univerza v ljubljani

ekonomska fakulteta

DISPOZICIJA MAGISTERSKEGA DELA

Priložnosti uporabe analitike masovnih podatkov za trajnostno upravljanje z vodnimi viri

Ljubljana, marec 2021 KLEMEN KOŽELJ, 19189377

**Naslov magistrskega dela**

Priložnosti uporabe analitike masovnih podatkov za trajnostno upravljanje z vodnimi viri

Opportunities of big data analytics for sustainable water resource management

**Predmet**

Statistika ocenjuje, da naj bi vsak evropejec porarabil pribljižno 100 litrov pitne vode dnevno. Tudi Evropska agencija za okolje ugotavlja, da je Evropa z vodo bogat kontinent. A zaskrbljujoča je ugotovitev, da se je od leta 1960 do 2010 zaloga obnovljivih virov vode na prebivalca v Evropi zmanjšala v povprečju za kar 24 odstotkov. Temu botrjuje več dejavnikov, od povečanja prebivalstva do rasti turizma in drugih gospodarksih dejavnosti. Stanje je najslabše na jugu Evrope kjer letna sušna obdobja postajajo stalnica, leta 2014 pa je kar 40 odstotkov ljudi na območju Mediterana živelo s kritično majhno količino vode. (European Environment Agency, 2020)

Žal je drugod po svetu stanje še slabše svetovna zdravstvena organizacija WHO ocenjuje, da trenutno 2 miljardi ljudi po svetu nima dostopa do sveže pitne vode. Hkrati napovedujejo tudi poslabšanje trenutne situacije in v bližni prihodnosti leta 2025 bi se lahko ta številka povspela na kar 4 miljarde oziroma skoraj polovico svetovnega prebivalstva. (World Health Organization, 2019)

Upravljanje z vodnimi viri je področje pod katerega spadajo aktivnosti načrtovanja in izgradnje vodne infrastrukture ter tudi nadzor in distribucija vodnih virov. Evropska unija se že dolgo zaveda problematike povezane z vodnimi viri in se zato z različnimi iniciativami in zakoni že dolgo zavzema, za njihovo učinkovito in trajnostno upravljanje. Leta 2012 je tako Evropska komisija predstavila načrt za varovanje evropskih vodnih virov v katerih postavlja minimalne standarde za kvantiteto in kvaliteto voda. (Kurrer, 2021)

Ključen del rešitve, ki nam bo pomagala bolje upravljati in nadzirati vodne vire, bo odigrala informacijska tehnologija. To prepoznava tudi Italiansko podjetje Acea, ki je vodilni ponudnik vodovodnih in storitev v centralni Italiji. Podjetje od leta 1999 kotira na Italianski borzi in ima trenutno 9 milionov odjemalcev. Acea je decembra 2020 na spletnem portalu Kaggle objavilo izziv v katerem pozivajo raziskovalce, da z podatki njihovih vodnih virov izdelajo napovedovalne modele za napovedovanje količine vode. Z uspešnimi napovednimi modeli bi podjetje lahko bolje preprečevalo poplave ali pa pričeli opozarjati na varčevanje z vodo še preden njena raven pade pod kritično mejo. O tem, kako pomembna se jim zdi rešitev tega problema, priča tudi dejstvo da so na izzvu za primerno rešitev ponudili $25,000 nagrade. (Acea Group, 2020)Različni raziskovalni programi Evropske unije ponujajo še več priložnosti za inovativne in interdisciplinarne rešitve ki bi pomagali izboljšati upravljanje z vodnimi viri. Tako recimo Evropska vesoljska agencija ESA v okviru programa Copernikus s kar 30-imi posebnimi sateliti izvaja slikanje celotne zemlje, posebno zanimiva sta 2 satelita imenovana Sentinel-2, ki kar s 13 senzorji zajemata različne podatke iz površja zemlje (European Space Agency, 2015). Zbrani podatki programa Copernikus so nato prosto dostopni javnosti v raziskovalne namene.

Podatki s satelitov Sentinel-2 so primarno namenjeni analiziranju zemeljske površine, izjemno pogosto pa se uporabljajo v sinergiji z različnimi algoritmi strojnega učenja. Učinkovitost oziroma zmožnost uporabe različnih algoritmov strojnega učenja na podatkih Sentinel-2 so na primer raziskovali v letu 2020 na Švedski univerzi Lund (Abdi, 2020) in ugotovili ogromen potencial. Podobno raziskavo so v istem letu izvajali na Universitat Politècnica de València kjer so zaznavali majnše vodne površine in zemljišča ki so ogrožena v primeru poplavljanja (Pena-Regueiro, Sebastiá-Frasquet, Estornell, & Aguilar-Maldonado, 2020).

