Analiza možnosti uporabe odprtih podatkov pri določanju okoljskih vplivov na razvoj bolezni

Lan Sevčnikar[1], Jurij Nastran[1], Vito Drofenik[1], Ana Trobec[2], Urška Vajngerl[2], Hana Kifle[3], Anja Pivaš[3], Luka Kovač[4], Damjan Popovski[1], Miha Jan[1], Ana Halužan Vasle[1], Tadeja Režen[5], Marjanca Starčič Erjavec[2], Stane Moškon[6], Matevž Juvančič[3], Špela Verovšek[3], Miha Moškon[1]

[1]*Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia,*
[2]*Biotechnical Faculty, University of Ljubljana, Ljubljana, Slovenia,*
[3]*Faculty of Architecture, University of Ljubljana, Ljubljana, Slovenia,*
[4]*Faculty of Education, University of Ljubljana, Ljubljana, Slovenia,*
[5]*Center for Functional Genomics and Bio-Chips, Institute of Biochemistry and molecular genetics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia,*
[6]*Nirek d.o.o., Vrhnika, Slovenia*
E-mail: miha.moskon@fri.uni-lj.si

Towards the application of open data for the assessment of environmental impacts on the development of a disease

In the proposed research, our aim was to identify various open datasets that can be utilized to quantify environmental factors influencing the development of a specific disease within a particular region. Furthermore, we implemented a computational framework to collect data from the identified sources, store this data and integrate it into an inference pipeline to identify and quantify the factors that influence the development of a specific disease. The framework was demonstrated on different European regions and can be applied to an arbitrary location, for which the applied datasets are available.

1 Uvod

Okolje in organiziranost urbanega okolja ključno vplivata na zdravje posameznika. Nedavno poročilo Evropske okoljske agencije (EEA Report No 21/2019, [1]) velik delež bolezni pripisuje onesnaženju okolja, ki izhaja iz aktivnosti človeka. V letu 2018 je ocena števila prezgodnjih smrti zaradi onesnaženosti zraka znašala 400.000 ljudi. Dejavnikom, kot so prekomerna onesnaženost zraka, toplotna obremenitev v poletnih mesecih, svetlobna onesnaženost in hrup, so izpostavljeni predvsem ljudje, ki živijo in/ali delajo v mestih. Znano je, da navedeni dejavniki vplivajo na razvoj nenalezljivih bolezni. Na primer, svetlobna onesnaženost vpliva na rušenje cirkadianih ritmov, kar vodi v razvoj nenalezljivih bolezni, kot je rak [2]. Delež zaradi onesnaženja povzročenih tovrstnih obolenj in mehanizmi njihovega razvoja zaenkrat še niso znani in popolnoma pojasnjeni. Obenem zaznavamo pomanjkanje inovativnih rešitev oziroma nezadostno infrastrukturno podporo za zbiranje in integracijo podatkovnih virov, ki lahko omogočijo kvantifikacijo vpliva posegov v urbano okolje na zdravje ljudi. Takšna infrastruktura bi odločevalcem omogočila podatkovno podprto sprejemanje odločitev o posegih v urbanem okolju in spodbujala prehod v bolj zdravo družbo.

V pričujoči raziskavi smo naredili pomemben korak k omogočanju kvantifikacije povezav med parametri urbanega okolja in razvojem nenalezljivih bolezni. Osredotočili smo se na iskanje in zbiranje podatkov iz javno dostopnih in odprtih podatkovnih baz, tako s področja pokazateljev in glavnih povzročiteljev onesnaženosti v urbanem okolju, kot tudi s področja zdravja in razširjenosti določenih bolezni. Primere okoljskih podatkov predstavljajo števci prometa, vremenski podatki in podatki merilnih postaj o onesnaženosti zraka. Ti so večinoma javno dostopni za različna mesta in regije, tako preko podatkovnih portalov in iniciativ državljanske znanosti, kot so UTD19 [3] in Sensor Community [4], kot tudi preko državnih agencij, kot je npr. Agencija Republike Slovenije za okolje (ARSO), ter mestnih občin, kot je npr. Mestna občina Ljubljana. Zdravstveni podatki so občutljivejši, zato smo jih po eni strani pridobili iz javno dostopnih podatkovnih baz na nacionalnem nivoju, prav tako pa smo raziskali odprte repozitorije zdravstvenih podatkov za tuja mesta, države in skupnosti. Na ta način smo pridobili več heterogenih podatkov, ki jih lahko uporabimo pri iskanju povezanosti med večletnimi trendi okoljskih pokazateljev, pojavnosti nenalezljivih bolezni in smrtnosti prebivalcev. Pri tem je ključno, da razpolagamo s podatki, ki med seboj povezujejo izbrane lokacije.

