoa Italy	44.42568	8.848253	Via Alfredo D'Andrade 7, 16154 Genoa Genoa, Italy
8	La Taverna del Buongustaio		39.55034	16.07224	SS283, 87013 Fagnano Castello Cosenza, Italy
9	Agriturismo Montebeltrano		39.19718	16.294022	Strada Statale 108, 87040 Belsito Cosenza, Italy
10	Chalet Bosco Difesa		40.01399	15.98272	SP46, 85040 Castelluccio Superiore Potenza, Italy
11	Osteria TerraMasci nuova gestione		39.809677	18.35394	Via Ennio Quinto in Località Leuca 182, 73040 Castrignano del Capo Le...

Vir: lastno delo.

6.5.7 Zamenjani zemljepisna širina in dolžina

Restavracija, ki je na sliki 17 označena z rjavo barvo, se nahaja sredi morja. Opazil sem, da sta koordinati nekoliko nesmiselni glede na preostale koordinate in bi bili veliko bolj smiselni, če bi bili zamenjani. Gre za napako, opisano v poglavju Napačna lokacija, ki pravi, da sta lahko koordinati zamenjani. Zato sem ju zamenjal in ugotovil, da se je točka pravilno preslikala v Italijo. To je prikazano na sliki 22. Za tem sem pravilnost novih koordinat potrdil tudi na spletni strani Tripadvisor.

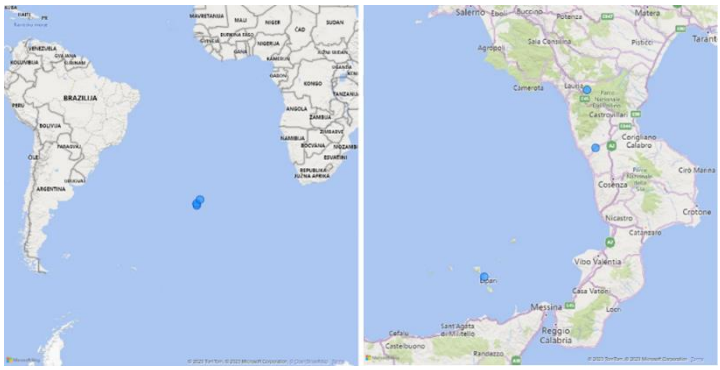
Slika 22: Napačen (levo) in popravljen (desno) vrstni red koordinat

Vir: lastno delo.

6.5.8 Napačen predznak koordinat

Pri restavracijah, ki so na sliki 17 označene z zeleno barvo, sem ugotovil, da gre za težavo z napačnim predznakom, ki je opisana v poglavju Napačna lokacija. Koordinate bi s pozitivnim predznakom bile blizu ostalim restavracijam. Zato sem koordinatam popravil predznak in te restavracije ponovno prikazal na zemljevidu. Tokrat so se točke prikazale v Italiji. To je prikazano na sliki 23. Po preverjanju na spletni strani Tripadvisor sem pravilnost novih lokacij potrdil.

Slika 23: Napačen (levo) in pravilen (desno) predznak koordinat



Vir: lastno delo.

6.5.9 Koordinate izven dovoljenega obsega

Zemljepisna širina in dolžina restavracije, ki je na sliki 17 označena z oranžno barvo, sta preveliki in izven dovoljenega obsega. To težavo sem opisal v poglavju Neveljavne koordinate. Opazimo lahko tudi, da se koordinate te restavracije razlikujejo od koordinat preostalih restavracij v tem, da ne vsebujejo decimalnih vejic. Koordinate so bile lahko napačne že na sami spletni strani ali pa je do njih prišlo pri spletnem strganju ali nalaganju v Power BI Desktop. Koordinatam sem decimalno vejico dodal na smiselno mesto, tako da

so bile podobne koordinatam preostalih restavracij. Zemljepisno širino sem spremenil iz 3919718 na 39,19718, zemljepisno dolžino pa iz 16294022 na 16,294022. Po popravku se je restavracija pravilno prikazala v Italiji. To je vidno na sliki 24. Pravilnost nove lokacije sem potrdil tudi na spletni strani Tripadvisor.

Slika 24: Popravljene koordinate izven obsega

Vir: lastno delo.

6.5.10 Zemljepisna širina in dolžina imata vrednost nič

Pri 2 restavracijah, ki sta na sliki 17 označeni z rdečo barvo, sem opazil, da imata zemljepisno širino in zemljepisno dolžino enako 0. Ta lokacija ni pravilna, saj se ne nahaja v Italiji, ampak sredi morja. Gre za napako, opisano v poglavju Napačna lokacija. Ker nisem imel na voljo nobenih podatkov, s katerimi bi si lahko pomagal oz. iz njih dinamično pridobil naslov ali koordinate, sem uporabil vrednosti atributa restaurant_link in z njuno pomočjo obiskal profila obeh restavracij na spletni strani Tripadvisor. Na teh sem pridobil naslova in ju ročno dodal v podatkovni vir.

6.5.11 Izpeljanke in agregacija

V poglavju Preoblikovanje sem opisal, da atributi, ki jih imamo na voljo, niso vedno dovolj za reševanje našega problema. Zato moramo kdaj iz obstoječih atributov ustvariti nove. Kot enega od načinov sem omenil agregacijo, s katero vnose smiselno združujemo. V mojem primeru sem dodaten stolpec potreboval pri prikazovanju tega, katera lastnost je v posamezni regiji najslabše ocenjena. Potrebna sta bila 2 nova stolpca: ime najslabše ocenjene lastnosti in njena povprečna ocena. Zato sem podatkovni vir kopiral in nad kopijo izvedel agregacijo, pri kateri sem vnose združeval glede na atribut region, nad stolpci food, value, service in

atmosphere pa izračunal povprečje. Nato sem nad to tabelo ustvaril nova stolpca. Formula za kreiranje stolpca s povprečjem najslabše ocenjene lastnosti je prikazana na sliki 25.

Slika 25: Ustvarjanje stolpca s povprečjem najslabše ocenjene lastnosti

```
min_avg =
VAR ratings = {
    worst_rating_by_region[avg_food],
    worst_rating_by_region[avg_service],
    worst_rating_by_region[avg_value],
    worst_rating_by_region[avg_atmosphere]
}
RETURN
    MINX ( ratings, [Value] )
```

Vir: lastno delo.

Drugi stolpec hrani ime lastnosti z najslabšo povprečno oceno. Formula je prikazana na sliki 26.

Slika 26: Ustvarjanje stolpca z imenom lastnosti z najslabšo povprečno oceno

```
min_avg_name =
SWITCH(
    worst_rating_by_region[min_avg],
    worst_rating_by_region[avg_food], "Hrana",
    worst_rating_by_region[avg_service], "Postrežba",
    worst_rating_by_region[avg_value], "Vrednost",
    worst_rating_by_region[avg_atmosphere], "Vzdušje"
)
```

Vir: lastno delo.

Novo ustvarjena tabela ima 20 vrstic in 6 stolpcev. Prikazana je na sliki 27.

Slika 27: Nova tabela z najslabšimi povprečji

region	avg_food	avg_service	avg_value	avg_atmosphere	min_avg	min_avg_name
Friuli-Venezia Giulia	4,0811174688657	3,96244131455399	3,90628149143433	3,89236319903788	3,89236319903788	Vzdušje
Trentino-South Tyrol	4,137031408308	4,02681944094686	3,93161746151905	4,04333490343853	3,93161746151905	Vrednost
Emilia-Romagna	4,08379963898917	3,94253287695911	3,89269632564841	3,82776600945087	3,82776600945087	Vzdušje
Molise	4,13500931098696	3,97769516728625	4,03252788104089	3,93295019157088	3,93295019157088	Vzdušje
Sicily	4,15405340677763	4,02597014925373	4,01010765550239	3,9201226993865	3,9201226993865	Vzdušje
Lazio	4,08959220101606	3,99157534246575	3,92973306800247	3,85981846075173	3,85981846075173	Vzdušje
Marche	4,15353697749196	4,00053475935829	4,00655080213904	3,90906909788868	3,90906909788868	Vzdušje
Basilicata	4,18279022403259	4,03857868020305	4,05736040609137	3,94130434782609	3,94130434782609	Vzdušje
Umbria	4,19242902208202	4,08003597122302	4,06393001345895	3,9992125984252	3,9992125984252	Vzdušje
Lombardy	4,00515568649362	3,90307622136631	3,83002028397566	3,81343498273878	3,81343498273878	Vzdušje
Piedmont	4,06576021027093	3,97127133191962	3,9236834125139	3,87605971249539	3,87605971249539	Vzdušje
Sardinia	4,11712277178006	4,00874344241819	3,9472236118059	3,87130392632089	3,87130392632089	Vzdušje
Apulia	4,11203319502075	3,96143231682075	3,96393533232002	3,90782196470898	3,90782196470898	Vzdušje
Tuscany	4,14857119793348	4,02939993564994	3,99625603864734	3,93149516770893	3,93149516770893	Vzdušje
Liguria	4,0925321888412	3,96443228454172	3,92360992301112	3,87299703264095	3,87299703264095	Vzdušje
Abruzzo	4,17166921898928	4,00152346130408	4,03740458015267	3,91234498308906	3,91234498308906	Vzdušje
Calabria	4,16577450652392	4,02591973244147	4,05070281124498	4,00387596899225	4,00387596899225	Vzdušje
Veneto	4,07695702056306	3,97141097818438	3,89297364013048	3,89036725801432	3,89036725801432	Vzdušje
Campania	4,1418096723869	3,99968899025503	3,98763764803657	3,9328845369237	3,9328845369237	Vzdušje
Aosta Valley	4,06642941874259	3,98576512455516	3,85900473933649	4,02434077079107	3,85900473933649	Vrednost

Vir: lastno delo.

Pri primeru prikazovanja najbolje ocenjenih restavracij v posamezni regiji sem v nadaljevanju izbral najbolje ocenjene restavracije, ki imajo največje število ocen. Tudi za ta primer sem podatkovni vir kopiral. Za vsako regijo sem pripadajoče restavracije razvrstil padajoče glede na povprečno oceno in število ocen. Novi vir je vseboval 3 restavracije za vsako regijo, kar znaša 60 restavracij.

6.5.12 Revizija informacij

Pri preoblikovanju podatkov sem opisal, da je na koncu čiščenja podatkov te treba preveriti s prvotnimi in s tem preveriti, da ni prišlo med čiščenjem do kakšne napake. To preverjanje sem v nadaljevanju storil tako, da sem preveril, ali je število vrstic skladno glede na procese, ki sem jih nad podatki izvajal. Število vrstic po čiščenju znaša 224.700.

Pred začetkom čiščenja je bilo v podatkovnem viru 1.083.349 restavracij. Pri začetnem odstranjevanju vrstic sem odstranil 858.633 restavracij, za katere sem bil prepričan, da niso iz Italije. Pri čiščenju podatkov sem odstranil 2 restavraciji, ki sta bili podvojeni, 13 restavracij, ki so bile bodisi iz tujine bodisi napačni vnosi in 1 restavracijo, ki je tako kot v podatkovnem viru tudi na spletni strani hranila napačno lokacijo. Po odštevanju vseh odstranjenih vrstic od prvotnega števila ugotovimo, da je število vrstic enako 224.700, kar se ujema s preoblikovanim podatkovnim virom.