Obe zgoraj navedeni raziskavi sta pokazala, da so podatki iz satelitov Sentinel-2 odličen vir podatkov za širše opazovanje in različne makro-analize zemeljskega površja. Odprto pa ostaja vprašanje, ali bi lahko podatke uporabili tudi za bolj natančna opazovanja vodnih virov ter tako poskusili z njimi reševati podobne probleme, kot jih ima podjetja Acea. Acea trenutno pridobiva podatke o rekah in jezerih iz različnih senzorjev in v kolikor bi lahko senzorje oziroma strojno opremo nadomestili zgolj s programsko opremo za procesiranje slik s satelitov, bi to za njih pomenilo ogromne operativne prihranke, hkrati pa bi lahko analizirali veliko širše geografsko področje ter tako dobili bolj celovito analizo stanja na terenu.

Na tej podlagi se postavlja tudi osrednje raziskovalno vprašanje tega magistrskega dela, ki je preučitev možnosti uporabe strojnega učenja in masovnih podatkov, kot so satelitski posnetki, za zagotavljanje trajnostnega upravljanja z vodnimi viri.

**Namen**

Opolnomočenje upravljalcev vodnih virov s kakovostnimi napovednimi modeli bi lahko pripomoglo k bolj pametni, predvsem pa naravi prijazni in trajnostni izrabi vodnih virov, saj bi lahko odločevalci svoje odločitve sprejemali na podlagi matematičnih modelov in ne subjektivnih mnenj.

Podoben pozitiven učinek bi lahko imelo izkoriščanje prosto dostopnih podatkovnih baz kot jih na primer ponuja Evropski vesoljski program Copernicus. Z uporabo podatkov s satelitov Sentinel-2 bi lahko izboljšali prvotne napovedi in v najbolšem primeru celo nadomestili trenutne senzorje za gladino vode ter tako močno znižali operativne stroške nadzorovanja vodne gladine.

**Cilj**

Cilj magistrskega dela bo sprva preučiti obstoječe relevantne raziskave na temo strojnega učenja in preučiti, kako so drugi raziskovalci uporabljali Sentinel-2 satelitske posnetke za analiziranje vodne gladine na zemlji, ter tako pridobiti širše razumevanje, kako so se podobnih problemov lotili drugi raziskovalci.

V drugem delu bom poskusil z različnimi tehnikami razviti inovativno programsko rešitev za analiziranje satelitskih slik in napovedovanje količine vode. Celoten razvoj prototipa bo sledil naslednjim korakom:

* Analiza podatkov in njihova priprava za strojno učenje.
* Razvoj napovednega modela za napovedovanje količine vode glede na zgodovinske in druge konvencionalne podatke.
* Razvoj programa za pridobivanje satelitskih posnetkov ter iz pridobljenih podatkov izluščiti uporabne informacije za izboljšavo zgoraj omenjenega napovednega modela.
* Izvedel bom eksperiment ter ugotovil, ali lahko satelitski posnetki služijo kot alternativa trenutnim senzorjem za izmero vodne gladine.

**Opis metodologije**

Za izhodišče raziskave bom vzel podatkovno zbirko, ki jo je na portalu Kaggle objavil zgoraj omenjeni Italianski ponudnik vodovodnih storitev Acea. Podatkovna zbirka zajema podatke od leta 1998 do sredine 2020 in vključuje različne tipe vodnih teles kot so jezera, reke, podzemna zajetja in izviri. Vsako vodno telo je opisano z različnimi atributi o količini vode, priloženi pa so tudi nekateri drugi osnovni podatki kot so na primer količina padavine.

Za obdelavo podatkov, izgradnjo modelov in vizualizacijo rezultatov bom uporabljal programski jezik Python 3 z grafičnim vmesnikom Jupyter. Za programski jezik Python sem se odločil ker ponuja izjemno bogat odprtokodni ekosistem različnih programskih knjižnic in orodij za delo z podatki. Vmesnik Jupyter pa omogoča pisanje programske kode ter njeno interaktivno izvajanje in izpisovanje rezultatov.