Vzporedno z zbiranjem podatkov smo se lotili vzpostavitve metodologije, ki omogoča smiselno združitev okoljskih in zdravstvenih podatkov. Vzpostavili smo računalniško ogrodje, ki podpira kvantifikacijo povezav med določenim okoljskim dejavnikom oziroma kombinacijo teh in določenim zdravstvenim pokazateljem. V ta namen smo uporabili določanje korelacijskih koeficientov in parcialnih korelacij ter regresijo na podlagi naključnih gozdov (angl. *random forest regression*). Razvoju in implementaciji metodologije je sledila vzorčna uporaba razvitega računalniškega orodja na testnih primerih iz okolja izbranih Evropskih regij po NUTS 2 [5].

Pridobljeni rezultati ponujajo številne možnosti za izvedbo analiz v širšem kontekstu s perspektive različnih področij, kot so podatkovno vodeno načrtovanje urbanega prostora in vpliv slednjega na zdravo družbo. Iz-

vorna koda razvitega orodja je dostopna na povezavi `https://github.com/SusTra/IDA4Health` pod licenco GPL-3.0, ki omogoča enostavno in brezplačno uporabo in nadgradnjo s strani vseh zainteresiranih deležnikov.

2 Metode

2.1 Izbor podatkov

Da bi zbrali informacije o razpoložljivih naborih podatkov, smo pregledali prosto dostopne podatkovne zbirke in ustrezne raziskovalne članke. Zbirke podatkov smo izbirali po naslednjih merilih: odprti dostop, natančnost, prostorska ločljivost, razpoložljivost zgodovinskih podatkov, razpoložljivost za vsaj eno evropsko državo, strojno berljiva oblika in časovna ločljivost vsaj enega leta. Zdravstveni podatki o nenalezljivih boleznih so morali zajemati glavne zdravstvene kazalnike, kot so umrljivost, incidenca, razširjenost ali dejavniki tveganja za izbrano bolezen. Okoljski podatki so morali zajemati dejavnike, ki domnevno vplivajo na zdravje ljudi (npr. onesnaženost zraka in hrup). Nato smo parametre razvrstili glede na njihovo pomembnost za naše raziskave, pri čemer je bila visoka prostorska ločljivost najpomembnejša. Na koncu smo pripravili prednostni seznam zbirk podatkov.

Razpoložljivi zdravstveni in okoljski podatki izhajajo iz posameznih epidemioloških študij (dostopnih preko portalov, ko je npr. *European Open Science Cloud* – EOSC [6]) ter velikih zbirk podatkov na nacionalni in mednarodni prostorski ravni (npr. podatkovne zbirke Nacionalnega inštituta za javno zdravje – NIJZ). Podatkovne zbirke na področju zdravja, ki ustrezajo našim prednostnim merilom, nudijo časovno ločljivost na letni ravni in imajo prostorsko ločljivost na ravni mesta/občine/regije. Zbirke, ki vključujejo nekatere zbirke podatkov o umrljivosti, prevalenci in incidenci, so določeni podatkovni seti iz zbirk EUROSTAT (angl. *European statistics*) [7], SLORA (Register raka Republike Slovenije) [8, 9], EUROCAT (angl. *European network of population-based registries for the epidemiological surveillance of congenital anomalies*)
[10], NIJZ [11] in podatkovne zbirke GBD (angl. *Global Burden of Disease*) [12].

Podatkovne zbirke o okolju so na voljo z dnevno/mesečno/letno časovno ločljivostjo in večinoma omejene na mesta, vključno s podatki o kakovosti zraka iz portala Sensor Community [4], European Air Quality Portal [13] in WHO Air Quality Database [14], podatki o hrupu iz Sensor Community, podatki o svetlobnem onesnaževanju iz LAADS DAAC [15] ter podatki o sončnem sevanju in indeksu UV iz portala Visual Crossing [16].