6.6 Implementacija vizualizacij

6.6.1 Povprečna ocena restavracij v posamezni regiji

V tem poglavju sem reševal prvi problem iz poglavja Predstavitev problema. V njem sem izdelal vizualizacijo, ki prikazuje povprečno oceno restavracij v posamezni regiji.

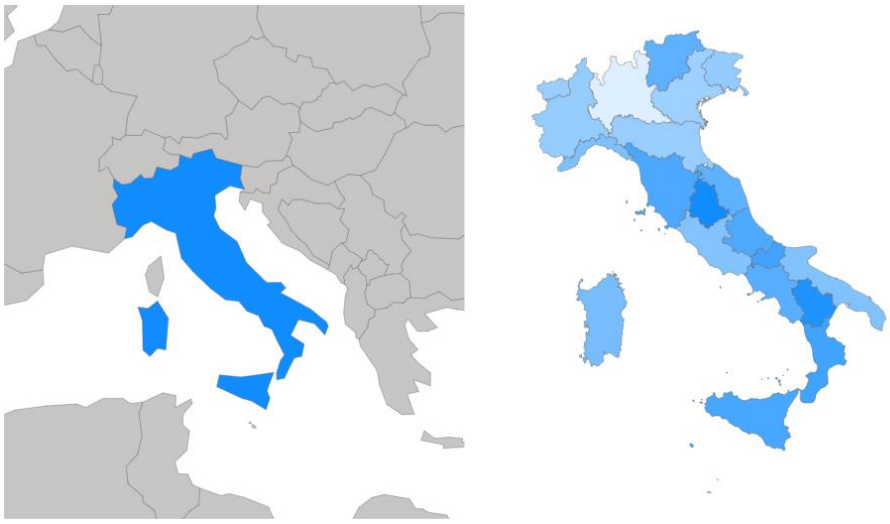
6.6.1.1 Predhoden razmislek

Kot sem omenil pri kakovosti vizualizacij, se je treba pred začetkom izdelave vizualizacij vprašati, katere informacije oz. sporočilo želimo prikazati in na kakšen način jih bomo prikazovali. Cilj tega zemljevida je primerjava povprečnih ocen restavracij med posameznimi regijami. V poglavju Kakovost vizualizacij z zemljevidi sem pri izbiri pravega zemljevida opisal, da moramo zemljevid izbirati tudi na podlagi tega, ali so naši prostorski podatki točke, črte ali poligoni. Ker je treba prikazati primerjavo vrednosti med poligoni in ne točnih lokacij restavracij, sem za ta primer uporabil horoplet zemljevid, ki je temu namenjen. Uporabil sem tudi tabelo in stolpčni diagram.

6.6.1.2 Izbira ustrezne ravni

V poglavju Izbira ustrezne ravni pri horoplet zemljevidih sem opisal, da je pri uporabi horoplet zemljevida treba določiti, na kateri ravni bomo prikazovali območja. V tem primeru nas zanima samo primerjava med regijami, zato bi bili primerjavi med državami ali pa med občinami nesmiselni. Pri primerjavi med državami bralcu ne bi sporočali informacije, ki ga zanima, pri primerjavi med občinami, pa bi informacijo težje razbral. Primer neustrezne in ustrezne ravni je prikazan na sliki 28, kjer je na levi strani prikazana primerjava med državami (v mojem primeru hranim podatke samo za eno), na desno pa ustrezna primerjava med regijami.

Slika 28: Izbira neustrezne (levo) in ustrezne ravni (desno)



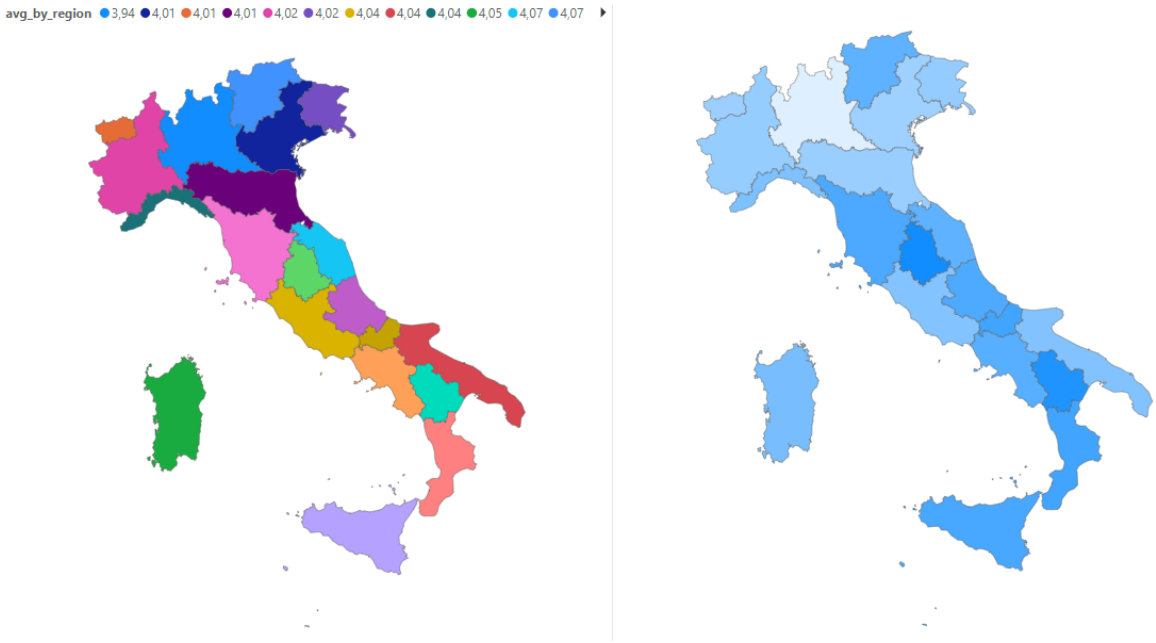
Vir: lastno delo.

Uporabljen privzeti horoplet zemljevid v Power BI Desktopu ne vsebuje posameznih evropskih držav. Zato sem za levi zemljevid na sliki 28 s spleta naložil datoteko tipa json, ki hrani podatke o mejah vseh držav (Eldersveld, brez datuma).

6.6.1.3 Uporaba barv

Tako v poglavju Kakovost vizualizacij kot tudi v poglavju Kakovost vizualizacij z zemljevidi sem v podpoglavjih o barvah opisal, da moramo pri prikazovanju razpona vrednosti uporabiti isto barvo z različnimi odtenki in ne različnih barv, saj v slednjem primeru uporabnik ne ve, katera barva pomeni večjo in katera manjšo vrednost. Primer uporabe različnih barv za različne povprečne ocene lahko vidimo na levem zemljevidu na sliki 29, kjer je zelo težko razbrati, katera regija ima večjo in katera nižjo povprečno oceno. Na desnem zemljevidu pa je uporabljena ista barva, le da imajo višje vrednosti temnejši odtenek, nižje vrednosti pa svetlejši odtenek. Na desnem zemljevidu lahko takoj vidimo, katera regija ima višjo in katera nižjo vrednost.

Slika 29: Izbira različnih barv (levo) in različnih odtenkov iste barve (desno)



Vir: lastno delo.

V poglavju Barve sem tudi omenil, da je priporočljivo uporabiti barve, ki jih razumejo tudi barvno slepi ljudje. Zato sem končni zemljevid iz poglavja 6.6.1.9 preveril s spletnim orodjem Coblis in ugotovil, da je zemljevid razumljiv pri različnih barvnih slepotah (Colblindor, brez datuma).

6.6.1.4 Izbira ustreznega barvnega intervala

V poglavju Izbira ustreznega barvnega intervala sem opisal, da moramo pri uporabi horoplet zemljevidov uporabiti ustrezen barvni interval, saj premajhen interval lahko vodi v preveč različnih skupin/odtenkov, ki jih je med seboj težko ločiti, prevelik interval pa v to, da preveč vrednosti pripada enakemu odtenku. Na to sem bil pozoren tudi pri svojem zemljevidu. Po preizkušanju različnih intervalov sem ugotovil, da je pri mojem zemljevidu najbolj primeren interval 0.05. Pri tem intervalu se razlike med regijami lepo razločijo. To je vidno pri končnem rezultatu v poglavju 6.6.1.9.

6.6.1.5 Izbira ustreznega razpona vrednosti pri samodejnem dodeljevanju barv

V Power BI Desktopu imamo pri horoplet zemljevidu tudi možnost samodejnega dodeljevanja odtenkov posameznim vrednostim. Pri uporabi je možno določanje, med katero najmanjšo in največjo vrednost naj se odtenki barve razporedijo. Tu lahko pride do enake težave kot pri izbiri neustreznega barvnega intervala, tj. da so si različne vrednosti med seboj preveč podobne ali pa da se med seboj preveč razlikujejo.

Za pridobitev zemljevida z lepo vidnimi razlikami med regijami sem preizkušal različne razpone vrednosti. Najmanjša in največja možna ocena, ki ju atribut avg_rating lahko zaseda, znašata 1 in 5. Nekatere od restavracij so ti 2 oceni tudi prejele. Najmanjša in največja povprečna ocena glede na regijo pa znašata 3,94 in 4,15. Na sliki 30 sem horoplet zemljevid prikazal s 3 različnimi razponi vrednosti. Na levem zemljevidu sem barvo razpotegnil med vrednostma 1 in 5, kar pomeni, da ima vrednost 1 najsvetlejši odtenek, vrednost 5 pa najtemnejši odtenek. Ker so si povprečne vrednosti regij zelo blizu in zasedajo majhen del tega razpona, tj. od 3,94 do 4,15, so si odtenki med seboj zelo podobni in jih med seboj ni mogoče ločiti. Na sredinskem zemljevidu sem razpon nekoliko zmanjšal, vendar je bilo odtenke med seboj še vedno težko ločiti. Na desnem zemljevidu pa sem najmanjšo in največjo vrednost razpona barve enačil z najnižjo in najvišjo povprečno oceno in tako prišel do zemljevida, na katerem se hitreje opazi, katere regije so bolj svetle in katere bolj temne.

Slika 30: Horoplet zemljevid z različnimi razponi vrednosti

Vir: lastno delo.

