PROFILIRANJE NAPOVEDLJIVOSTI KEMIJSKIH PARAMETROV V VODAH

POLONA KREN,[1] NIKA FLAKUS,[1] EVA KUHAR,[1]
POLONA MIHORKO[3] IN DRAGO BOKAL[1,2]

[1] Univerza v Mariboru, Fakulteta za naravoslovje in matematiko, Maribor, Slovenija.
E-pošta: polona.kren@student.um.si, nika.flakus@student.um.si, eva.kuhar@student.um.si, drago.bokal@um.si
[2] DataBitLab d.o.o., Maribor, Inštitut za matematiko, fiziko in mehaniko, Ljubljana, Slovenija.
E-pošta: drago.bokal@um.si
[3] Agencija Republike Slovenije za okolje, Ministrstvo za okolje in prostor, Ljubljana, Slovenija.
E-pošta: polonca.mihorko@gov.si

Povzetek Agencija Republike Slovenije za okolje že od leta 1987 spremlja kakovost podzemne vode v Sloveniji. V merilno mrežo so vključeni vodnjaki, vrtine in kraški izviri. Na vsaki postaji se vzorčijo vzorci podzemne vode in v njih analizirajo osnovni parametri, kovine, pesticidi, lahkohlapne organske spojine, v zadnjih letih tudi novodobna onesnaževala (ostanki zdravil, hormoni, …). Za prispevek so izbrani podatki o vsebnosti kisika, nitratov, kloridov in sulfatov v vodi ter podatki o električni prevodnosti. V prispevku predstavimo profiliranje napovedljivosti koncentracij kemijskih parametrov. Ugotoviti želimo, pri katerih postajah so trendi naraščanja oz. padanja bolj izraziti in bolj stabilni in opredeliti, kakšni morajo biti podatki za zanesljivo oceno trenda. Zanima nas torej, kako pogosto in v kakšnem časovnem razmiku je najbolj smiselno analizirati parameter v vodi, da bomo lahko čim bolj zanesljivo napovedali trend spreminjanja.

Ključne besede: odločitvena drevesa, linearna regresija, napoved trendov, dendrogram.

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.fov.3.2022.34
ISBN 978-961-286-583-2

PROFILING PREDICTABILITY OF CHEMICAL PARAMETERS IN WATERS

POLONA KREN,[1] NIKA FLAKUS,[1] EVA KUHAR,[1]
POLONA MIHORKO[3] & DRAGO BOKAL[1,2]

[1] University of Maribor, Faculty of natural sciences and mathematics, Maribor, Slovenia.
E-mail: polona.kren@student.um.si, nika.flakus@student.um.si,
eva.kuhar@student.um.si, drago.bokal@um.si
[2] DataBitLab d.o.o., Maribor, Institute of Mathematics, Physics and Mechanics,
Ljubljana, Slovenia.
E-mail: drago.bokal@um.si
[3] Slovenian Environment Agency, Ministry of the Environment and Spatial Planning,
Ljubljana, Slovenia.
E-mail: polonca.mihorko@gov.si

Abstract Agency for environment of the Republic of Slovenia has been monitoring the quality of groundwater in Slovenia since 1987. The quality control includes wells, holes and karst springs. On every station, samples of underground water are being sampled and analyzed to the content of standard compounds like metals, pesticides, volatile organic compounds and, since last few years, modern contaminants (medicine residue, hormons, ect.). For this article, there is given data about containment of oxygen, nitrates, chlorides, sulphates in the water and its electrical conductivity. In this article is presented profiling of prediction of concentrations of chemical parameters. We wanted to find out, in which stations the trends are more prominent and more stable and to evaluate, what makes the data reliable to estimate the trend. In other words, we wanted to find out, how often it is reasonable to analyze the water in order to reliably predict the trend. We propose an answer using a decision tree obtained by analysing several statistics of the analysed population of measured time series data.

Keywords: linear regression, decision tree, trend predictability, dendrogram.

DOI https://doi.org/10.18690/um.fov.3.2022.34
ISBN 978-961-286-583-2

University of Maribor Press

1 Uvod

Za spremljanje kemijskega in ekološkega stanja voda v Sloveniji je odgovorna Agencija Republike Slovenije za okolje (ARSO). Monitoring stanja voda se izvajata na podlagi Zakona o vodah (Republika Slovenija, 2002), Zakona o varstvu okolja (Republika Slovenija, 2004) ter vrste podzakonskih aktov, ki v slovenski pravni red prenašajo zahteve evropskih direktiv s področja kakovosti površinskih in podzemnih voda. Slovenski predpisi, ki določajo način izvajanja monitoringa so Pravilnik o monitoringu stanja površinskih voda (Republika Slovenija, 2009a) in Pravilnik o monitoringu podzemnih vod (Republika Slovenije, 2009b), kriterije za oceno stanja voda pa predpisujeta Uredba o stanju površinskih voda (Republika Slovenije, 2009c) in Uredba o stanju podzemnih voda (Republika Slovenija, 2009d).

Vodna direktiva za vse države članice Evropske skupnosti postavlja enotne zahteve tako glede izvajanja monitoringa kot tudi glede ocenjevanja stanja voda. V Sloveniji program monitoringa kakovosti voda v skladu z zahtevami Vodne direktive poteka od leta 2007 dalje. Z uvedbo Vodne direktive so se spremenili tudi kriteriji in način ocenjevanja kakovosti voda. Za površinske vode se določa ekološko in kemijsko stanje, za podzemne vode pa kemijsko in količinsko stanje (Slika 1).