Najobsežnejši demografski podatki so na voljo na portalu EUROSTAT [7], podatki o kakovosti življenja pa v zbirki Numbeo [17]. Zdravstveni in okoljski podatki iz posameznih nacionalnih podatkovnih zbirk imajo približno enako kakovost in precej natančnejšo prostorsko ločljivost kot zgoraj omenjene zbirke, vendar je za zbiranje podatkov iz nacionalnih zbirk potrebno več truda.

Kot vodilni vir raziskave smo izbrali podatke o boleznih s portala EUROSTAT [7]. EUROSTAT zagotavlja celovite in zanesljive podatke, ki so v svoji metodologiji poenoteni med državami. Eden glavnih razlogov za izbiro portala EUROSTAT kot primarnega vira podatkov je vsebovanost baz z visoko prostorsko ločljivostjo (npr. na ravni regij NUTS 2 [5]). Ugotovili smo, da zdravstvenim podatkom iz drugih virov pogosto manjka stopnja prostorske ločljivosti, ki je potrebna za dosego zastavljenih ciljev. Z dajanjem prednosti naborom podatkov z visoko prostorsko ločljivostjo želimo pridobiti natančnejši vpogled v odnos med okoljskimi dejavniki in nenalezljivimi boleznimi. Primer analize, ki je služila kot prikaz koncepta, temelji na EUROSTAT-ovih regijah NUTS 2 [5]. Vsi drugi nabori podatkov so bili nato prilagojeni tej obliki, pri čemer smo izračunali povprečje razpoložljivih podatkov na ravni regije, v kateri je bil lociran posamezen vir. Tak primer so bili podatki o kvaliteti zraka E1a in E2a (angl. (*Air quality time series (E1a & E2a data sets)*), ki so dostopni preko portala *European Environment Agency* [18] in vsebujejo zgodovinske podatke o onesnaženosti zraka za številna večja evropska mesta.

Ker so bili zdravstveni podatki EUROSTAT naše izhodišče, so zahtevali najmanj predhodne obdelave. Posledično je bila s tega portala pridobljena tudi velika količina podatkov o prebivalstvu, na primer bruto domači proizvod (BDP) na prebivalca. Regije smo omejili na tiste, ki imajo manj kot 5.000.000 prebivalcev, da bi poskušali zagotoviti čim večjo stopnjo homogenosti znotraj vsake regije.

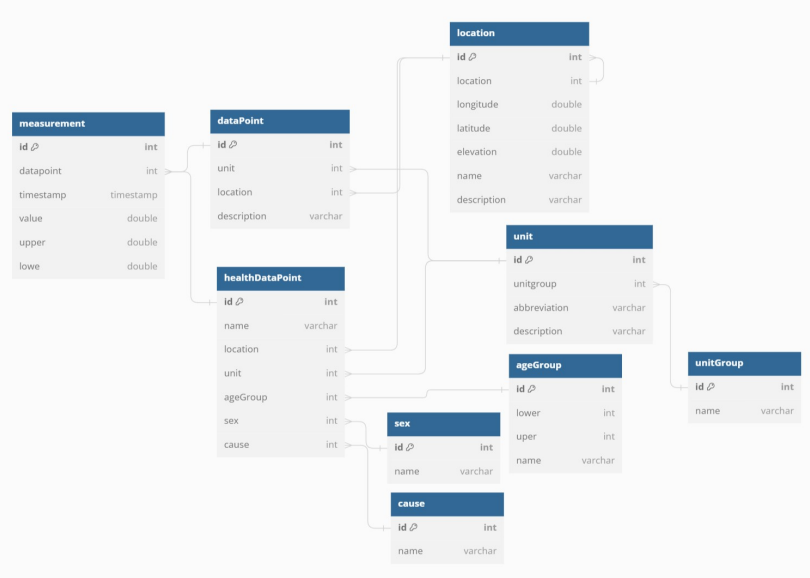
V naslednjih korakih so bili izbrani nabori podatkov umeščeni v računalniški okvir za pridobivanje ter integracijo podatkov in sklepanje vpliva okoljskih podatkov na zdravstvene pokazatelje, kar opišemo v nadaljevanju.