Na začetku razvoja prototipa bom uporabljal le podatke podjetja Acea, osredotočil se bom le na površinske vodna telesa saj za podzemne ne morem kasneje uporabiti satelitskih posnetkov. Podatke bom analiziral, iz njih izločil šume in jih ustrezno pripravil za izgradnjo različnih regresijskih napovelovalnih modelov.

Natančnost vseh modelov in algoritmov bom ocenjeval glede na primerjavo s pravimi vrednostmi, razliko bom izračunal z navadno in kvadrirano absolutno povprečno napako saj je bil tako določen tudi kriterij uspešnosti v objavljenemu Kaggle izzivu.

Podatke iz satelitov bom pridobil od ponudnika oblačnih storitev Sentinel Hub, ki omogoča dostopanje do Sentinel-2 podatkov preko več spletnih interaktivnih programskih vmesnikov. Ponudnik omogoča tudi poizvedovanje po času in kraju ter minimalno pre-procesira podatke na primer odstrani šum ob stičišču dveh satelitskih slik. Podatki so naposled podani kot multidimenzionalne matrike števil.

Podobno kot z podatki podjetja Acea bom tudi v primeru Sentinel-2 satelitskih posnetkov analiziral pridobljene podatke, odstranil šume ter nato poskusil razviti algoritem za ugotavljanje količine vode v vodnemu telesu. Naposled bom analiziral kako lahko z pridobljenimi informacijami izboljšam prvotno izdelani napovedovalni model ter oceniti ali bi lahko razviti algoritem nadomestil strojne senzorje podjetja Acea.

**Struktura dela:**

**UVOD**

1. **UPRAVLJANJE Z VODNIMI VIRI**
   1. Upravljanje z vodnimi viri
   2. Vodna bilanca
2. **RAZVOJ NAPOVEDOVALNEGA MODELA**
   1. Analiza podatkov v surovi obliki
   2. Priprava podatkov za strojno učenje
   3. Izbira in izgradnja napovedovalnega modela
   4. Ocenjevanje modela
3. **SATELITSKI POSNETKI**
   1. Program Copernikus in satelita Sentinel-2
   2. Pridobivanje satelitskih posnetkov
   3. Procesiranje satelitskih posnetkov
   4. Detekcija količine vode
   5. Ocenjevanje pridobljenih informacij
   6. Obogatitev napovedovalnega modela
   7. Substitucija merilcev za gladino vode
4. **DISKUSIJA**
   1. Ključne ugotovitve
   2. Priporočila za nadaljno raziskovanje