Slika 1: Shematični prikaz ocenjevanja stanja voda

Na vodnih telesih, ki ležijo na posebnih varstvenih območjih, se izvaja dodatni monitoring. Dodatni monitoring poteka tudi s sosednjimi državami in v okviru mednarodnih konvencij.

Vsako leto ARSO pripravi program monitoringa kakovosti voda za vsako vodno kategorijo posebej (reke, jezera, morje, podzemne vode), katerih posamezne dele nato v skladu s pogodbo izvedejo različni zunanji izvajalci. V programu so definirana merilna mesta, parametri in pogostost analiz. Predpisana je tudi metodologija vzorčenja in zahteve za uporabljene analitske metode. Podatki se po predhodni kontroli shranjujejo v bazi podatkov in so osnova za ocene kemijskega in ekološkega stanja.

Ocena stanja voda predstavlja izhodišče za pripravo ukrepov, na osnovi katerih bodo vodna telesa površinskih in podzemnih voda dosegla dobro stanje. Za vrednotenje pripravljenih ukrepov pa je potrebno razumeti predvidljivost njihovih posledic na izboljšanje stanja. Študiji te predvidljivosti je namenjen pričujoč prispevek, v katerem s pomočjo modeliranja časovnih vrst merjenih parametrov na izbranih lokacijah preverjamo statistične karakteristike njihovih trendov.

Intuitivno domnevamo, da več kot bo podatkov, bolj kakovostne ocene trenda bomo lahko na njihovi osnovi pridobili. Več podatkov lahko dosežemo z daljšim merilnim obdobjem ali s pogostejšimi meritvami. Za podatke, ki jih imamo na voljo, nimamo vpliva na trajanje obdobja zbiranja ali na dolžino časovnega intervala med meritvami, zanima pa nas, ali lahko iz variabilnosti obeh parametrov na znani populaciji izluščimo koristno informacijo o vplivu parametrov na kakovost napovedi.

V literaturi podobne raziskave kakovosti ocenjevanja trendov na velikem številu časovnih vrst, ki bi jo lahko vzeli za izhodišče naše raziskave, nismo zasledili. Ker je podatkov za vsako časovno vrsto relativno malo (največ nekaj deset) in ker je časovnih vrst relativno veliko (skoraj 200), obenem pa iz nekaj meritev na leto ne moremo izluščiti ključne, letne sezonskosti, smo za ocenjevanje trendov vzeli najpreprostejši model – linearno regresijo. Podatki, nad katerimi regresijo izvajamo, so opisani v razdelku 2. Metodologija, ki smo jo razvili za vrednotenje kakovosti ocenjevanja trendov časovnih vrst, je opisana v razdelku 3. Rezultati analize podatkov so predstavljeni v razdelku 4. Ključne ugotovitve predstavimo v razdelku

5. Med drugim ugotovimo, da že parametri kakovosti preprostega modela napovedovanja vrednosti z linearno regresijo porodijo zanimive, relevantne skupine časovnih vrst primerljivih stopenj kakovosti – skupina časovnih vrst z visoko nepojasnjeno varianco (najslabša kakovost), skupina časovnih vrst z nizko varianco in visoko vrednostjo p (srednja kakovost) in skupina časovnih vrst z nizko varianco in nizko vrednostjo p (najvišja kakovost). Uvrščanje časovnih vrst v te skupine je odvisno predvsem od števila meritev in dolžine merilnega obdobja, v manjši meri pa tudi od merjenega parametra, ki ga časovna vrsta spremlja.

2 Vhodni podatki

Vhodni podatki so pridobljeni iz baze državnega monitoringa kakovosti voda Agencije RS za okolje, ki jih dobimo za analizo v obliki Excel tabele. Vsaka vrstica v tabeli predstavlja meritev na določeni postaji v določenem času. Stolpci tabele vsebujejo podatke o šifri postaje, koordinatah postaje, datumu vzorčenja in podatke o količinah snovi v vodi za kisik, nitrate, sulfate, kloride ter električno prevodnost.

Nekateri podatki o količinah v tabeli niso v številski obliki, ampak so oblike npr. <1. Ta oznaka pomeni, da je količina pod mejo LOQ (ang. "limit of qualification"), torej jih v postopku meritve ne moremo numerično ovrednotiti. Vsem takšnim podatkom smo z namenom, da lahko z njimi računamo, priredili vrednost 0.

Naše podatke razporedimo v dveh nivojih. V prvem nivoju gre za podatke naštetih količin na izbranih postajah. Vsaka količina se na vsaki postaji meri večkrat; vse te meritve predstavljajo eno časovno vrsto. Iz razlogov, ki so pojasnjeni v naslednjem razdelku, za vsako tako časovno vrsto z linearno regresijo ocenimo trend. V procesu pridobimo statistike, ki opredeljujejo kakovost linearne regresije za modeliranje trenda vsake od časovnih vrst.

Naša populacija, ki je dejansko predmet raziskav, pa je populacija tako pridobljenih časovnih vrst. Katere podatke o njej zajamemo, je že predmet metodologije, ki je predstavljena v naslednjem razdelku.

3 Metodologija

Iz Excel datoteke prebrane in urejene (LOQ vrednosti so zamenjane z 0) spremenljivke standardiziramo, da imajo vse enako zalogo vrednosti na intervalu [0,1]. Pri standardizaciji vsako izmerjeno količino delimo z maksimalno izmerjeno vrednostjo te količine med vsemi vzorci vseh merilnih postaj.