6.6.1.6 Velikost besedila

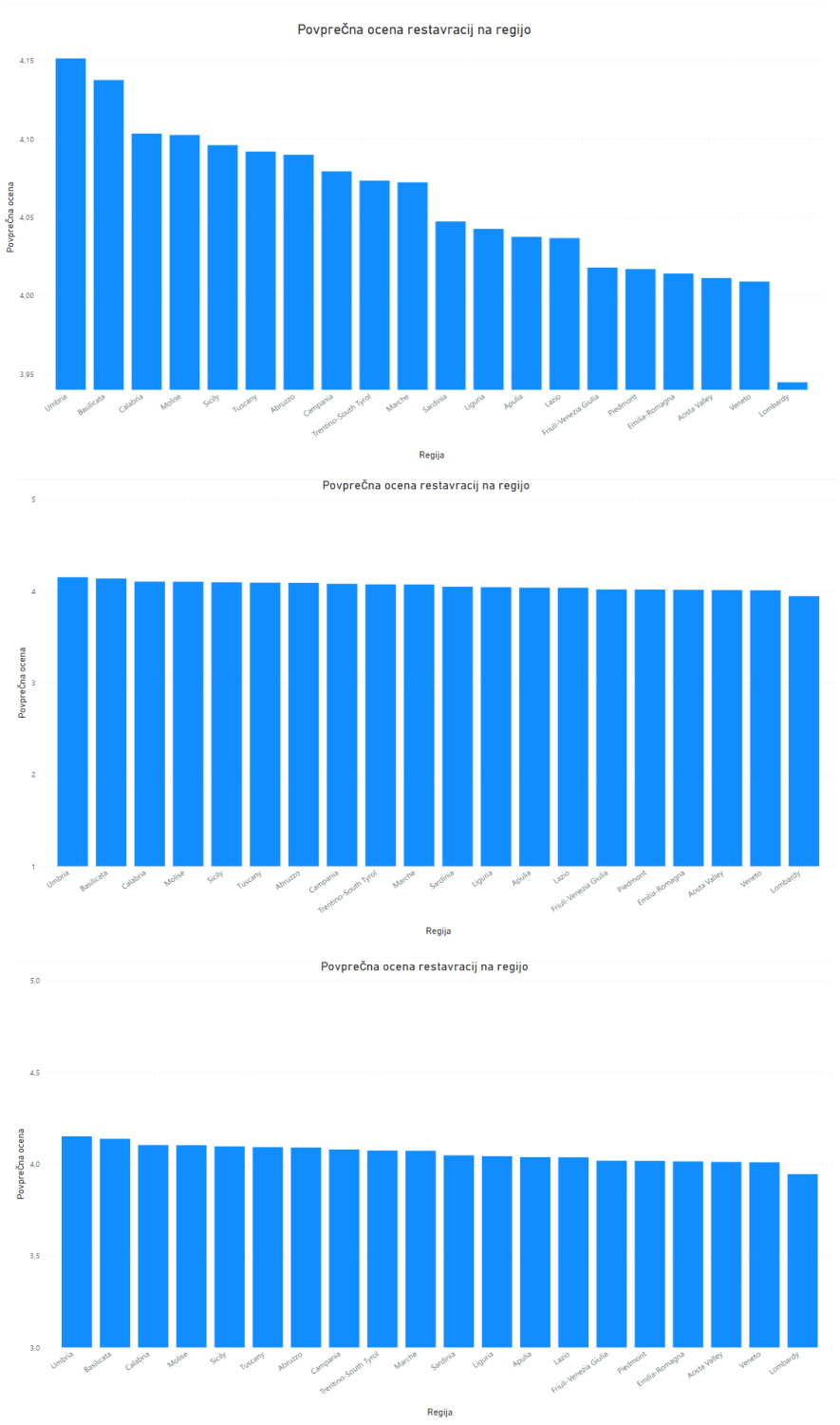
Moj horoplet zemljevid vsebuje besedilo v naslovu in legendi. Poskrbel sem, da sta naslov in besedilo legende dovolj velika, da ju bralec lahko brez težav prebere. To je vidno pri končnem rezultatu v poglavju 6.6.1.9.

6.6.1.7 Koordinatne osi

Pri kakovosti vizualizacij z zemljevidi sem v podpoglavju Koordinatne osi opisal, da je v večini primerov dobro, da pri številskih vrednostih koordinatnih osi začenjamo z vrednostjo nič, saj se nam v nasprotnem primeru lahko zgodi, da sporočamo napačno informacijo in

prikazane razlike med vrednostmi izgledajo večje, kot so v resnici. Na sliki 31 sem stolpčni diagram prikazal s 3 različnimi razponi vrednosti na osi y.

Slika 31: Stolpčni diagrami z različnim minimumom in maksimumom povprečne ocene na osi y



Vir: lastno delo.

Na sliki 31 os x predstavlja posamezne regije, os y pa povprečno oceno restavracij v regiji. Na zgornjem diagramu je razpon osi med 3.94, tj. najmanjša povprečna ocena, in 4.16, tj. največja povprečna ocena, na srednjem diagramu med 1, tj. najmanjša možna ocena, in 5, tj. največja možna ocena, na spodnjem diagramu pa med 3 in 5. Na vseh 3 diagramih so vrednosti razvrščene z leve proti desni od največje do najmanjše.

Na zgornjem diagramu lahko vidimo, da je precej zavajajoč. V resnici se največja vrednost na levi od najmanjše na desni razlikuje za samo za okoli 0,22, kar znaša približno 4 % celotnega možnega razpona. Na diagramu izgleda, da je največja vrednost veliko večja od najmanjše. Če primerjamo največjo vrednost, ki jo ima regija Umbria, na diagramu izgleda, kot da je približno 2-krat večja od vrednosti v Sardiniji, čeprav to ni res. To zavajanje je odpravljeno na sredinskem diagramu, kjer je razpon med najmanjšo in največjo možno vrednostjo. Ker pa so si vrednosti zelo blizu in zasedajo le približno 4 % razpona, so te tudi narisane zelo blizu in je med njimi težko ločiti. Zato sem na spodnjem diagramu kot začetno vrednost uporabil 3. Tako so razlike med vrednostmi nekoliko bolj očitne, vseeno pa ne tako velike, da bi bile zavajajoče. Vsi 3 diagrami pravilno vsebujejo enakomerne razmike med vrednostmi na osi y, kot je to opisano v omenjenem podpoglavju Koordinatne osi.

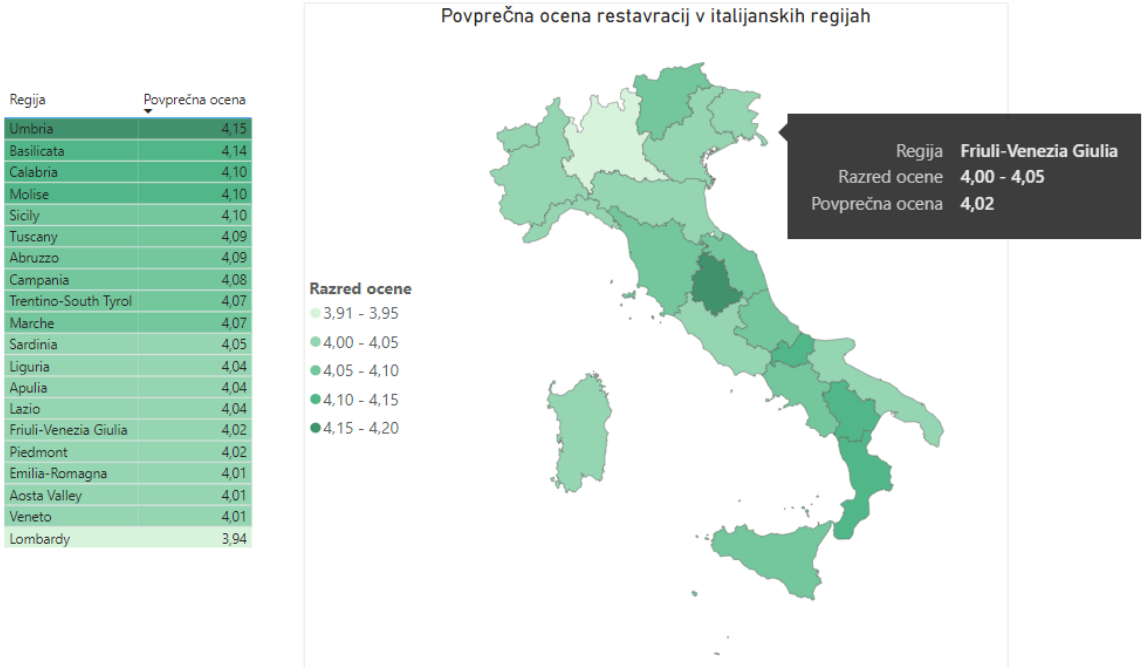
6.6.1.8 Interaktivnost

V poglavju Kakovost vizualizacij z zemljevidi sem v podpoglavju Interaktivnost opisal interaktivnost v vizualizacijah. To sem dodal tudi zemljevidu, tako da se uporabniku s premikom miške na posamezno regijo prikaže ime regije, povprečna ocena v tej regiji in v kateri razred ocena spada. Poleg zemljevida sem ustvaril tabelo, ki vsebuje 2 stolpca: ime regije in njeno povprečno oceno. Iz te tabele lahko uporabnik brez interaktivnosti razbere, kolikšno povprečno oceno ima katera od regij in kako so te razvrščene od največje do najmanjše. To je prikazano tudi pri končnem rezultatu v poglavju 6.6.1.9.

6.6.1.9 Ravnotežje elementov in končni rezultat

Pri končnem rezultatu, prikazanem na sliki 32, sem pri kakovosti vizualizacij z zemljevidi upošteval opisano postavitev elementov in ravnotežje med njimi. Tako sem naslov zemljevida, ki nam pove, kaj zemljevid predstavlja, postavil na vrh, da ga uporabniki najprej preberejo. Zemljevid sem postavil približno na sredino, da je izpostavljen. Ob strani je dodana tudi legenda, s katero si uporabnik lahko pomaga pri branju. Tabela je prikazana ob strani zemljevida kot ločena vizualizacija.

Slika 32: Povprečna ocena restavracij v italijanskih regijah



Vir: lastno delo.

6.6.2 Najslabše ocenjena lastnost v posamezni regiji

V tem poglavju sem reševal drugi problem iz poglavja Predstavitev problema. V njem sem izdelal vizualizacijo, ki prikazuje katera lastnost restavracij je v posamezni regiji najslabše ocenjena.

6.6.2.1 Predhoden razmislek

Cilj tega zemljevida je identifikacija najslabše ocenjene lastnosti restavracij v posamezni regiji. Ker gre za prikaz različnih kategorij, sem uporabil zemljevid s kategorijami. Za ta primer sem uporabil stolpca min_avg in min_avg_name, ki sem ju ustvaril pri čiščenju podatkov.

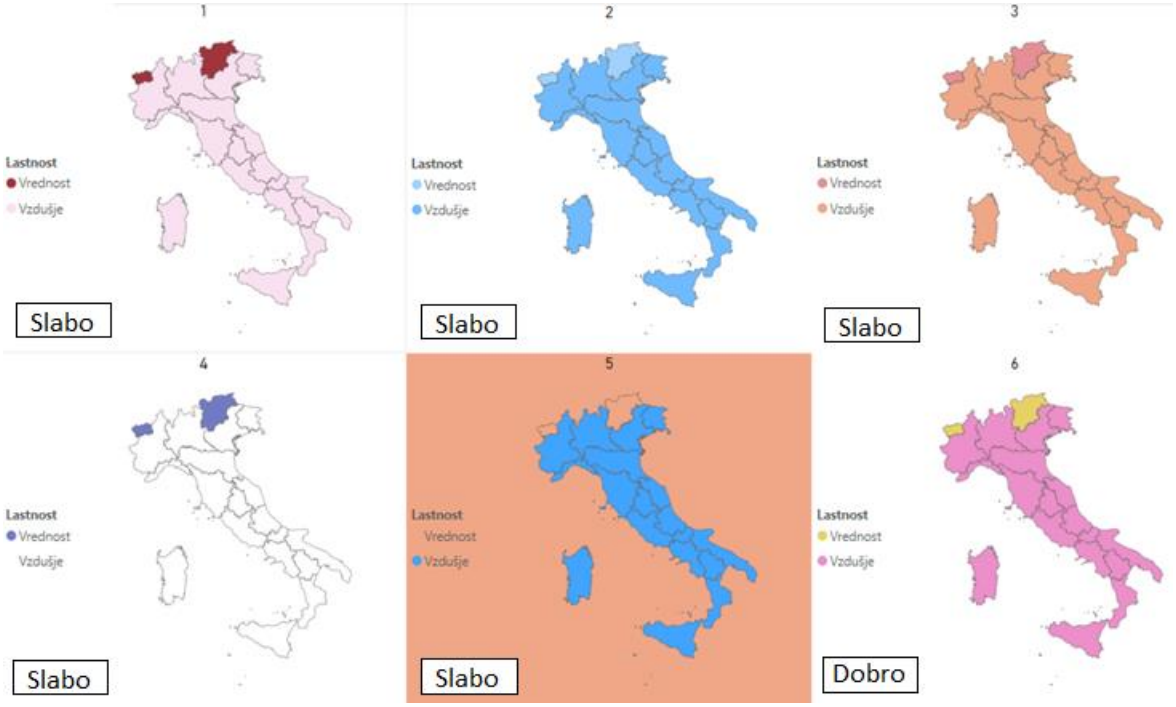
6.6.2.2 Barve

V poglavjih Kakovost vizualizacij in Kakovost vizualizacij z zemljevidi sem opisal napotke pri uporabi barv, ki se lahko aplicirajo na uporabo zemljevida s kategorijami. To so:

– Ni dobro, da katero od barv preveč poudarimo, saj lahko zasenči ostale. To se vidi na sliki 33 na zemljevidu 1. Regiji s kategorijo »Vrednost« sta bolj vpadljivi in zato lahko vodita do napačnega sklepanja, da sta bolj pomembni od ostalih regij.