2.2 Zajem in shranjevanje podatkov

Za vsak vir podatkov smo implementirali lastne zajemalce podatkov (angl. *data scrapers*), ki so skrbeli za zajem in predobdelavo razpoložljivih podatkov ter njihovo shranjevanje v bazo. Iz slednje je orodje za sklepanje črpalo podatke za analizo.

Kompleksnost zajemalca je bila odvisna od zahtevnosti pridobivanja podatkov in zahtevnosti predprocesiranja. Nizi podatkov o kvaliteti zraka E1a in E2a vsebujejo zgodovinske meritve senzorjev v evropskih mestih. Geografska širina in dolžina sta bili za vsako mesto določeni s postopkom klicanja programskega vmesnika API (angl. *application programming interface*) za geokodiranje. Dobljeni rezultat smo nato potrdili s kosinusnimi podobnostmi (angl. *cosine similiarity*) in primerjavami endonimov, saj so se viri podatkov razlikovali v načinu poimenovanja mest (npr. Wien, Dunaj, Vienna). Za vsako zemljepisno širino in dolžino smo določili pripadajočo regijo po NUTS 2, za katero smo povprečili odčitke vseh senzorjev, ki se nahajajo v tej regiji. S tem smo pridobili 59 regij, ki vključujejo meritve za SO_2, 114 za PM_{10}, 101 za O_3, 124 za NO_4, 63 za CO in 101 za NO.

Pridobljene podatke smo shranili v centralizirano relacijsko bazo podatkov, iz katere je orodje za sklepanje pridobivalo podatke za analizo. Shema relacijske baze podatkov razvite rešitve je prikazana na sliki 1.



Slika 1: Shema centralizirane relacijske baze podatkov.

Najpomembnejši del baze so podatkovne točke (angl. *data points*). Posamezna podatkovna točka je opredeljena z enoto (kot na primer pojavnost na časovno enoto ali znesek na prebivalca) in lokacijo (regija/mesto, v katerem je ta meritev bila narejena) ter lahko vsebuje več različnih meritev in/ali podatkov (npr. starostna skupina, spol, vzrok). Primer podatkovne točke so podatki o smrtnosti zaradi raka na pljučih v zahodni Sloveniji za vse moške starejše od 65 let. Enota je v tem primeru stopnja pojavnosti (angl. *standardized rate*) glede na EUROSTAT, starostna skupina *65+*, spol *moški*, vzrok *rak na pljučih* in lokacija *zahodna Slovenija*. Za vsako leto, za katero imamo meritev, obstaja vnos *measurement* z datumom meritve ter izmerjeno vrednostjo.

2.3 Orodje za sklepanje

Orodje za sklepanje in analizo razmerja med okoljskimi dejavniki in indikatorji bolezni smo pripravili v jeziku Python in okolju Jupyter Notebook. Uporabnik najprej določi parametre, ki ga zanimajo, nato pa orodje dostopa do centralizirane baze podatkov in izbere ustrezne podatke. Parametri, med katerimi lahko uporabnik izbira, so bolezen, okoljski dejavniki, leto, spol in starostna skupina.

Predstavitev podatkov z razsevnim grafom omogoča celovito razumevanje razmerja med okoljskimi dejavniki in nenalezljivimi boleznimi, saj lahko uporabniki opazujejo trende, vzorce in morebitna odstopanja. Poleg tega orodje vključuje različne statistične analize za oceno stopnje povezanosti in pomembnosti opazovanih korelacij. Ena izmed teh je parcialna regresijska analiza, ki omogoča preučevanje prispevka posameznega okoljskega dejavnika ob nadzorovanju drugih spremenljivk (Slika 2). S tem lahko izoliramo specifičen vpliv posameznih dejavnikov na izid bolezni. Poleg tega orodje omogoča izračun ključnih statističnih mer, kot sta korelacijski koeficient (R), ki določi stopnjo povezanosti spremenljivk, in koeficient determinacije (R^2), ki ovrednoti količino variabilnosti, pojasnjene z modelom.