Za vsako od skoraj 200 časovnih vrst zgradimo regresijski model z metodo najmanjših kvadratov, kjer je neodvisna spremenljivka število dni od prve meritve, odvisna spremenljivka pa je vrednost izmerjene količine. Sledi izračun parametrov kakovosti linearnega modela te časovne vrste. Za vsako od njih izračunamo korelacijski koeficient - R, signifikanco linearne regresije - p (to je signifikanca, s katero lahko zavrnemo hipotezo, da je pri obravnavani časovni vrsti regresijski koeficient neodvisne spremenljivke – časa – enak 0) in standardno napako modela - s.

Dodatno izračunamo vsoto kvadratnih odklonov – vsota oddaljenosti podatkov od regresijske premice in varianco – vsoto kvadratnih odklonov delimo s številom podatkov.

V projektu nas zanima tudi primerjava rezultatov, ki jih dobimo s surovimi in s povprečnimi vrednostmi. Surove vrednosti nam predstavljajo vsako posamezno meritev, torej eno vrstico v Excel datoteki. Povprečne vrednosti nam predstavljajo povprečje meritev posameznega leta. Zato za vsako leto posebej izračunamo povprečje vrednosti vsake količine v izbranem letu ter tako dobimo za vsako postajo samo eno vrstico podatkov letno. Za primerjavo dela s surovimi vrednostmi in letnimi povprečji se za vsako količino zgradita dva linearna modela za vsako postajo. Vse izračunane vrednosti zberemo v dveh ločenih razpredelnicah, v kateri vsak zapis (vrstica) predstavlja eno časovno vrsto; atributi tega zapisa so prej naštete statistike, ki smo jih o časovni vrsti zapisali (R, p, s, SSE, Var). Podatke projiciramo na nekaj parov dimenzij v razsevnih diagramih in intuitivno preverimo, ali se podatki izrazito združujejo v več skupin. Združevanje eksaktno preverimo z aglomerativnim razvrščanjem z uporabo evklidske metrike, katerega rezultat je dendrogram. Dendrogram je grafična predstavitev hierarhije skupin podatkov, ki na vsakem koraku poveže/združi dva podatka, najbližja po izbrani metriki. Tako ustvarjene

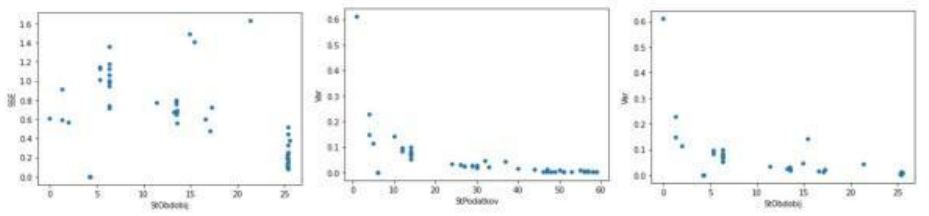
skupine podatkov shranimo v tabeli, kjer vsakemu podatku priredimo skupino, v kateri se nahaja.

Za dobljene skupine izdelamo tudi uvrščevalni model, ki z uporabo parametrov časovne vrste (število obdobij meritev, število podatkov meritev, merjena količina in število LOQ vrednosti) izdela odločitveno drevo, ki časovno vrsto uvrsti v eno od skupin kakovosti ocenjevanja. Med možnimi odločitvenimi drevesi izpišemo drevo z največjo natančnostjo, med enakovrednimi drevesi pa iščemo tistega s čim manjšo razvejanostjo. Pri izdelavi odločitvenega drevesa dopuščamo decimalna števila za število obdobij, saj je število obdobij definirano kot trajanje obdobja med prvo in zadnjo meritvijo na postaji, izraženo v letih. Metriki, ki se uporabljata za gradnjo odločitvenega drevesa, sta gini-indeks in entropija. Gini-indeks lahko interpretiramo kot pričakovano stopnjo napake. Najmanjša vrednost gini indeksa je 0, kar se zgodi, ko so vsi elementi istega razreda. Odločitveno drevo se na vsakem koraku razcepi, da se zmanjša gini-indeks vozlišča in to počne dokler je mogoče. Entropija meri, kako razpršeni so podatki v množici. Množica, kjer so vsi elementi enake skupine, ima vrednost entropije enako 0. Če je enako elementov dveh skupin, ima entropija vrednost 0,5. Podobno kot pri gini-indeksu se delitev zgodi, da se zmanjša entropija, dokler je to mogoče.

