Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Miha Zidar

Dostop do podatkov Svetovne banke v orodju Orange

diplomsko delo

univerzitetni študijski program računalništvo in informatika

Mentor: prof. dr. Blaž Zupan

Ljubljana, 2016

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Na spletu je veliko odprtih baz podatkov, za katere bi bilo koristno, da bi bile dostopne v orodjih za podatkovno analitiko. Primer take baze so spletne strani in programski vmesniki Svetovne banke, preko katerih lahko dostopamo do demografskih, gospodarskih in klimatskih podatkov. V nalogi razvijte knjižnico in komponente z grafičnimi vmesnikom za program Orange, s katerimi lahko enostavno in hitro pridobimo podatke iz Svetovne banke in jih z obstoječimi analitičnimi gradniki v Orange-u tudi analiziramo.

Zahvalil bi se mentorju, prof. dr. Blažu Zupanu in članom laboratorija za bioinformatiko za pomoč in usmerjanje med izdelavo diplomskega dela. Prav tako bi se zahvalil svojemu partnerju, staršem in prijateljem za spodbudo.

Kazalo

Povzetek

Naslov: Dostop do podatkov Svetovne banke v orodju Orange

Avtor: Miha Zidar

Program Orange je orodje za podatkovno rudarjenje, v katerem lahko za namene analiz uporabimo različne podatkovne vire. Sam program Orange vsebuje predpripravljene zbirke podatkov, dodatne zbirke podatkov si lahko pripravi in uvozi tudi uporabnik sam, ali pa uporabi katerega od že obstoječih dodatkov za uvoz podatkov. Za namen diplomske naloge smo izdelali dodatek imenovan Orange Data Sets, ter v njem razvili gradnike za dostop do podatkov s programskega vmesnika Svetovne banke. Svetovna banka omogoča uporabo štirih različnih programskih vmesnikov: gospodarski indikatorji, finančni podatki, projekti Svetovne banke in podnebni podatki. V dodateku Orange Data Sets smo razvili gradnike za branje in uporabo podatkov indikatorjev in podnebnih podatkov. S tem bo uporabnikom programa Orange omogočena enostavnejša uporaba velikega števila podatkov iz omenjenih dveh programskih vmesnikov.

Ključne besede: podatkovno rudarjenje, programski vmesnik, Svetovna banka, gospodarski indikatorji, podnebni podatki, Orange.

Abstract

Title: Access to World Bank Data with Orange

Author: Miha Zidar

Orange is an open source data-mining software, capable of using multiple sources for data analysis. There are a few test data samples already present in Orange, and the user can import their own data sets with the use of one of Orange input widgets. For this thesis, we created an add-on Orange Data Sets, that includes widgets for accessing free data from World Bank application program interface (API). The World Bank exposes four different data APIs; indicator, project, finance and climate. Widgets included in the Orange Data Sets add-on are be able to read data from the World Bank Indicators and World Bank Climate APIs. This will enable Orange users a quick and easy access to data from the two previously mentioned APIs.

Key words: Data mining, API, World Bank, indicators, climate, Orange.

Poglavje 1

Uvod

Na svetovnem spletu je dosegljivih vedno več prosto dostopnih programskih vmesnikov (angl. *application programming interface*). Ti vmesniki omogočajo dostop do zelo raznolikih zbirk podatkov. Primeri takih zbirk so seznam stopnje ogroženosti živali po državah[1], podatki meritev in slike vesolja agencije NASA[2], seznam knjig z ocenami in povezavami med uporabniki[3], zgodovina meteoroloških meritev[4], razni indikatorji stopenj razvoja držav[5].

Programski vmesniki so oblikovani tako, da je omogočena raznolika uporaba vmesnika, s katerim določimo, katere podatke želimo pridobiti. Ta fleksibilnost pa ima tudi slabost, saj je podatke potrebno predhodno obdelati za vsak namen posebej. Tako bi na primer moral vsak uporabnik programa Orange, sicer splošno uporabnega okolja za podatkovno analitiko, podatke predhodno pretvoriti v obliko, primerno za dani problem in cilje zastavljene analize.

[1] `http://apiv3.iucnredlist.org/api/v3/docs`
[2] `https://api.nasa.gov/`
[3] `https://www.goodreads.com/api`
[4] `http://climatedataapi.worldbank.org/`
[5] `http://api.worldbank.org/`

1.1 Motivacija

Povezava programskega vmesnika za dostop do podatkov in orodja za analizo podatkov je pogosto prezapletena za končnega uporabnika. Razviti želimo knjižnice in dodatka za program Orange, s katerimi bi podatke s programskega vmesnika Svetovne banke pripravili v obliki primerni za nadaljnjo uporabo v orodju Orange in drugih programih za obdelavo podatkov. S tem bi dobili enostavnejši dostop do preko 16.000 indikatorjev in številnih podnebnih meritev, s čimer bomo lažje analizirali in iskali morebitne zakonitosti v podatkih. Če bi imeli en sam ustrezen dodatek za dostop do podatkov programskega vmesnika Svetovne banke, bi se poenostavilo tudi posodabljanje in vzdrževanje kode v primeru sprememb programskega vmesnika. S tem odpravimo potrebo, da bi moral vsak uporabnik sam skrbeti za uskladitvene posodobitve, ampak se vmesnik posodobi enkrat in za vse uporabnike.