Z uporabo Pythonovih knjižnic in ogrodij, kot so pandas, NumPy in scikit-learn, je naše orodje enostavno nadgraditi z drugimi naprednimi statističnimi analizami in implementacijami algoritmov strojnega učenja. Jupyter Notebook omogoča enostavno uporabniško izkušnjo, vizualizacijo rezultatov in dokumentacijo analize.

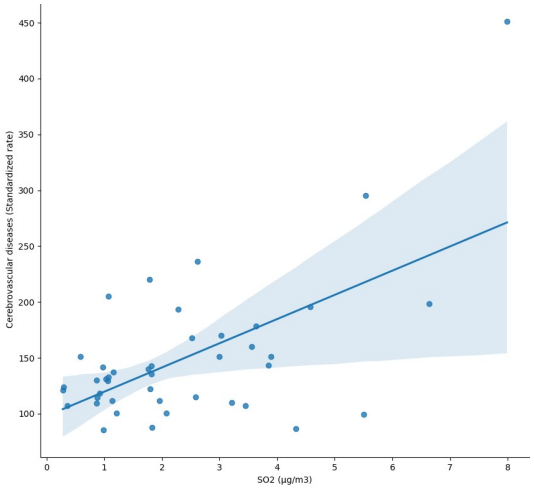
3 Primer uporabe: Povezanost koncentracij SO_2 v zraku in smrtnosti zaradi cerebrovaskularnih bolezni

S pripravljenim orodjem smo analizirali vpliv onesnaženosti zraka na smrtnost zaradi cerebrovaskularnih bolezni v okviru 59 regij. Izračun parcialnih korelacij je pokazal, da na smrtnost zaradi cerebrovaskularnih bolezni najmočneje vpliva koncentracija SO_2 (slika 2), kar je skladno z obstoječo literaturo [19].

V nadaljevanju smo postavili regresijski model na osnovni naključnih gozdov, pri čemer smo regresijo standardizirane smrtnosti zaradi cerebrovaskularnih bolezni izvajali na osnovi koncentracije NO_2, O_3, SO_2 in BDP na prebivalca. Uporabili smo implementacijo scikit-learn naključnih gozdov, pri čemer smo učenje izvajali na gozdovih s pet drevesi z maksimalno globino štirih vozlišč. Model smo testirali z uporabo prečnega preverjanja, pri čemer je bila povprečna vrednost metrike R^2 na učnih množicah enaka 0,70, na testnih pa 0,59. Kot daleč najbolj pomembni značilki sta se pri napovedovanju izkazali BDP na prebivalca in koncentracija SO_2 v zraku.

4 Zaključki

Računalniško ogrodje za integrativno analizo okoljskih in zdravstvenih podatkov omogoča avtomatski zajem, predprocesiranje, shranjevanje in analizo heterogenih odprtih podatkov. Razvito ogrodje zaenkrat podpira zgolj osnovne statistične analize, vendar je enostavno nadgradljivo z



Slika 2: Parcialna korelacija med koncentracijo SO_2 in smrtnostjo zaradi cerebrovaskularnih bolezni. Parcialni korelacijski koeficient je znašal 0,48, kar nakazuje zmerno korelacijo. Modro območje predstavlja interval zaupanja regresijske premice.

naprednejšimi metodami strojnega učenja.

Kljub temu, da so podobna orodja v preteklosti že bila razvita, so slednja omejena na predhodno zbrane podatke in/ali vnaprej definirane lokacije (npr. Environment and Health Atlas [20]) ali pa omogočajo sklepanje zgolj na že vnaprej pripravljenih modelih (npr. AirQ+ [21]). Za razliko od teh, naše orodje temelji izključno na realnih odprtih podatkih, kar omogoča njegovo enostavno prilagodljivost na poljubne nabore podatkov za poljubne lokacije.

Omejena dostopnost odprtih zdravstvenih podatkov otežuje izvajanje analiz, zaradi česar bi bilo smotrno učenje modela izvajati tudi na zaprtih podatkih. Možna razširitev vključuje tudi integracijo zajemalcev podatkov na podlagi vsebine znanstvenih člankov in objavljenih raziskav, kar bi bilo moč izvesti s tehnikami za obdelavo naravnega jezika (angl. *natural language processing*, NLP). Na ta način bi lahko v napovedne modele vključili dodatna strokovna znanja, ki bi zvišala njihovo napovedno moč in s tem uporabnost v različnih scenarijih.