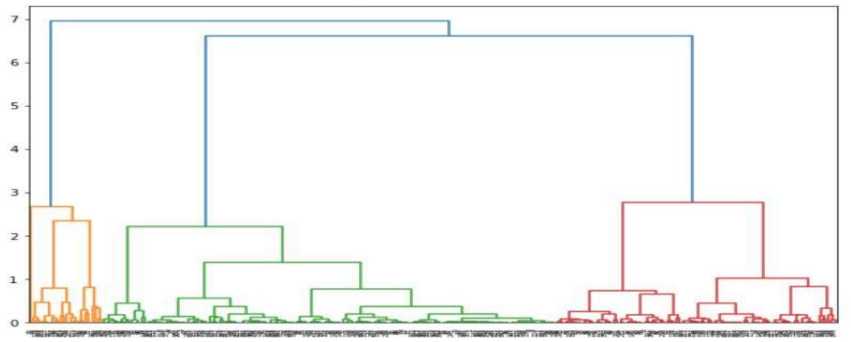
4 Analiza podatkov

4.1 Trendi iz surovih vrednosti

Najprej naredimo analizo parametrov kakovosti ocen trendov časovnih vrst, pridobljenih iz surovih podatkov. To pomeni, da v časovno vrsto zajamemo vse meritve neke količine na izbrani postaji. Najprej predstavimo razsevne diagrame, ki prikazujejo projekcije podatkov o časovnih vrstah na izbrane pare obravnavanih vplivnih faktorjev. Slika 2 prikazuje najprej razsevni diagram s številom obdobij in vsoto kvadratov napak, nato razsevni diagram s številom podatkov in nepojasnjeno varianco časovnih vrst in nazadnje razsevni diagram s številom obdobij in nepojasnjeno varianco časovnih vrst..

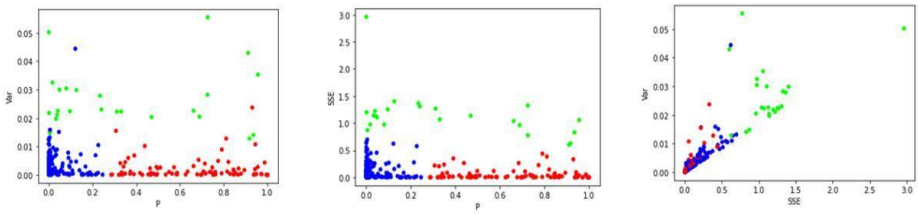


Slika 2: Razsevni diagram



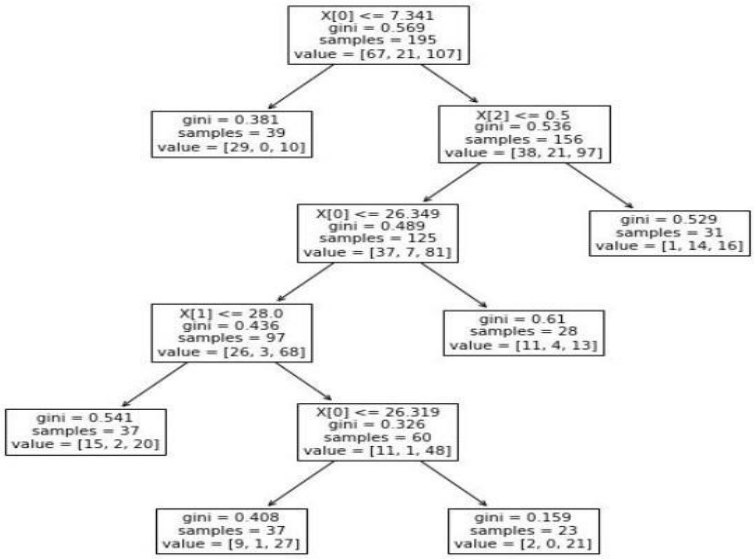
Slika 3: Dendrogram

Dendrogram na sliki 3 pokaže, da lahko vse časovne vrste razdelimo v tri bistveno različne skupine, od katerih vsaka vsebuje relativno bližnje časovne vrste. Na sliki 4 te tri skupine izrišemo na treh dodatnih razsevnih diagramih. Na vseh treh je vsaka skupina dendrograma obarvana z drugačno barvo: zelena barva predstavlja najslabšo skupino, za katero so značilna visoka odstopanja meritev od linearnega modela. V tej skupini najdemo časovne vrste s statistično značilnimi kot s statistično neznačilnimi trendi. Rdeča barva predstavlja srednjo skupino kakovosti trendov. V tej so odstopanja meritev od modela manjša, trend pa statistično značilno ni prisoten (p-vrednosti so visoke). Najboljša skupina je predstavljena z modro barvo. Ta vsebuje časovne vrste, katerih linearni modeli imajo nizka odstopanja meritev od modela, kot tudi koeficient modela statistično značilno različen od 0.



Slika 4: Razsevni diagram z upoštevanimi skupinami dendrograma

Razumevanje vpliva števila parametrov časovne vrste (število podatkov, število merilnih obdobij, merjena količina) na kakovost napovedanega trenda raziščemo z modelom uvrščanja. Na sliki 5 je odločitveno drevo, ki pojasni, kakšne časovne vrste se uvrstijo v katero od omenjenih treh z dendrogramom pridobljenih skupin.



Slika 5: Odločitveno drevo

Pri izdelavi dreves smo preverjali možnosti izdelave z izbiro odločitvenih atributov bodisi gini metriko, bodisi entropy metriko. Atributi, uporabljeni za izdelavo, so X[0] - število obdobij merjenja, X[1] - število meritev, X[2] - ali gre za meritev kisika (1) ali ne (0), X[3] - ali gre za meritev nitratov (1) ali ne (0), X[4] - ali gre za meritev sulfatov (1) ali ne (0), X[5] - ali gre za meritev kloridov (1) ali ne (0), X[6] - ali gre za