- Ko prikazujemo različne kategorije, je dobra uporaba popolnoma različnih barv, saj se te bolj razlikujejo kot odtenki iste barve. To se vidi na sliki 33 na zemljevidu 2, kjer je za obe kategoriji uporabljena ista barva z različnima odtenkoma. Regije z različno kategorijo zato težje ločimo med sabo.
- Ob prikazu zemljevida s kategorijami ne smemo prikazovati preveč kategorij oz. barv, saj je zato zemljevid težje berljiv in mora uporabnik večkrat preverjati legendo. V mojem primeru te težave ni, saj prikazujem le 2 različni kategoriji.
- Ni priporočljiv prikaz z barvami, ki imajo nizek medsebojni kontrast, saj potem težje ločimo med posameznimi elementi. To je vidno na sliki 33 na zemljevidu 3, kjer sem za različni kategoriji uporabil različni barvi, vendar imata ti 2 nizek medsebojni kontrast in zato med njima težko ločimo.
- Ospredje (zemljevid) mora biti dovolj izrazito poudarjeno in ločeno od ozadja. Na sliki 33 je na zemljevidu 4 prikazan slab primer, ko je barva ospredja enaka barvi ozadja. Sicer v ozadju ni zemljevida in ima ospredje obrobo, ki ga ločuje od ozadja, vendar lahko dobimo občutek, da sta ospredje in ozadje nekoliko povezana. Na zemljevidu 5 je prikazana podobna težava, le da sem tu tudi ozadju dodal barvo in je zato regiji s kategorijo »Vrednost« težko ločiti od ozadja.
- Priporočljivo je, da uporabimo barve, ki jih razumejo tudi barvno slepi ljudje. Zato sem uporabil integrirano barvno temo aplikacije Power BI Desktop, ki je prijazna barvno slepim. Ta barvna tema je uporabljena tudi pri končnem rezultatu v poglavju 6.6.2.5. Da je zemljevid razumljiv pri različnih barvnih slepotah, sem preveril s spletnim orodjem Coblis (Colblindor, brez datuma).

Slika 33: Primeri slabih in dobrega zemljevida s kategorijami



Vir: lastno delo.

Na sliki 33 je na zemljevidu 6 prikazana dobra uporaba barv. Med kategorijama je lahko ločiti, nobena od barv ne izstopa, ospredje pa je vidno ločeno od ozadja.

6.6.2.3 Interaktivnost

Tako kot pri horoplet zemljevidu sem tudi tu dodal interaktivnost, s katero lahko uporabnik s premikom miške na posamezno regijo vidi ime regije, ime lastnosti, ki je v tej regiji najslabše ocenjena, in koliko znaša ta ocena. To je vidno pri končnem rezultatu v poglavju 6.6.2.5.

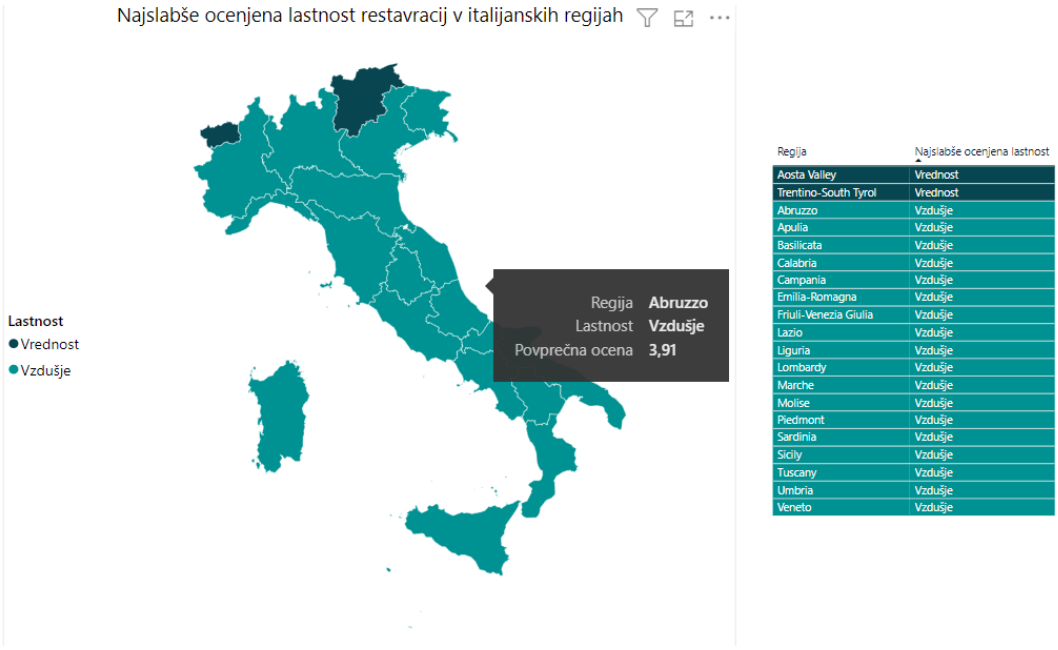
6.6.2.4 Velikost besedila

Prav tako kot pri horoplet zemljevidu tudi ta zemljevid vsebuje besedilo v naslovu in legendi. Poskrbel sem, da sta naslov in besedilo legende dovolj velika, da ju bralec lahko brez težav prebere. To je vidno pri končnem rezultatu v poglavju 6.6.2.5.

6.6.2.5 Ravnotežje elementov in končni rezultat

Prav tako kot pri horoplet zemljevidu sem tudi tukaj upošteval napotke za uravnoteženo postavitev elementov vizualizacije. Poleg zemljevida sem ustvaril tudi tabelo, ki vsebuje 2 stolpca: ime regije in njeno najslabšo ocenjeno lastnost. Iz te tabele lahko uporabnik brez interaktivnosti razbere, katera lastnost je v kateri regiji najslabša. To je prikazano na sliki 34.

Slika 34: Najslabše ocenjena lastnost restavracij v italijanskih regijah



Vir: lastno delo.

6.6.3 Prikaz lokacij najbolje ocenjenih restavracij v posamezni regiji

V tem poglavju sem reševal tretji problem iz poglavja Predstavitev problema. V njem sem prikazal lokacije in naslove treh najbolje ocenjenih restavracij v posamezni regiji.

6.6.3.1 Predhoden razmislek

Cilj tega zemljevida je prikaz točnih lokacij najbolje ocenjenih restavracij v posamezni regiji. Ker gre za prikaz točnih lokacij, sem uporabil zemljevid s točkami. Ker imajo povprečne ocene restavracij v viru diskreten interval 0,5 in niso zvezne, lahko več restavracij zaseda največjo povprečno oceno. Zato sem poleg povprečne ocene upošteval tudi atribut total_reviews_count. Za vsako regijo sem izbral 3 najbolje ocenjene restavracije, ki so prejele največ ocen. Za podatke sem uporabil novo tabelo, ustvarjeno pri čiščenju podatkov v podpoglavju Izpeljanke in agregacija.

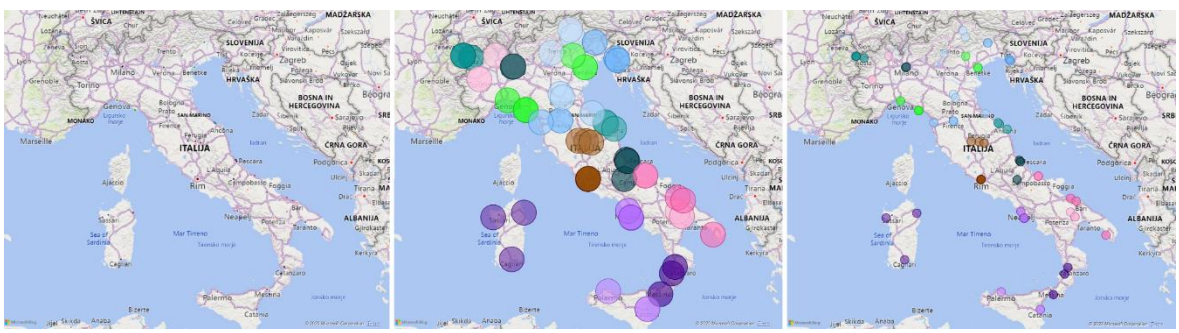
6.6.3.2 Barve

Točke restavracij sem barvno označil tako, da imajo barvo pripadajoče regije. Tako se lažje loči, katere točke pripadajo kateri regiji. Tako kot pri prejšnjih 2 primerih sem tudi pri tem zemljevidu uporabil barvno temo, ki je prijazna barvno slepim ljudem.

6.6.3.3 Izbira in velikost simbolov

Pri Kakovosti vizualizacij z zemljevidi sem v podpoglavju Izbira in velikost simbolov opisal, da simboli, ki jih uporabljamo kot točke, ne smejo biti preveč kompleksni in da morajo biti zaradi lažje berljivosti dovolj veliki. Za simbole sem uporabil privzete kroge, saj so ti preprosti za razumevanje, njihova vloga pa je, da podajo informacijo, kje se restavracije nahajajo. Na sliki 35 je prikazana uporaba različnih velikosti točk. Na levem zemljevidu je primer premajhne, na sredinskem zemljevidu prevelike, na desnem pa ustrezne velikosti točk.

Slika 35: Uporaba različnih velikosti točk



Vir: lastno delo.