Modularna zasnova orodja omogoča preprosto nadaljnjo nadgradnjo in razširjanje nabora metod ter podatkov za integrativno analizo okoljskih in zdravstvenih podatkov v prihodnosti. Nadaljnje delo bo usmerjeno v iskanje dodatnih podatkovnih zbirk in vrst podatkov (npr. svetlobno onesnaženje, hrup in vreme) in njihovo vključitev v orodje. Hkrati bomo preučili možnosti implementacije kompleksnejših statističnih metod (npr. faktorska analiza, večnivojska analiza, strukturno-enačbeni modeli) in metod strojnega učenja (npr. metoda podpornih vektorjev, Bayesova omrežja) v kombinaciji z metodami razlagalne umetne inteligence, ki bodo omogočale bolj poglobljeno analizo.

Zahvala

Pričujoče delo je nastalo v okviru projekta Integrativna analiza zdravstvenih in okoljskih podatkov za zdravo družbo, ki ga sofinancirata Republika Slovenija in Evropska unija iz Evropskega socialnega sklada.

Literatura

[1] European Environment Agency, "EEA Report No 21/2019." https://www.eea.europa.eu/publications/healthy-environment-healthy-lives, 2020.

[2] A. A. Shafi and K. E. Knudsen, "Cancer and the circadian clock," *Cancer research*, vol. 79, no. 15, pp. 3806–3814, 2019.

[3] A. Loder, L. Ambühl, M. Menendez, and K. W. Axhausen, "Understanding traffic capacity of urban networks," *Scientific reports*, vol. 9, no. 1, p. 16283, 2019.

[4] Sensor.Community Team, "Sensor.Community." https://sensor.community/, 2023.

[5] Eurostat, "Statistical regions in the european union and partner countries: NUTS and statistical regions 2021: 2020 edition," tech. rep., European Union, 2020.

[6] EOSC Portal, "European Open Science Cloud." https://eosc-portal.eu/, 2023.

[7] European Commission, "Eurostat." https://ec.europa.eu/eurostat/, 2023.

[8] V. Zadnik, M. P. Zakelj, K. Lokar, K. Jarm, U. Ivanus, and T. Zagar, "Cancer burden in slovenia with the time trends analysis," *Radiology and oncology*, vol. 51, no. 1, pp. 47–55, 2017.

[9] V. Zadnik and T. Žagar, "SLORA: Slovenija in rak. Epidemiologija in register raka. Onkološki inštitut Ljubljana.." http://www.slora.si/, 2023.

[10] European Commission, "EUROCAT: European network of population-based registries for the epidemiological surveillance of congenital anomalies." https://eu-rd-platform.jrc.ec.europa.eu/eurocat_en, 2023.

[11] NIJZ, "Podatkovne zbirke in raziskave." https://nijz.si/podatki/podatkovne-zbirke-in-raziskave/podatkovne-zbirke-in-raziskave/, 2023.

[12] Institute for Health Metrics and Evaluation, "Global Burden of Disease (GBD)." https://www.healthdata.org/gbd, 2023.

[13] "EEA: European Air Quality Portal." https://aqportal.discomap.eea.europa.eu/, 2023.

[14] "WHO Air Quality Database." https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database, 2023.

[15] "LAADS DAAC." https://ladsweb.modaps.eosdis.nasa.gov/, 2023.

[16] "Visual Crossing." https://www.visualcrossing.com/, 2023.

[17] Mladen Adamovic, "Numbeo: Cost of Living." https://www.numbeo.com/, 2023.

[18] European Environment Agency, "European Environment Agency Portal." https://www.eea.europa.eu/, 2023.

[19] N. Sang, Y. Yun, H. Li, L. Hou, M. Han, and G. Li, "SO2 inhalation contributes to the development and progression of ischemic stroke in the brain," *Toxicological Sciences*, vol. 114, no. 2, pp. 226–236, 2010.

[20] A. L. Hansell, L. A. Beale, R. E. Ghosh, L. Fortunato, D. Fecht, L. Jarup, and P. Elliott, *The environment and health atlas for England and Wales*. Oxford University Press, USA, 2014.

[21] World Health Organization, "AirQ+: software tool for health risk assessment of air pollution," 2018.