**SKLEP**

**LITERATURA IN VIRI**

**Literatura in viri**

1. Abdi, A. M. (2020). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience and Remote Sensing*, *57*(1). <https://doi.org/10.1080/15481603.2019.1650447>
2. Acea Group. (2020, december 10). Acea Smart Water Analytics. Pridobljeno 3. januar 2021 iz <https://www.kaggle.com/c/acea-water-prediction>
3. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Müller, A. C., Grisel, O., … Varoquaux, G. (b. d.). *API design for machine learning software: experiences from the scikit-learn project*. Pridobljeno iz <https://github.com/scikit-learn>
4. eo-learn Revision. (2018). eo-learn 0.8.0 documentation. Pridobljeno 28. februar 2021 iz https://eo-learn.readthedocs.io/en/latest/
5. European Environment Agency. (2018, november 12). Sustainable water management. Pridobljeno 16. februar 2021 iz <https://www.eea.europa.eu/themes/water/european-waters/water-management>
6. European Environment Agency. (2020, november 23). Water use and environmental pressures. Pridobljeno 16. februar 2021 iz <https://www.eea.europa.eu/themes/water/european-waters/water-use-and-environmental-pressures/water-use-and-environmental-pressures#toc-2>
7. European Environment Agency. (2018). Water in the economy : users and abusers. Pridobljeno 11. februar 2021 iz <https://www.eea.europa.eu/signals/signals-2018-content-list/signals-2018-content-list/articles>
8. European Space Agency. (2015). *Sentinel-2 User Handbook* (1. izd.; B. Hoersch, Ur.). European Space Agency. <https://doi.org/10.1021/ie51400a018>
9. Eurostat. (2018). Water productivity. Pridobljeno 15. februar 2021 iz <https://ec.europa.eu/eurostat/web/products-datasets/-/t2020_rd210>
10. Geissen, V., Mol, H., Klumpp, E., Umlauf, G., Nadal, M., van der Ploeg, M., … Ritsema, C. J. (2015). Emerging pollutants in the environment: A challenge for water resource management. *International Soil and Water Conservation Research*, *3*(1), 57–65. <https://doi.org/10.1016/j.iswcr.2015.03.002>
11. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with {NumPy}. *Nature*, *585*(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
12. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
13. Kurrer, C. (2021, januar). Water protection and management. Pridobljeno 25. februar 2021 iz <https://www.europarl.europa.eu/factsheets/en/sheet/74/water-protection-and-management>
14. Lettenmaier, D. P., Wood, A. W., Palmer, R. N., Wood, E. F., & Stakhiv, E. Z. (1999). Water resources implications of global warming: A U.S. regional perspective. *Climatic Change*, *43*(3), 537–579. <https://doi.org/10.1023/A:1005448007910>
15. McKinney, W. (2010). Data Structures for Statistical Computing in Python. V S. van der Walt & J. Millman (Ur.), *Data Structures for Statistical Computing in Python* (str. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
16. Metzger, E., Reig, P., H. Wen, W., S. Young, R., & Owens, B. (2016). *Water-Energy Nexus: Business Risks and Rewards* (P. Mary & L. Jen, Ur.). Washington, DC 20002, ZDA: World Resources Institute. Pridobljeno iz <https://www.wri.org/publication/water-energy-nexus?downloaded=true>
17. Molden, D. (2020). Scarcity of water or scarcity of management? *International Journal of Water Resources Development*, *36*(2–3), 258–268. <https://doi.org/10.1080/07900627.2019.1676204>
18. Moss, K., & Debora, F. (2016, januar 25). Solving the Twin Crises of Energy and Water Scarcity. Pridobljeno 15. februar 2021 iz <https://hbr.org/2016/01/why-energy-and-water-scarcity-are-connected-chall>
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
20. Pena-Regueiro, J., Sebastiá-Frasquet, M. T., Estornell, J., & Aguilar-Maldonado, J. A. (2020). Sentinel-2 application to the surface characterization of small water bodies in Wetlands. *Water (Switzerland)*, *12*(5). <https://doi.org/10.3390/w12051487>
21. Pérez, F., & Granger, B. E. (2007). IPython: a System for Interactive Scientific Computing. *Computing in Science and Engineering*, *9*(3), 21–29. <https://doi.org/10.1109/MCSE.2007.53>
22. Perktold, J., Seabold, S., & Taylor, J. (2021, februar 2). API Reference — statsmodels. Pridobljeno 28. februar 2021 iz <https://www.statsmodels.org/stable/api.html>
23. Project Jupyter. (2021, februar 8). Jupyter Documentation. Pridobljeno 28. februar 2021 iz <https://jupyter.org/documentation>
24. S. Lubber, Mi. (2009, marec 10). Growing Business Risks from Global Water Scarcity. Pridobljeno 15. februar 2021 iz <https://hbr.org/2009/03/growing-business-risks-from-gl>
25. Taylor, S. J., & Letham, B. (2017). Business Time Series Forecasting at Scale. *PeerJ Preprints 5:e3190v2*, *35*(8), 48–90. Pridobljeno iz <https://peerj.com/preprints/3190/%0Ahttp://ezproxy.bangor.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=108935824&site=ehost-live%0Ahttps://peerj.com/preprints/3190/%0Ahttps://peerj.com/preprints/3190.pdf>
26. The European Federation of National Water Services. (2017). *Europe’s water in figures*. The European Federation of National Water Services. Pridobljeno od <https://www.danva.dk/media/3645/eureau_water_in_figures.pdf>
27. The pandas development team. (2020, februar). *pandas-dev/pandas: Pandas*. Zenodo. <https://doi.org/10.5281/zenodo.3509134>
28. Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
29. World Health Organization. (2019, junij 14). Drinking water. Pridobljeno 15. februar 2021 iz <https://www.who.int/news-room/fact-sheets/detail/drinking-water>
30. Zubaidi, S. L., Ortega-Martorell, S., Al-Bugharbee, H., Olier, I., Hashim, K. S., Gharghan, S. K., … Al-Khaddar, R. (2020). Urban water demand prediction for a city that suffers from climate change and population growth: Gauteng province case study. *Water (Switzerland)*, *12*(7), 1–17. <https://doi.org/10.3390/W12071885>