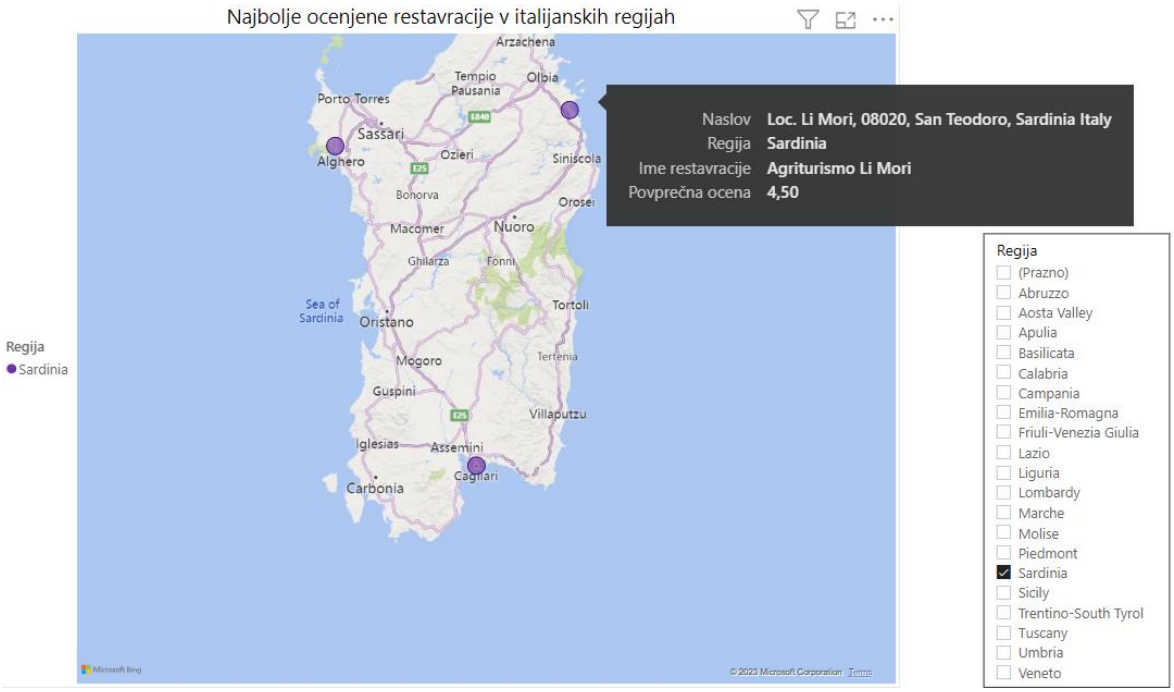
6.6.3.4 Velikost besedila

Prav tako kot pri horoplet zemljevidu in zemljevidu s kategorijami tudi ta zemljevid vsebuje besedilo v naslovu in legendi. Poskrbel sem, da sta naslov in besedilo legende dovolj velika, da ju bralec lahko brez težav prebere. To je vidno pri končnem rezultatu v poglavju 6.6.3.6.

6.6.3.5 Interaktivnost

Za lažje pridobivanje informacij z zemljevida se ob premiku miške na eno od točk prikažejo ime in naslov restavracije, ime regije in povprečna ocena restavracije. Za lažjo identifikacijo restavracij po posameznih regijah sem dodal tudi razčlenjevalnik, s katerim lahko uporabnik izbere regijo, ki ga zanima. Primer prikaza podatkov o restavraciji je na sliki 36.

Slika 36: Interaktivni elementi na zemljevidu s točkami



Vir: lastno delo.

6.6.3.6 Ravnotežje elementov in končni rezultat

Prav tako kot pri horoplet zemljevidu in zemljevidu s kategorijami sem tudi tukaj upošteval napotke za uravnoteženo postavitev elementov vizualizacije. Dodal sem tudi naslov in legendo. V legendi so z barvami označene različne regije, tako da so barve skladne z barvami točk na zemljevidu.

Slika 37: Najbolje ocenjene restavracije v italijanskih regijah

Vir: lastno delo.

7 UGOTOVITVE

Pri uporabi zemljevida s točkami sem ugotovil dve težavi, ki se lahko pojavita ob uporabi napačne velikosti točk. Pri premajhnih točkah so te lahko tako majhne, da sploh niso vidne. Še posebej pride težava s premajhnimi točkami do izraza, če imamo na zemljevidu tudi druge označbe, npr. imena držav ali teren. Na drugi strani pa ne smejo biti točke prevelike, saj se te lahko med seboj prekrivajo. Težava s preveliko velikostjo točk pride najbolj do izraza, ko želimo prikazati točne lokacije. V tem primeru zaradi prevelikega premera točke ne prikazujejo točnih lokacij, ampak območja. V tem primeru se moramo odločiti, kolikšna natančnost je za nas še sprejemljiva.

Pri uporabi zemljevida kategorij in horoplet zemljevida se nam lahko zgodi, da ne podpirata prikaza imen območij, npr. imen regij. To predstavlja težavo, saj uporabniki lahko ne poznajo imen območij. Težave ne moremo rešiti z dodajanjem barve območjem, saj so barve že uporabljene za prikaz drugih vrednosti. Ko želimo prikazati, kolikšno vrednost zaseda neko območje, lahko zemljevid podpremo z dodatno vizualizacijo, kot je tabela. V mojem primeru sem uporabil tabelo s padajoče razvrščenimi povprečnimi ocenami restavracij, s katere je hitro razvidno, katera regija ima največjo in katera najmanjšo vrednost. Prav tako lahko težavo rešimo tako, da zemljevid podpremo z interaktivnostjo. Pri svojem primeru sem zemljevidu dodal pojavno okno, ki se pokaže s premikom miške na posamezno regijo. V njem so izpisani vsi potrebni podatki.

Prav tako se nam težava z nepodprtim izpisom imen območij pojavi na zemljevidu s točkami. Tako vidimo točne lokacije točk, v primeru nepoznavanja prikazanega zemljevida pa ne vemo, katera točka se nahaja v katerem območju. Tudi tu lahko težavo z rešimo z dodatno tabelo in interaktivnostjo, za razliko od horoplet zemljevidov in zemljevidov kategorij pa lahko regije ločimo z barvami. V mojem primeru sem skupku točk vsake regije dodal unikatno barvo in regije s pripadajočimi barvami dodal v legendo. Uporabil sem tudi interaktivnost, ki s premikom miške na posamezno točko prikaže pojavno okno, na katerem je med drugim izpisano tudi ime regije.

Pri naslovih moramo biti pozorni, saj ti lahko pravilno vsebujejo imena tujih držav. V mojem primeru je restavracija iz Francije imela naslov, ki je pravilno vseboval besedilo Italy. Dvoumne naslove, pri katerih ne vemo, ali gre za napako, lahko ročno preverimo z resničnimi podatki.

Zemljevidi kategorij in horoplet zemljevidi hranijo šifrante z imeni območij, ki jih pričakujejo. Imena območij, ki jih vsebuje naš podatkovni vir, morajo biti enaka tem imenom. V nasprotnem primeru zemljevid območij ne prepozna in jih zato ne prikaže. Težavo z napačnimi imeni lahko rešimo tako, da vrednosti v podatkovnem viru nadomestimo z vrednostmi, ki jih zemljevid pričakuje. V mojem primeru sem imel težavo s prikazom regij, saj je zemljevid pričakoval regije v angleškem, moj podatkovni vir pa je hranil imena v italijanskem jeziku. V Power BI Desktopu lahko pričakovana imena regij preverimo.

Pri nalaganju podatkov lahko koordinate, ki so zapisane kot decimalno število, izgubijo decimalno vejico ali piko. Do težave pride zaradi območnih nastavitev, ki niso skladne s formatom podatkov. Primera: če je območna nastavitev nastavljena na »slovenščino (Slovenija)«, morajo podatki vsebovati decimalne vejice, v primeru »angleščine (Združeno kraljestvo)« pa morajo podatki vsebovati decimalne pike. Težava pri uporabi decimalnih pik pri slovenščini je, da pomenijo ločevanje tisočic od nižjih enot števila. To pa lahko predstavlja težavo, če so podatki shranjeni v datoteki tipa CSV, saj so vejice rezervirane za ločevanje atributov in zato decimalnih pik ne moremo nadomestiti z vejicami. Zato je priporočljivo, da pred nalaganjem podatkov območno nastavitev spremenimo na tisto, s katero so podatki predstavljeni.

Po čiščenju moramo preveriti, da je ime podatkovnega vira še vedno ustrezno. Pri mojem primeru se je podatkovni vir imenoval »tripadvisor_european_restaurants«, po odstranitvi restavracij, ki niso iz Italije, pa bi bilo bolj ustrezno ime »tripadvisor_italian_restaurants«. Ime podatkovnega vira ne vpliva na analizo podatkov v primeru, da želimo, pa ga lahko po čiščenju ustrezno popravimo.

Če zemljevid podpremo s stolpčnim diagramom, moramo paziti na razpon vrednosti na koordinatni osi, ki predstavlja količino. V primeru prevelikega razpona se nam lahko zgodi težava, da težko razločimo, katere vrednosti so manjše in katere večje. To je še posebej opazno, če so vrednosti podatkov skoncentrirane v majhnem deležu celotnega razpona. Na

to lahko vpliva tudi uporaba najmanjše vrednosti 0, saj pripomore k večjemu razponu na koordinatni osi kot, če bi bila najmanjša vrednost večja. Zato je v primerih, ko imamo vrednosti skoncentrirane na manjšem deležu razpona, treba uporabiti večjo najmanjšo vrednost od 0. Vseeno pa moramo paziti, da prikazana razlika med vrednostmi ne postane prevelika in zavajajoča.

Nekateri informacijski sistemi za izdelavo vizualizacij z zemljevidi vsebujejo teme, ki so prijazne barvno slepim ljudem. Paziti moramo, saj lahko katera od tem ne podpira vseh barvnih slepot. V mojem primeru sem uporabil takšno temo v Power BI Desktopu in z orodjem Coblis ugotovil, da 1 od 8 barvnih slepot ni podprta. Zato je kljub uporabi barvno slepim ljudem prijazne teme treba dodatno preveriti, ali je zemljevid res razumljiv pri vseh barvnih slepotah. Za preverjanje lahko uporabimo različna orodja.

Podatkovni vir ima lahko prelomljene vrstice in se zato 1 vrstic ob nalaganju podatkov doda kot 2 ali več vrstic z zamaknjenimi vrednostmi. Takšne vrstice lahko hitro identificiramo, saj imajo vrednosti pri nekaterih stolpcih prazne ali nesmiselne. Težavo lahko odpravimo tako, da v podatkovnem viru takšne vrstice poiščemo in ustrezno združimo ter podatke ponovno naložimo.

Pri naslovih se nam lahko zgodi, do so ti sicer zapisani v podatkovnem viru, vendar niso dovolj natančni. Pri mojem primeru so nekateri naslovi vsebovali samo podatek o mestu in ne o ulici. V tem primeru lahko uporabimo koordinate, če so te na voljo, ali pa pridobimo bolj natančne vrednosti iz drugega (bolj natančnega) podatkovnega vira.

Pri nekaterih neveljavnih koordinatah in koordinatah, ki se nahajajo na napačni lokaciji, lahko sami logično sklepamo, za kakšno napako gre in kako napako odpraviti. Primera sta koordinate brez decimalnih ločil in zamenjani koordinati. Ob pogledu v podatkovni vir lahko sami opazimo, v čem je težava. V prvem primeru hitro opazimo, da so koordinate prevelike in bi bile bolj smiselne z decimalnimi vejicami, v drugem primeru pa lahko koordinate primerjamo z ostalimi v podatkovnem viru in opazimo, da bi bile bolj skladne z ostalimi, če bi bili zemljepisna širina in dolžina zamenjani. Nekatere neveljavne in napačne koordinate niso popolnoma neuporabne in lahko napake sami identificiramo ter popravimo.

Ugotovil sem, da lahko tudi pravilni podatkovni vir hrani napačno lokacijo. V mojem primeru sem na začetku predpostavil, da so na spletni strani Tripadvisor podatki resnični in zato nekatere manjkajoče ter nepravilne podatke v mojem podatkovnem viru nadomestil s tamkajšnjimi vrednostmi. Ugotovil pa sem, da so lahko tudi tam lokacije napačne in je treba lokacije dodatno preveriti, preden jih uporabimo. V mojem primeru je ena od restavracij tudi na spletni strani Tripadvisor hranila napačen naslov.