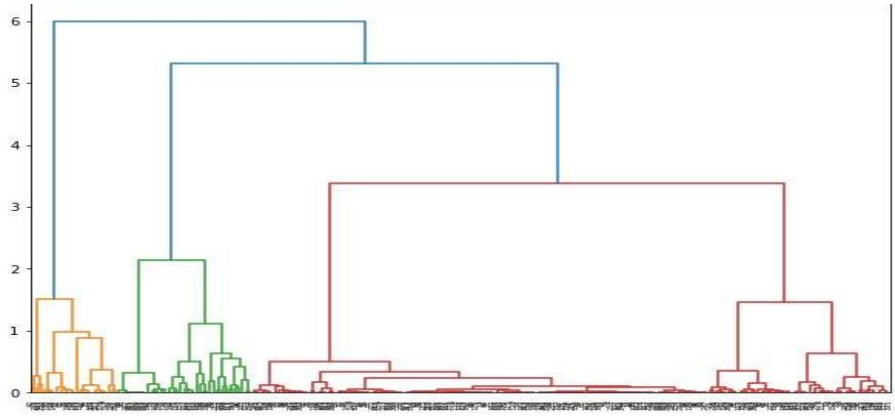
meritev električne prevodnosti (1) ali ne (0) in X[7] – število LOQ vrednosti. Med vsemi drevesi smo iskali tisto, ki je imelo največjo natančnost uvrščanja v skupine, med vsemi takimi pa najmanj razvejano.

Prikazano drevo temelji na gini metriki in za delitev upošteva število obdobij (X[0]), število podatkov (X[1]) ter ali gre za meritev kisika ali katere druge količine (X[2]). Natančnost drevesa je približno 69%. Listi drevesa predstavljajo šest skupin, ki skušajo biti karseda homogene glede na oznake, pridobljene z dendrogramom.

Prva delitev se zgodi glede na število obdobij. Če imamo manj ali enako kot 7,341 obdobij, ima časovna vrsta v vsakem primeru nizko nepojasnjeno varianco regresijskega modela in s 74% verjetnostjo spada v skupino z visoko vrednostjo p, s 26% verjetnostjo pa ima tudi nizko vrednost p in s tem statistično značilen trend. Če pa je število obdobij večje od 7.341, pride do naslednje delitve, ki kot atribut upošteva, ali gre za meritev kisika. Če je količina kisik, s 45% verjetnostjo gre za visoko nepojasnjeno varianco linearnega modela, s 3% verjetnostjo za nizko nepojasnjeno varianco in visoko vrednost p, z 52% verjetnostjo pa gre za nizko nepojasnjeno varianco in nizko vrednost p. Če podatek ni meritev kisika, se izvede naslednja delitev glede na število obdobij. Če je število obdobij večje od 26.349, ima časovna vrsta 14% verjetnostjo visoko nepojasnjeno varianco linearnega modela, s 39% verjetnostjo nizko nepojasnjeno varianco in visoko vrednost p, s 47% verjetnostjo pa nizko nepojasnjeno varianco in nizko vrednost p. Če je število meritev manjše ali enako 26.349, dobimo naslednjo delitev. Ta delitev upošteva število meritev. Če je število meritev manjše ali enako od 28, ima 5% časovnih vrst visoko nepojasnjeno varianco, 41% časovnih vrst nizko nepojasnjeno varianco in visoko vrednost p ter 54% časovnih vrst nizko nepojasnjeno varianco in nizko vrednost p. Če je število meritev večje od 28, pridemo do zadnje delitve. Ta se zgodi glede na število obdobij in sicer pri številu 26.319. Omeniti je potrebno, da imamo na tem koraku samo podatke, ki imajo število obdobij med 7.341 in 26.349. Torej, če je število obdobij znotraj tega obsega, za podatek število obdobij merjenja manjše ali enako 26.319 je 3% verjetnosti, da gre za visoko nepojasnjeno varianco, 24% verjetnosti za nizko nepojasnjeno varianco in visoko vrednost p ter 73% verjetnosti za nizko nepojasnjeno varianco in nizko vrednost p. Sicer pa gre v vsakem primeru za nizko nepojasnjeno varianco, pri čemer je z 91% verjetnostjo nizka tudi vrednost p, z 9% verjetnostjo pa je vrednost p visoka.

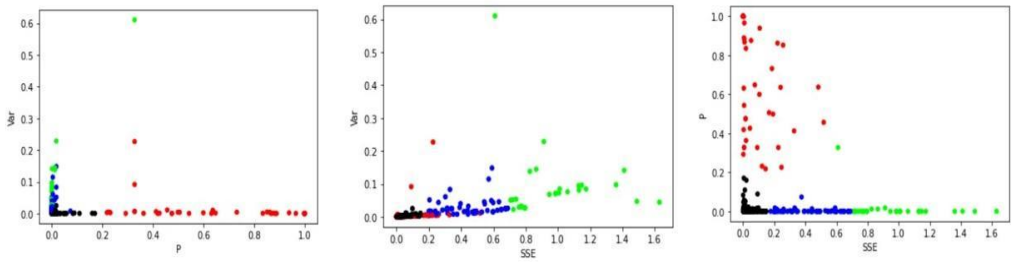
4.2 Trendi iz povprečnih letnih vrednosti

Poleg prej opisane analize, v katerih časovne vrste obsegajo vse meritve, lahko naredimo tudi analizo, v kateri podatke istega časovnega obdobja (leta) povprečimo in v časovni vrsti dobimo po en podatek za vsako obdobje, ki predstavlja povprečje meritev tega obdobja. Ker se pri analizi surovih vrednosti pokaže dokaj jasno razvidna delitev časovnih vrst v tri skupine glede na kakovost izračunanih linearnih trendov, predpostavljamo, da bo podobno tudi v primeru uporabe povprečnih letnih vrednosti. Pri preverjanju te predpostavke si spet pomagamo z aglomerativnim razvrščanjem novih časovnih vrst z uporabo istih atributov kot v prejšnjem primeru.



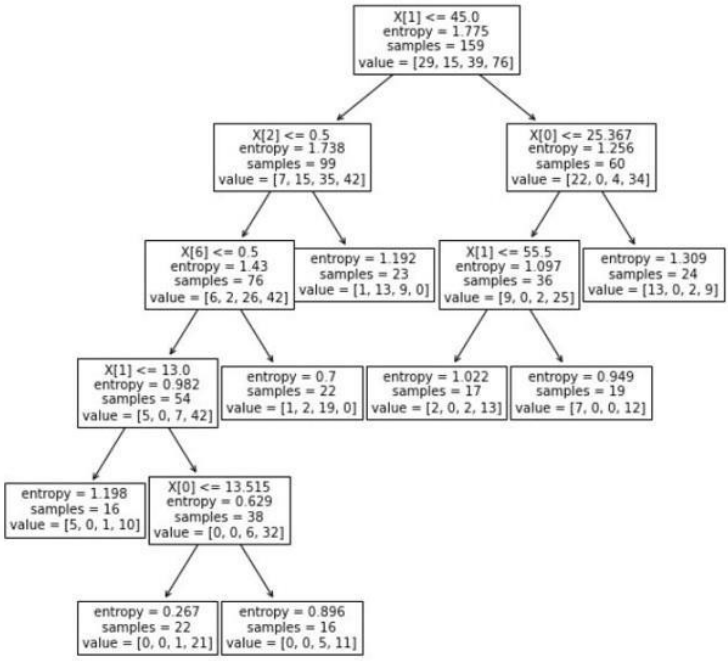
Slika 6: Dendrogram

Dendrogram na sliki 6 pokaže, da se časovne vrste delijo v štiri večje skupine - torej nekoliko drugače kot pri surovih podatkih. Na naslednjih slikah so prikazani razsevni diagrami, na katerih je vsaka skupina dendrograma obarvana z drugačno barvo: rdeča barva predstavlja skupino 0, za katero so značilne visoke p vrednosti, zelena barva predstavlja skupino 1 (nizke p vrednosti, visoka vsota kvadratov napak), modra barva predstavlja skupino 2 (nizke p vrednosti, srednja vsota kvadratov napak) in črna barva predstavlja skupino 3 (nizke p vrednosti, nizka vsota kvadratov napak, nizka nepojasnjena varianca).



Slika 7: Razsevni diagram z upoštevanimi skupinami dendrograma

Skupine, ki smo jih pridobili z aglomerativnim razvrščanjem, spet pojasnimo z modelom uvrščanja, odločitvenim drevesom, in pridobimo oceno verjetnosti, da z uporabo vnaprej znanih atributov časovne vrste (število podatkov, število merilnih obdobij, merjena količina), le-to uvrstimo v eno od štirih opisanih skupin.



Slika 8: Odločitveno drevo

Izbrano drevo temelji na metriki entropije in pri delitvi uporablja štiri različne kriterije: število obdobij (X[0]), število meritev (X[1]), podatek ali gre za meritev kisika ali ne (X[2]) in podatek, ali gre za meritev električne prevodnosti ali ne (X[6]). Natančnost drevesa je približno 80%. Drevo ima osem listov, torej dobimo osem kombinacij kriterijev, na podlagi katerih lahko z določeno verjetnostjo posamezne časovne vrste uvrstimo v eno od skupin dendrograma.

Drevo začne deliti časovne vrste s kriterijem število meritev.

Poglejmo najprej možnost, da je število meritev manjše ali enako 45. Naslednji kriterij, ki ga vzame drevo je, ali gre za meritev kisika ali ne. Če gre za meritev kisika, potem je taka časovna vrsta s 4% verjetnostjo v skupini z visokimi p-vrednostmi, s 57% verjetnostjo v skupini z nizkimi p-vrednostmi in visokimi vsotami kvadratov napak in z 39% verjetnostjo v skupini z nizkimi p-vrednostmi in srednjimi vsotami kvadratov napak. Sicer drevo naredi naslednjo delitev glede na kriterij, ali gre za meritev električne prevodnosti ali ne. Če gre za meritev električne prevodnosti, potem lahko s 5% verjetnostjo rečemo, da časovna vrsta spada v skupino z visokimi p vrednostmi, z 9% verjetnostjo, da spada v skupino z nizkimi p vrednostmi in visokimi vsotami kvadratov napak, in s 86% verjetnostjo, da je časovna vrsta v skupini z nizkimi p vrednostmi in srednjimi vsotami kvadratov napak. Če ne gre za meritev električne prevodnosti, pridemo do naslednje delitve drevesa, ki upošteva število meritev. Če je število meritev manjše ali enako 13, potem časovna vrsta z verjetnostjo 31% spada v skupino z visokimi p vrednostmi, s 6% verjetnostjo v skupino z nizkimi p vrednostmi in srednjimi vsotami kvadratov napak in s 63% verjetnostjo v skupino z nizkimi p-vrednostmi, nizkimi vsotami kvadratov napak in nizkimi nepojasnjenimi variancami . Če je število meritev večje od 13, pridemo do zadnje delitve v tej veji drevesa, kjer imajo vse skupine nizke p vrednosti. Ta delitev se zgodi glede na število obdobij. Če je število obdobij manjše ali enako 13.515, dobimo samo 2 skupini: 5% je verjetnosti, da časovna vrsta spada v skupino s srednjo vsoto kvadratov napak in 95% verjetnosti, da časovna vrsta spada v skupino z nizko vsoto kvadratov napak in nizko nepojasnjeno varianco. Tudi pri številu obdobij večjem od 13.515 dobimo le ti dve skupini: z 31% verjetnostjo lahko rečemo, da časovna vrsta spada v skupino s srednjo vsoto kvadratov napak in z 69% verjetnostjo, da je časovna vrsta v skupini z nizko vrednostjo kvadratov napak in nizko nepojasnjeno varianco.