Horplet zemljevidi v aplikacijah za vizualizacijo lahko ne vsebujejo podatkov o nekaterih mejah poligonov. Nekatere meje so prednaložene. V mojem primeru je horoplet zemljevid v Power BI Desktopu vseboval meje italijanskih regij, ni pa vseboval meje evropskih držav. Ker pa zemljevid podpira dodajanje novih meja, sem meje držav lahko poiskal na svetovnem

spletu in jih naložil v horoplet zemljevid. Datoteke z mejami so lahko shranjene v različnih formatih, zato moramo uporabiti format, ki ga uporablja zemljevid.

Zemljevid s poligoni v Power BI Desktopu, ki se lahko uporablja kot zemljevid s kategorijami ali horoplet zemljevid, podatke vedno prikaže kot kategorije oz. z različnimi barvami. Ko sem pri mojem primeru želel prikazati povprečne ocene regij s horoplet zemljevidom, so bile posamezne vrednosti prikazane z različnimi barvami. Zato sem moral izdelati dodatno tabelo, ki sem jo uporabil za legendo in barve zemljevida ročno nastaviti glede na vrednosti, ki jih je nova tabela vsebovala.

Pri horoplet zemljevidu lahko uporabimo samodejno dodeljevanje barv. Zemljevid na podlagi vrednosti poligonom sam dodeli ustrezne odtenke barve. Paziti moramo, da izberemo ustrezen razpon vrednosti, saj je v primeru prevelikega razpona interval med vrednostmi lahko premajhen in so zato odtenki preveč podobni med seboj. Pri mojem primeru so bili, ob uporabi razpona med najmanjšo in največjo možno vrednostjo 1 in 5, odtenki preveč podobni, saj so vsi vsebovali vrednosti v majhnem deležu tega razpona – med 3,94 in 4,15. Zato sem razpon zmanjšal na ti 2 vrednosti in odtenki so se med seboj lepo razlikovali. Pri dodeljevanju razpona moramo poskusiti različne vrednosti in videti, pri katerem se razlike v odtenkih lepo razloči.

Ko prikazujemo točne lokacije, so nekatere točke lahko precej skupaj, v primeru duplikatov pa celo na isti lokaciji. V mojem primeru sem imel 2 točki, ki sta se zaradi nenatančnega naslova prikazovali napačno v Združenih državah Amerike. Bili sta zelo skupaj, zato sta od daleč izgledali, kot da je tam samo 1 točka. Zato je 1 točko priporočljivo navidezno približati in preveriti, ali gre res samo za 1. Prav tako sem imel točke, ki so zasedale enako lokacijo. Zemljevid takšne točke ni ločil, ampak jih je pokazal kot 1 točko. V tem primeru približevanje ne razkrije, za koliko točk gre. Takšne točke lahko poiščemo tako, da v podatkovnem viru preverimo, ali je katera od lokacij podvojena ali pa število vrstic primerjamo s številom točk, ki so prikazane na zemljevidu.

Ugotovil sem, da v primeru podvojenih primarnih ključev ni nujno, da gre za podvojene vnose. Lahko gre za posledico prelomljenih vrstic v podatkovnem viru, kjer sta se 2 ali več vrstic prelomili pri istem atributu, ki hrani enako vrednost. Recimo, da se 2 vrstici prelomita pred atributom »mesto« in ta atribut pri obeh vrednostih zaseda vrednost »Ljubljana«, bosta s tem nastala 2 vnosa, ki pri prvem atributu (to je lahko tudi primarni ključ) zasedata vrednost »Ljubljana«. Zato moramo vrstice s podvojenimi primarnimi ključi pred izbrisom dodatno preveriti. Kako takšne vrstice poiščemo in popravimo, sem opisal pri ugotovitvah višje.

Podvojeni vnosi ne nujno vplivajo na informacije, ki jih vizualizacija sporoča. V mojem primeru so bili v Power BI Desktopu podvojeni vnosi na zemljevidu s točkami prekriti, zato so bili pravilno prikazani kot 1 točka, pri horoplet zemljevidu pa se sprememba v odtenku območja ni opazila, saj je šlo za prikaz velikega števila vnosov. Vseeno pa je treba podvojene vnose odstraniti, da imamo podatke čiste. Ob njihovi ponovni porabi v drugi aplikaciji ali pa

že ob samem brskanju bi nam podvojeni vnosi lahko predstavljali lažno sliko. Če bi imeli pri horoplet zemljevidu malo podatkov oz. velik delež podvojenih vnosov, bi to lahko vplivalo tudi na odtenek območja in s tem informacijo.

Ugotovil sem, da so lahko atributi, ki jih ne potrebujemo za analizo, vseeno uporabni prav pri čiščenju podatkov. Zato jih je priporočljivo odstraniti v kasnejših korakih čiščenja, da si lahko z njimi do takrat pomagamo.

Zaradi lažjega pregleda sem zgoraj opisane ugotovitve in pripadajoče usmeritve obnovil ter združil. Prikazane so v tabeli 6.

Tabela 6: Ugotovitve

Ugotovitev	**Usmeritve**
Pri zemljevidu s točkami te ne smejo biti premajhne ali prevelike.	Točke morajo biti dovolj velike, da se jih hitro vidi in lahko razbere ter dovolj majhne, da se med seboj ne prekrivajo in prikazujejo dovolj točne lokacije za naše potrebe.
Zemljevidi kategorij in horoplet zemljevidi lahko ne podpirajo prikaza imen območij.	Zemljevid lahko podpremo z dodatno tabelo, ki vsebuje razvrščene vrednosti. Zemljevidu lahko dodamo tudi interaktivnost, ki posamezna območja podpre s pripadajočimi informacijami v pojavnem oknu.
Zemljevidi s točkami lahko ne podpirajo prikaza imen območij.	Zemljevid lahko podpremo z dodatno tabelo, interaktivnostjo, npr. pojavno okno s potrebnimi podatki, ali pa za vsako območje določimo unikatno barvo, s katero so točke na tem območju obarvane. Barve nato dodamo tudi v legendo.
Naslovi lahko vsebujejo imena tujih držav.	Dvoumne naslove, ki vsebujejo imena tujih držav, moramo ročno preveriti, saj so ti lahko tudi pravilni.
Pri uporabi zemljevida kategorij ali horoplet zemljevida se morajo imena območij v podatkovnem viru ujemati z imeni, ki jih zemljevid pričakuje.	Preveriti je treba, kakšne vrednosti zemljevid pričakuje in z njimi posodobiti vrednosti v podatkovnem viru.
Napačne območne nastavitve lahko koordinatam odstranijo decimalne pike ali vejice.	V informacijskem sistemu moramo uporabiti območne nastavitve, ki uporabljajo enaka decimalna ločila, kot jih imajo podatki, ki jih nalagamo.
Po čiščenju podatkov lahko ime podatkovnega vira ni več ustrezno.	Ime podatkovnega vira lahko po potrebi ročno popravimo.
Prevelik razpon na koordinatni osi s količino lahko vodi v težjo primerjavo med stolpci.	Razpon na koordinatni osi, ki predstavlja količino, moramo ustrezno prilagoditi tako, da vidimo, katere vrednosti so večje in katere manjše. Paziti moramo, da diagram ni zavajajoč in še vedno prikazuje približno enake razlike med stolpci.

Tabela 6: Ugotovitve (nad.)

Ugotovitev	**Usmeritve**
Teme za barvno slepe ljudi lahko ne podpirajo vseh barvnih slepot.	Če uporabimo temo, ki je prijazna barvno slepim ljudem, je priporočljivo, da dodatno preverimo, ali je zemljevid res razumljiv pri vseh barvnih slepotah. Za preverjanje lahko uporabimo različna orodja.
Vrstice v podatkovnem viru so lahko prelomljene in se zato 1 entiteta naloži kot 2 ali več entitet.	Te vrstice moramo najprej identificirati. To lahko storimo v orodju za čiščenje podatkov, kjer poiščemo vrstice, ki imajo veliko praznih oz. nesmiselnih vrednosti. Nato te vrstice v podatkovnem viru ustrezno združimo in ponovno naložimo.
Naslovi lahko niso shranjeni dovolj natančno, npr. vsebujejo samo podatke o mestu in ne o ulici.	Če so na voljo, lahko uporabimo koordinate ali pa podatek o ulici pridobimo iz drugega podatkovnega vira.
Pri nekaterih neveljavnih in napačnih koordinatah lahko sami logično sklepamo, za kakšno napako gre in kako jo odpraviti, takšne koordinate niso nujno neuporabne.	Koordinate preverimo v podatkovnem viru in poskušamo ugotoviti, za kakšno napako gre. Lahko jih primerjamo tudi s preostalimi pravilnimi koordinatami.
Čeprav imamo na voljo pravilni podatkovni vir, lahko ta hrani napačne lokacije.	Pred uporabo lokacij iz pravilnega podatkovnega vira je pred uporabo priporočljivo preveriti njihovo pravilnost.
Horplet zemljevidi lahko ne vsebujejo podatkov o nekaterih mejah poligonov.	Če zemljevid to podpira, lahko podatke o mejah poiščemo na svetovnem spletu in jih naložimo v zemljevid. Ker so datoteke z mejami lahko shranjene v različnih formatih, moramo paziti, da uporabimo format, ki ga uporablja naš zemljevid.
Nekateri zemljevidi s poligoni ob dodajanju legende barve spremenijo v popolnoma različne, čeprav gre za horoplet zemljevid.	Ustvarimo lahko novo tabelo in jo uporabimo za legendo ter barve vrednosti ročno nastavimo za horoplet zemljevid.
Pri samodejnem dodeljevanju barv na horoplet zemljevidu je treba določiti primeren razpon vrednosti. Pri prevelikem razponu so si lahko odtenki različnih vrednosti med seboj preveč podobni in je med njimi težko razlikovati.	Ne uporabimo prevelikega razpona. Poizkušamo lahko različne razpone in vidimo, kateri najbolj ustreza.
Ko prikazujemo točne lokacije, so nekatere točke lahko precej skupaj, v primeru duplikatov pa celo na isti lokaciji. Zato so te lokacije zavajajoče prikazane kot 1 točka.	S tem ko posamezne točke na zemljevidu približamo, lahko razločimo točke, ki so zelo blizu in na daleč prikazane kot 1. V primeru duplikatov pa lahko preverimo, ali se na tej lokaciji nahaja več točk v podatkovnem viru. V primeru, ko nimamo veliko točk, lahko preverimo tudi, ali se število vrstic v podatkovnem viru ujema s številom točk na zemljevidu.
V primeru podvojenih primarnih ključev ni nujno, da gre za podvojene vnose, ampak so ti posledica preloma vrstic v podatkovnem viru.	Vrstic s podvojenimi primarnimi ključi ne smemo odstraniti, ampak najprej preveriti, ali gre res za podvojene vrstice. Kako vrstice, ki so posledica preloma, poiščemo in popravimo, sem že opisal.

Tabela 6: Ugotovitve (nad.)