Preveriti moramo še, kaj se zgodi, če je število meritev večje od 45. Naslednja delitev v drevesu je glede na kriterij števila obdobij. Če je število obdobij večje od 25.367, potem je časovna s 54% verjetnostjo v skupini z visokimi p vrednostmi, z 8% verjetnostjo v skupini z nizkimi p vrednostmi in srednjimi vsotami kvadratov napak in z 38% verjetnostjo v skupini z nizkimi p vrednostmi, nizkimi vsotami kvadratov napak in nizko nepojasnjeno varianco. Sicer se zgodi delitev glede na število meritev. Če je število meritev manjše ali enako 55.5 (in hkrati večje od 45 zaradi zgornjega pogoja), potem je časovna vrsta v skupini z visoko p vrednostjo z verjetnostjo 12%, nizko p vrednostjo in srednjo vsoto kvadratov napak z verjetnostjo 12% in z nizkimi vsemi tremi količinami z verjetnostjo 76%. Če pa je število meritev večje od 55.5, je časovna vrsta z verjetnostjo 39% v skupini z visoko p vrednostjo in z verjetnostjo 61% pa ima nizke vse tri omenjene količine.

4.3 Primerjava surovih in povprečnih vrednosti

4.3.1 Upoštevani kriteriji

Glavna kriterija, ki ju upošteva odločitveno drevo v obeh primerih sta število obdobij merjenja in število meritev. Opazimo, da se drevo pri surovih vrednostih največkrat posluži kriterija število obdobij merjenja, drevo pri povprečnih vrednostih pa največkrat upošteva kriterij število meritev.

Dodaten kriterij, ki ga upoštevata obe drevesi, je, ali gre za meritev kisika ali ne. Drevo pri povprečnih vrednostih upošteva še kriterij, ali gre za meritev električne prevodnosti. Ostali kriteriji v večini primerov niso upoštevani. Naša domneva je, da do tega pride zaradi razlik v vrednostih meritev. Pri kisiku se merjene vrednosti gibljejo med 0.8 in 12.3 miligramov kisika na liter vode. Pri nitratih, sulfatih in kloridih se vrednosti meritev gibljejo vse od 1 pa do 100 miligramov na liter, torej so tukaj razlike veliko večje in za drevo ti kriteriji niso relevantni. Električna prevodnost ima še nekoliko drugačno skalo in sicer so njene vrednosti med 288 in 1110 mikro Siemens na centimeter. Ta skala za surove vrednosti ni preveč ugodna, saj so razlike še vedno zelo velike, pri povprečnih vrednostih pa se te razlike po izračunu povprečij zmanjšajo in kriterij postane pomemben za odločitveno drevo.

Pri primerjavi velja omeniti tudi, da je odločitveno drevo, pridobljeno nad podatki s povprečnimi letnimi vrednostmi, nekoliko natančnejše (80%) v primerjavi z drevesom uvrščanja časovnih vrst nad surovimi podatki (69%).

4.3.2 Število meritev

Pri odločitvenem drevesu glede na povprečne vrednosti opazimo, da se veliko postaj z velikim številom meritev uvršča v skupino, za katero so značilna visoka odstopanja meritev od linearnega modela (visoka vrednost p in SSE). Podobno se zgodi tudi pri drevesu glede na surove vrednosti. Razlog za to vidimo v dejstvu, da zaradi večjega števila meritev lahko pride do posameznih izstopajočih podatkov, ki ne sledijo trendu zaradi več možnih razlogov, npr. letni čas meritve, vremenske razmere pred meritvijo itd. Pri nadaljnji analizi bi bilo smiselno izključiti te izstopajoče podatke in opazovati, kako se spremeni položaj meritev glede na skupine dendrograma.

4.3.3 Število obdobij merjenja

Kriterij število obdobij merjenja kaže na to, da se z večanjem obdobij merjenja podatki izboljšujejo, torej trendi dobijo manjšo p-vrednost in manjšo napako ali nepojasnjeno varianco. Se pa tudi tukaj v listih odločitvenega drevesa (predvsem pri drevesu surovih vrednosti) pri večjem številu merilnih obdobij pojavijo tudi slabši podatki. Do tega pride, ker se z večanjem števila obdobij poveča tudi število meritev in pride do večje možnosti odstopanja.

5 Zaključek

V prispevku predstavimo metodo, s katero ovrednotimo kakovost izračunanih trendov za časovne vrste kemijskih parametrov površinskih voda. Trende ocenimo z linearno regresijo in s tem dobimo opisne statistike posameznih časovnih vrst. Vektorji petih statistik časovnih vrst (korelacijski koeficient, p-vrednost regresijskega koeficienta, standardna napaka regresije, vsota kvadratov napak linearnega modela, nepojasnjena varianca linearnega modela) nastopajo kot vhodni podatek v aglomerativno gručenje časovnih vrst. Če smo trende računali na vseh razpoložljivih meritvah, pridobimo tri kakovostno različne skupine (visoka nepojasnjena varianca, nizka nepojasnjena varianca in hkrati visoka p vrednost ter nizka nepojasnjena

varianca in hkrati nizka p vrednost). Če smo trende računali na podatkih, v katerih smo meritve znotraj istega leta povprečili, pa dobimo štiri kakovostno različne skupine (visoka p vrednost, nizka p vrednost in visoka vsota kvadratov napak, nizka p vrednost in srednja vsota kvadratov napak ter nizka p vrednost in hkrati nizka vsota kvadratov napak in tudi nizka nepojasnjena varianca).