Ugotovitev	Usmeritve
Podvojene vnose je priporočljivo odstraniti, čeprav ti ne vplivajo na vizualizacije. Dobro je, da imamo čiste podatke za njihovo nadaljnjo uporabo.	Čeprav lahko v nekaterih primerih podvojeni vnosi ne vplivajo na vizualizacije in s tem informacije, pa lahko v drugih primerih ob ponovni uporabi podatkov ti sporočajo napačno informacijo. Zato je podvojene vnose priporočljivo odstraniti.
Atributi, ki jih ne potrebujemo za analizo, so lahko vseeno uporabni pri čiščenju podatkov.	Odvečne atribute lahko odstranimo v kasnejših korakih čiščenja ali pa na koncu.

Vir: lastno delo.

SKLEP

V magistrskem delu sem naredil celovit pregled nad napotki za pripravo kakovostnih vizualizacij z zemljevidi, ki sem jih našel v literaturi in spletnih virih. Te sem upošteval tudi v empiričnem delu, kjer sem izvedel vizualizacije z zemljevidi na primeru. Dosegel sem vse cilje, ki sem si jih na začetku zastavil.

Cilj magistrskega dela, ki je bil na podlagi literature ugotoviti, kaj so merila kakovostne vizualizacije podatkov z zemljevidi, sem dosegel s preučevanjem literature in spletnih virov. Pri tem cilju sem merila enačil z vsemi smernicami in postopki, ki sem jih pridobil s preučevanjem literature, ter z mojimi ugotovitvami. Na začetku sem opisal ocenjevanje kakovosti in čiščenje podatkov na splošno, pri čemer sem naredil poudarek na čiščenju. Najbolj sem se osredotočil na operacije, ki se jih lahko aplicira tudi na prostorske podatke. V nadaljevanju sem opisal tudi ocenjevanje kakovosti in čiščenje prostorskih podatkov. Nadaljeval sem z opisovanjem zagotavljanja kakovostnih vizualizacij na splošno, ki sem jih prav tako nadgradil s specifikami vizualizacij z zemljevidi. Na koncu sem izvedel tudi vizualizacijo podatkov z zemljevidi na primeru. S tem sem prišel do ugotovitev, ki prav tako spadajo k merilom kakovostne vizualizacije.

Cilj magistrskega dela, ki je bil podati smernice za pripravo kakovostnih podatkov, sem dosegel z opisom smernic na podlagi ugotovitev iz literature kot tudi na podlagi ugotovitev pri proučevanju primera. Poleg ugotovitev iz literature, ki sem jih upošteval pri implementaciji primera, sem prišel tudi do novih izzivov, ki jih v literaturi in spletnih virih nisem zasledil. Zato sem to tudi opisal v obliki smernic, ki jih lahko bralec upošteva.

Cilj magistrskega dela, ki je bil podati smernice glede priprave kakovostne vizualizacije, sem prav tako dosegel na podlagi ugotovitev iz literature in ugotovitev pri proučevanju primera. Na podlagi literature sem opisal najbolj pogosto uporabljene vrste zemljevidov, pojasnil, kaj z njimi prikazujemo oz. kdaj lahko katerega od njih uporabimo, in podal smernice glede priprave kakovostnih vizualizacij. Te sem nadgradil z ugotovitvami, ki sem

jih pridobil z implementacijo vizualizacij na primeru. Tudi te ugotovitve sem zapisal v obliki smernic, s katerimi si bralec lahko pomaga pri izdelavi vizualizacij z zemljevidi.

Namen magistrskega dela, ki je bil prispevati k razumevanju, kako zagotoviti kakovostno vizualizacijo podatkov z zemljevidi, sem dosegel z združitvijo meril in smernic iz literature ter spletnih virov na enem mestu. Poleg združitve sem uporabo teh meril in smernic prikazal tudi na primeru s 3 različnimi vrstami zemljevidov, kjer sem izvedel postopek od nalaganja podatkov do implementacije vizualizacij. Pri izvedbi postopka sem naletel tako na izzive, opisane v teoretičnem delu, kot tudi na nove izzive. Kako se z njimi soočiti, sem sproti tudi opisal. Bralec si lahko z magistrskim delom pomaga tako, da upošteva smernice iz teoretičnega dela, sledi postopku v empiričnem delu in si pomaga z ugotovitvami, kjer so smernice izzivov, ki sem jih ugotovil v empiričnem delu, obnovljene.

Literatura in spletni viri, ki sem jih preučil, naštevajo različne smernice za zagotavljanje kakovostnih vizualizacij podatkov z zemljevidi. Nekatere smernice se ponovijo v različnih virih, v osnovi pa vsak vir poda nekaj novih unikatnih smernic, ki jih v drugih virih nisem opazil. Zato je dodana vrednost tega magistrskega dela združitev smernic v strukturirano obliko na enem mestu. Prav tako je dodana vrednost pridobitev novih ugotovitev, ki sem jih ugotovil ob izvajanju empiričnega dela. Gre za težave, s katerimi sem se srečal ob izvedbi. Te sem v poglavju Ugotovitve podkrepil tudi s smernicami, kako lahko te težave odpravimo. Prav tako sem v magistrskem delu preizkusil smernice iz literature in svetovnega spleta ter pri nekaterih ugotovil, da ne moremo vseh aplicirati na vse probleme, ampak jih moramo zaradi lažjega razumevanja vizualizacij prilagoditi glede na primer uporabe.

Zaradi narave primera v empiričnem delu sem za reševanje uporabil 3 zemljevide, ki zastavljeni problem najbolje rešujejo. To so horoplet zemljevid, zemljevid kategorij in zemljevid s točkami. Kot sem opisal v teoretičnem delu, poznamo veliko različnih vrst zemljevidov. Možnost za nadaljnje delo bi zato lahko predstavljal preizkus smernic na preostalih zemljevidih in s tem identifikacija dodatnih težav, ki se nam lahko pri njih pojavijo. Ker so nekatere ugotovitve magistrskega dela vezane na aplikacijo, v kateri sem izvedel čiščenje in izdelal vizualizacije, bi se lahko magistrsko delo nadgradilo z izvedbo teh procesov še v kateri drugi aplikaciji. Tako bi se kot v primeru Power BI Desktopa tudi pri ostalih aplikacijah lahko identificiralo morebitne izzive, na katere lahko naletimo in moramo biti nanje pozorni.

Pri izdelavi vizualizacij na primeru sem upošteval ugotovitve iz literature in spletnih virov. Kljub upoštevanju teh ugotovitev pa sem pri izdelavi ravnal subjektivno. Zato bi kot omejitev mojega magistrskega dela izpostavil to, da nisem pridobil povratne informacije o tem, ali so vizualizacije res kakovostne. Vizualizacije bi lahko pokazal drugim ljudem, ki bi iz njih poskusili razbrati informacije, ki sem jih želel sporočiti. S tem bi moje ugotovitve potrdili ali ovrgli.

LITERATURA IN VIRI

1. Ali, E. (2020). *Geographic information system (GIS): definition, development, applications & components.* Pridobljeno 8. aprila 2023 iz https://www.researchgate.net/publication/340182760_Geographic_Information_System_GIS_Definition_Development_Applications_Components
2. Ali, S. M., Gupta, N., Nayak, G. K. & Lenka, R. K. (2016, december). Big data visualization: Tools and challenges. V *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, (str. 656–660). IEEE.
3. Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G. & Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. *DAMA UK working group,* 432–435.
4. Barik, R. K., Lenka, R. K., Ali, S. M., Gupta, N., Satpathy, A. & Raj, A. (2017, maj). Investigation into the efficacy of geospatial big data visualization tools. V *2017 International Conference on Computing, Communication and Automation (ICCCA)* (str. 88–93). IEEE.
5. Barker, P. (2015, 27. november). *The five different types of map and their uses* [objava na blogu]. Pridobljeno 19. februarja 2023 iz https://www.here.com/learn/blog/the-five-different-types-of-map-and-their-uses
6. Baur, J., Moreno-Villanueva, M., Kötter, T., Sindlinger, T., Bürkle, A., Berthold, M. R. & Junk, M. (2015). MARK-AGE data management: Cleaning, exploration and visualization of data. *Mechanisms of ageing and development*, *151*, 38–44.
7. Bindzárová Gergel'ová, M., Kuzevičová, Ž., Labant, S., Gašinec, J., Kuzevič, Š., Unucka, J. & Liptai, P. (2020). Evaluation of selected sub-elements of spatial data quality on 3D flood event modeling: Case study of Prešov City, Slovakia. *Applied Sciences, 10*(3), 820.
8. Brown, C. (brez datuma). *How to Clean Map Data*. Pridobljeno 2. decembra 2022 iz https://mangomap.com/industries/web-mapping/data-sourcing-preparation/how-to-clean-map-data.html
9. Buckley, A. (2012). Make maps people want to look at: Five primary design principles for cartography. *ArcUser Winter*, 46–51.
10. Caliper. (brez datuma). *What is a thematic map? 7 Types of Thematic Maps*. Pridobljeno 10. junija 2022 iz https://www.caliper.com/glossary/what-is-a-dot-density-map.htm
11. Centers for Disease Control and Prevention. (brez datuma). *Types of Thematic Maps*. Pridobljeno 23. junija 2022 iz https://www.cdc.gov/dhdsp/maps/gisx/resources/thematic-maps.html
12. Chapman, A. D. (2005). *Principles and methods of data cleaning*. Copenhagen: GBIF.
13. Chaudhri, V. K. (2021, 4. januar). *General Principles of Visualization Design*. Pridobljeno 19. februarja 2023 iz https://www.linkedin.com/pulse/general-principles-visualization-design-vinay-k-chaudhri
14. Chaudhuri, S., Dayal, U. & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM, 54*(8), 88–98.

15. Chiang, Y. Y., Wu, B., Anand, A., Akade, K. & Knoblock, C. A. (2014, november). A system for efficient cleaning and transformation of geospatial data attributes. V *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (str. 577–580).
16. Colblindor. (brez datuma). *Coblis – Color Blindness Simulator*. Pridobljeno 8. aprila 2023 iz https://www.color-blindness.com/coblis-color-blindness-simulator
17. Cukier, K. (2010). A special report on managing information. *The Economist, 394*(8671), 3–18.
18. Datawrapper. (2021, 4. januar). *What to consider when creating choropleth maps*. Pridobljeno 19. februarja 2023 iz https://academy.datawrapper.de/article/134-what-to-consider-when-creating-choropleth-maps
19. Devillers, R. & Jeansoulin, R. (2006). *Fundamentals of spatial data quality*. London: ISTE.
20. Dong, L., Bai, Q., Kim, T., Chen, T., Liu, W. & Li, C. (2020, junij). Marviq: Quality-Aware Geospatial Visualization of Range-Selection Queries Using Materialization. V *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (str. 67–82).
21. Dougherty, J. & Ilyankou, I. (2021). *Hands-on data visualization*. Sebastopol: O'Reilly Media, Inc.
22. Duke University Libraries. (2022, 10. januar). *Data Visualization: Chart Dos and Don'ts*. Pridobljeno 19. februarja 2023 iz https://guides.library.duke.edu/datavis/topten
23. Eldersveld, D. (brez datuma). *TopoJSON Collection.* Pridobljeno 8. aprila 2023 iz https://github.com/deldersveld/topojson/blob/master/world-countries.json
24. Eldrandaly, K. A., Abdel-Basset, M. & Shawky, L. A. (2019). Internet of spatial things: A new reference model with insight analysis. *IEEE Access, 7*, 19653–19669.
25. Elena, C. (2011). Business intelligence. Journal of Knowledge Management, Economics and Information Technology, *1*(2), 1–12.
26. Ferster, C. J., Nelson, T., Roberston, C. & Feick, R. (2017). Current Themes in Volunteered Geographic Information. V *Gis Applications for Socio-Economics and Humanity* (str. 26–41). Elsevier Inc.
27. Few, S. (2009). *Now you see it: simple visualization techniques for quantitative analysis.* Oakland: Analytics Press.
28. Few, S. & Edge, P. (2007). Data visualization: past, present, and future. *IBM Cognos Innovation Center*.
29. Fonte, C. C., Antoniou, V., Bastin, L., Estima, J., Arsanjani, J. J., Bayas, J. C. L., See, L. & Vatseva, R. (2017). Assessing VGI data quality. *Mapping and the citizen sensor*, 137–163.
30. Ganti, V. & Sarma, A. D. (2013). Data Cleaning: A Practical Perspective. *Synthesis Lectures on Data Management, 5*(3), 1–85.
31. Gigante, M. (2019, 18. julij). *What Is GIS Mapping? (+How to Use the Different Types of GIS Maps)*. Pridobljeno 23. junija 2022 iz https://www.g2.com/articles/gis-mapping

32. GISGeography. (brez datuma a). *25 Map Types: Brilliant Ideas to Build Unbeatable Maps*. Pridobljeno 20. junija 2022 iz https://gisgeography.com/map-types
33. GISGeography. (brez datuma b). *Dot Distribution vs Graduated Symbols vs Proportional Symbol Maps*. Pridobljeno 20. oktobra 2022 iz https://gisgeography.com/dot-distribution-graduated-symbols-proportional-symbol-maps
34. GISGeography. (brez datuma c). *Latitude, Longitude and Coordinate System Grids*. Pridobljeno 8. aprila 2023 iz https://gisgeography.com/latitude-longitude-coordinates
35. Guldåker, N. (2020). Geovisualization and geographical analysis for fire prevention. *ISPRS International Journal of Geo-Information*, *9*(6), 355.
36. Hari's BI. (2018, 24. april). *Power BI Latitude and Longitude Function | Bing Map API* [YouTube]. https://www.youtube.com/watch?app=desktop&v=YxwU5UubWjI&ab_channel=Hari%27sBI
37. Hohnova, A. & Vondrakova, A. (2017). Spatial data visualization: The appropriate use of colours. *International Multidisciplinary Scientific GeoConference: SGEM*, *17*, 657–664.
38. How to Power BI (2021, 19. marec). *Import from CSV in Power BI | Fixing missing decimals and columns* [YouTube]. https://www.youtube.com/watch?v=wuLnv3QJCHg&ab_channel=HowtoPowerBI
39. ICSM. (brez datuma). *Types of Maps*. Pridobljeno 19. februarja 2023 iz https://www.icsm.gov.au/education/fundamentals-mapping/types-maps
40. International Organization for Standardization. (2013). *ISO 19157: 2013 Geographic information – Data quality*. Pridobljeno 8. aprila 2023 iz https://www.sis.se/en/produkter/information-technology-office-machines/applications-of-information-technology/it-applications-in-science/iso191572013
41. KAISPE. (brez datuma). *Reverse GeoLocation In Power BI*. Pridobljeno 8. aprila 2023 iz https://www.kaispe.com/reverse-geolocation-in-power-bi
42. Kang, H., Sehgal, V. & Getoor, L. (2007, julij). Geoddupe: A novel interface for interactive entity resolution in geospatial data. V *2007 11th International Conference Information Visualization (IV'07)* (str. 489–496). IEEE.
43. Kelleher, C. & Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, *26*(6), 822–827.
44. Kirk, A. (2016). *Data Visualisation: A Handbook for Data Driven Design*. United Kingdom: SAGE Publications.
45. Koshley, D. K. & Halder, R. (2015, avgust). Data cleaning: An abstraction-based approach. V *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (str. 713–719). IEEE.
46. Krzywinski, M., Birol, I., Jones, S. & Marra, M. (2012, oktober). *Getting Into Visualization of Large Biological Data Sets*. Pridobljeno 8. aprila 2023 iz http://mkweb.bcgsc.ca/biovis2012/krzywinski-visualizing-biological-data.pdf
47. Kulyk, O., Kosara, R., Urquiza, J. & Wassink, I. (2007). Human-centered aspects. V

Human-Centered Visualization Environments: GI-Dagstuhl Research Seminar, Dagstuhl Castle, Germany, March 5–8, 2006, Revised Lectures (str. 13–75). Springer Berlin Heidelberg.

48. Leone, S. (brez datuma). *TripAdvisor European restaurants*. Pridobljeno 8. aprila 2023 iz https://www.kaggle.com/datasets/stefanoleone992/tripadvisor-european-restaurants
49. Mallon, M. (2015). Data Visualization. *Public Services Quarterly*, *11*(3), 183–192.
50. Maptive. (2020, 19. oktober). *What is a Thematic Map? 6 Types of Thematic Maps* [objava na blogu]. Pridobljeno 23. junija 2022 iz https://www.maptive.com/thematic-map-examples
51. Marzouki, A., Lafrance, F., Daniel, S. & Mellouli, S. (2017, junij). The relevance of geovisualization in Citizen Participation processes. V *Proceedings of the 18th annual international conference on digital government research* (str. 397–406).
52. Microsoft. (2022, 21. julij). *Find a Location by Point.* Pridobljeno 8. aprila 2023 iz https://learn.microsoft.com/en-us/bingmaps/rest-services/locations/find-a-location-by-point
53. Microsoft. (2023, 12. januar). *What is Power BI Desktop?*. Pridobljeno 8. aprila 2023 iz https://learn.microsoft.com/en-us/power-bi/fundamentals/desktop-what-is-desktop
54. Midway, S. R. (2020). Principles of effective data visualization. *Patterns*, *1*(9), 100141.
55. Namibia Statistics Agency. (2016, 7. oktober). *Government gazette of the Republic of Namibia No. 6145*. Pridobljeno 8. aprila 2023 iz https://gazettes.africa/archive/na/2016/na-government-gazette-dated-2016-10-07-no-6145.pdf
56. National Geographic. (brez datuma). *GIS (Geographic Information System)*. Pridobljeno 23. junija 2022 iz https://www.nationalgeographic.org/encyclopedia/geographic-information-system-gis
57. Naumann, F. (2014). Data profiling revisited. *ACM SIGMOD Record*, *42*(4), 40–49.
58. Ong, I. L., Siew, P. H. & Wong, S. F. (2011). A five-layered business intelligence architecture. *Communications of the IBIMA*, *2011*, 1–11.
59. Parmar, V. & Sheoran, S. (2021). Context-based spatial metadata cleaning using QGIS. *Vidyabharati International Interdisciplinary Research Journal*, *12*(1), 55–62.
60. Rapid Insight. (brez datuma). *7 Data Cleanup Terms Explained Visually* [objava na blogu]. Pridobljeno 15. oktobra 2022 iz https://www.rapidinsight.com/blog/7-data-cleanup-terms-explained-visually
61. Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M. & Perry, R. (2016). Data visualization. *International Journal of Engineering Research And Advanced Technology (IJERAT)*, *2*(12), 11–16.
62. Satyanarayana, M. & Guptha, N. (2021, junij). A Systematic Approach For Data Cleansing Process of Geospatial Data to Perform Application Specific Data Analytics. V *Proceedings of the First International Conference on Computing, Communication and Control System, I3CAC 2021, 7–8 June 2021*. Chennai: Bharath University.

63. Sergieieva, K. (2021, 6. marec). *GIS Mapping: Types Of Interactive Maps & Applications* [objava na blogu]. Pridobljeno 23. junija 2022 iz https://eos.com/blog/gis-mapping
64. Sheoran, S. & Parmar, V. (2022). GeoWebCln: An Intensive Cleaning Architecture for Geospatial Metadata. *Quaestiones Geographicae, 41*(1), 51–62.
65. Sinar, E. F. (2015). Data visualization. V *Big Data at Work* (str. 115–157). New York: Routledge.
66. Słomska-Przech, K. & Gołębiowska, I. M. (2021). Do different map types support map reading equally? Comparing choropleth, graduated symbols, and isoline maps for map use tasks. *ISPRS International Journal of Geo-Information*, *10*(2), 69.
67. Spatial Data Science. (brez datuma). *Data preparation*. Pridobljeno 2. decembra 2022 iz https://rspatial.org/raster/sdm/2_sdm_occdata.html
68. Spencer, J. & Wilkes, B. (2019). *Assessing Spatial Data Quality Using Five Data Anomalies Speeding the Process for Master Facility*. Pridobljeno 2. decembra 2022 iz https://www.measureevaluation.org/resources/publications/wp-19-227/at_download/document
69. Starček, S. & Kovač, M. Š. (2019). Vpliv kakovosti prostorskih podatkov na učinkovitost sistema obdavčenja nepremičnin: primer nadomestila za uporabo stavbnega zemljišča. *Urbani Izziv*, *30*(1), 17–30.
70. Szafir, D. A. (2018). The good, the bad, and the biased: Five ways visualizations can mislead (and how to fix them). *Interactions*, *25*(4), 26–33.
71. Tableau. (brez datuma a). *A Guide To Geospatial Visualizations*. Pridobljeno 23. junija 2022 iz https://www.tableau.com/data-insights/reference-library/visual-analytics/geospatial
72. Tableau. (brez datuma b). *What Is Data Visualization? Definition, Examples, And Learning Resources*. Pridobljeno 23. junija 2022 iz https://www.tableau.com/learn/articles/data-visualization
73. Tennekes, M. (2018). tmap: Thematic Maps in R. *Journal of Statistical Software*, *84*, 1–39.
74. Trame, J. & Keßler, C. (2011). Exploring the lineage of volunteered geographic information with heat maps. *GeoViz, Hamburg, Germany*.
75. Tripadvisor. (brez datuma). *O Sorbet d'Amour Annecy*. Pridobljeno 8. aprila 2023 iz https://www.tripadvisor.com/g187260-d15324678
76. Triglav, J. (2012). Tretji klic k razmisleku in ducat drobnih idej. *Geodetski Vestnik*, *56*(3), 579.
77. Tufte, E. R. (2001). *The visual display of quantitative information.* Cheshire: Graphics Press.
78. UC Davis DataLab. (brez datuma). *Data Forensics and Cleaning: Geospatial Data*. Pridobljeno 2. decembra 2022 iz https://ucdavisdatalab.github.io/adventures_in_data_science/data-forensics-and-cleaning-geospatial-data.html
79. Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177–

189.
80. Ware, C. (2019). *Information visualization: perception for design*. Burlington: Morgan Kaufmann.
81. Wei, W. (2012). Research on the application of geographic information system in tourism management. *Procedia Environmental Sciences*, *12*, 1104–1109.
82. Wilke, C. O. (2019). *Fundamentals of data visualization: a primer on making informative and compelling figures*. Sebastopol: O'Reilly Media.
83. Yu, J., Zhang, Z. & Sarwat, M. (2018, julij). Geosparkviz: a scalable geospatial data visualization framework in the apache spark ecosystem. V *Proceedings of the 30th international conference on scientific and statistical database management* (str. 1–12).
84. Založnik, R. (2018). *Priprava in integracija podatkov za sisteme poslovne inteligence* (magistrsko delo). Ljubljana: Ekonomska fakulteta.
85. Zhu, A. X., Zhao, F. H., Liang, P. & Qin, C. Z. (2021). Next generation of GIS: must be easy. *Annals of GIS*, *27*(1), 71–86.