Za te skupine izdelamo tudi uvrstitveni model, s katerim uvrščanje v skupine pojasnimo z atributi, ki pripadajo le časovni vrsti in jih lahko opazimo brez računanja modela. Ta uvrstitveni model potrdi, da je višjo stopnjo kakovosti linearnih trendov mogoče zaznati na postajah, kjer se podatki pogosteje in dlje časa beležijo. Obstajajo pa tudi primeri, kjer so trendi meritev prav zaradi velikega števila podatkov slabši, saj se med veliko podatki pojavijo visoka odstopanja.

Če gledamo na kemijske parametre, lahko bolj zanesljivo napovemo trende kisika kot trende ostalih parametrov. Pri kisiku ne prihaja do velikih odstopanj od povprečne meritve, pri nitratih, sulfatih, kloridih in električni prevodnosti pa so ta odstopanja večja.

Predvidevamo, da bi lahko imelo število LOQ vrednosti za določeno postajo velik vpliv na zanesljivost trenda. Število LOQ vrednosti smo vključili v gradnjo uvrstitvenega modela, a v izdelanem drevesu ni uporabljeno. Menimo, da do takšnega rezultata pride zaradi relativno majhnega števila časovnih vrst, ki sploh vsebujejo LOQ vrednosti. Le-teh je samo 19 med našimi 265 podatki.

Vidimo, da je odprtih veliko možnosti za nadaljnje raziskave. Zanimivo bi bilo odstraniti meritve, ki zelo odstopajo od povprečja in preveriti, ali lahko v tem primeru trend bolj zanesljivo napovemo. Prav tako bi bilo dobro naše algoritme preveriti še na kakšnih drugih podatkih, ki bi lahko dali drugačne rezultate (npr. če bi bilo več LOQ vrednosti, bi lahko število le-teh znatno vplivalo na zanesljivost trenda). Morda bi bilo poleg napovedovanja z linearno regresijo smiselno uporabiti še kak kompleksnejši pristop računanja trendov časovnih vrst.

Literatura

ARSO. (5. januar 2022). Pridobljeno iz Vode - poročila in publikacije: https://www.arso.gov.si/vode/poro%C4%8Dila%20in%20publikacije/kakovost%20voda/Kakovost%20voda-SLO.pdf

Data Analytics. (10. december 2021). Pridobljeno iz https://vitalflux.com/visualize-decision-tree-python-sklearn-library/

DataCamp. (10. december 2021). Pridobljeno iz https://www.datacamp.com/community/tutorials/decision-tree-classification-python

Hasan, M. I. (11. november 2021). Mohammad Imran Hasan. Pridobljeno iz https://mohammadimranhasan.com/linear-regression-of-time-series-data-with-pandas-library-in-python/

Oman, Š. (7. januar 2022). BenSTAT. Pridobljeno iz https://www.benstat.si/blog/pearsonov-koeficient-korelacije/

QuantDare. (7. januar 2022). Pridobljeno iz https://quantdare.com/decision-trees-gini-vs-entropy/

SciKit Learn. (10. december 2021). Pridobljeno iz https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

SciKit Learn. (10. december 2021). Pridobljeno iz https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html

SciKit Learn. (10. december 2021). Pridobljeno iz https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

SciPy. (10. december 2021). Pridobljeno iz https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html

Statistik.si. (7. januar 2022). Pridobljeno iz https://www.statistik.si/p-vrednost/

Wikipedia. (10. december 2021). Pridobljeno iz https://en.wikipedia.org/wiki/Dendrogram

Editor, M. B. (2013). *How to Interpret Regression Analysis Results: P-values and Coefficients*. Pridobljeno iz Minitab: https://blog.minitab.com/en/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients

Enotna zbirka podatkov monitoringa kakovosti voda, Agencija Republike Slovenije za okolje, 2021

Republika Slovenija. (2002). *Zakon o vodah (ZV-1)*. Pridobljeno iz PisRS: http://pisrs.si/Pis.web/pregledPredpisa?id=ZAKO1244

Republika Slovenije. (2004). *Zakon o varstvu okolja (ZVO-1)*. Pridobljeno iz PisRS: http://pisrs.si/Pis.web/pregledPredpisa?id=ZAKO1545

Republika Slovenije. (2009). *Pravilnik o monitoringu podzemnih voda*. Pridobljeno iz PisRS: http://pisrs.si/Pis.web/pregledPredpisa?id=PRAV9521

Republika Slovenije. (2009). *Pravilnik o monitoringu stanja površinskih voda*. Pridobljeno iz PisRS: http://pisrs.si/Pis.web/pregledPredpisa?id=PRAV9315

Republika Slovenije. (2009). *Uredba o stanju podzemnih voda*. Pridobljeno iz PisRS: http://www.pisrs.si/Pis.web/pregledPredpisa?id=URED5121

Republika Slovenije. (2009). *Uredba o stanju površinskih voda*. Pridobljeno iz PisRS: http://www.pisrs.si/Pis.web/pregledPredpisa?id=URED5010