1.2 Cilji in struktura diplomske naloge

Cilj diplomske naloge je izdelati knjižnico za uporabo programskega vmesnika Svetovne banke za programski jezik Python ter izdelati dodatek za program Orange, ki s pomočjo omenjene knjižnice omogoča uporabniku dostop do podatkov Svetovne banke preko grafičnega vmesnika.

V diplomski nalogi najprej predstavimo spletna vira indikatorjev držav sveta in meritev podnebnih podatkov Svetovne banke ter opišemo delovanje njunih programskih vmesnikov. Nato podrobneje opišemo našo implementacijo knjižnice za dostop do programskega vmesnika Svetovne banke in gradnikov za program Orange, ki to knjižnico uporabljajo. V nadaljevanju prikažemo še nekaj praktičnih primerov uporabe razvitih gradnikov. Na koncu še popišemo opravljeno delo, navedemo vire programske kode in omenimo možne načine za izboljšavo ali nadgradnjo našega dodatka.

Poglavje 2

Podatkovne zbirke Svetovne Banke

Pri diplomski nalogi smo se osredotočili na dva programska vmesnika za dostop podatkov Svetovne banke. To sta "ClimateAPI", s katerim dostopamo do podatkovne zbirke meteoroloških meritev in "IndicatorAPI", s katerim dostopamo do zbirke podatkov raznih indikatorjev stopenj razvoja držav. Za uporabo podatkovne zbirke Svetovne banke smo se odločili, ker združuje in na enovit način predstavi podatke iz več različnih virov. Podatkovni viri za indikatorje stopnje razvoja držav so:

- Svetovni indikatorji razvoja [1],
- Globalni finančni razvoj [2],
- Afriški indikatorji razvoja [3],
- Poslovanje [4],
- Podjetniške raziskave [5],
- Razvojni cilji [6],
- Statistike izobraževanja [7],
- Statistike spolov [8],

- Statistike zdravja in prehranjevanja [9] in
- Rezultati meritev IDA [10].

Podatkovni vir zbirke podnebnih meritev pa je osnovan na podatkih oddelka za podnebne raziskave (angl. *Climatic Research Unit*) [11].

Svetovna banka omogoča dostop do podatkov preko programskega vmesnika predstavitvene arhitekture za prenos podatkov REST (angl. *Representational State Transfer*), ki ponuja veliko možnosti za iskanje in izbor rezultatov programskih poizvedb. Pri vsaki poizvedbi REST lahko določimo želeno obliko odgovora. Za poizvedbe o informacijah indikatorjev sta na voljo obliki razširljivega označevalnenga jezika XML (angl. *Extensible Markup Language*) in javascript objektne notacije JSON (angl. *JavaScript Object Notation*). Programski vmesnik meteoroloških meritev pa ponuja samo obliko JSON. Za konsistentnost in lažjo berljivost smo na obeh programskih vmesnikih uporabili obliko JSON. To na programskem vmesniku indikatorjev dosežemo tako da nastavimo parameter GET `format` na vrednost `json`.

2.1 Podatki indikatorjev razvoja držav

Programski vmesnik indikatorjev razvoja držav Svetovne banke omogoča dostop do podatkov preko 16.000 raznih indikatorjev. Podatki indikatorjev so merjeni v mesečnem, četrtletnem ali letnem intervalu. Začetek meritev podatkov posameznega indikatorja je odvisna od vira podatkov. Najstarejši podatki segajo do leta 1960. Poleg podatkov indikatorjev nam ta programski vmesnik omogoča tudi dostop do večine metapodatkov, s katerimi lahko presejamo in natančneje določimo našo poizvedbo in ki vključujejo::

- viri podatkov in njihovi opisi (angl. *Catalog Source Queries*[1]),
- seznam držav, skupin držav in regij z identifikatorji (angl. *Country Queries*[2]),

[1] `http://api.worldbank.org/sources?format=json`
[2] `http://api.worldbank.org/countries?format=json`

- razdelitev višin dohodkov z identifikatorji (angl. *Income Level Queries*[3]),
- seznam indikatorjev (angl. *Indicator Queries*[4]),
- seznam tipov posojil (angl. *Lending Type Queries*[5]),
- seznam tem (angl. *Topics*[6]).

Za pridobitev podatkov indikatorjev potrebujemo metapodatke o indikatorjih in državah. Primere teh metapodatkov si bomo podrobneje pogledali v nadaljevanju.

Ker je mogoče z eno poizvedbo dostopati do velike količine podatkov, ima programski vmesnik za dostop do podatkov indikatorjev implementirano paginacijo, s katero je omejeno število podatkov, ki jih lahko dobimo z eno poizvedbo. Tako so podatki razdeljeni na skupine, ki jih imenujemo strani.

Vsi odgovori na veljavne poizvedbe po podatkih in metapodatkih, ki so na voljo s programskim vmesnikom indikatorjev razvoja, imajo enako osnovno obliko. Poizvedbe vračajo seznam z dvema elementoma, kjer ima prvi element informacije o količini podatkov in trenutnem izboru podatkov, drugi element pa vsebuje seznam izbranih podatkov (primer 1). Privzeta vrednost števila elementov na stran je 50, kar lahko spremenimo tako, da poizvedbi nastavimo parameter GET `per_page` na poljubno vrednost. Če želimo pridobiti podatke z več strani, moramo za vsako stran poslati novo poizvedbo, v kateri podamo številko želene strani s parametrom GET `page`. Veljavne poizvedbe s sitom, ki ne vrača nobenih podatkov, imajo vrednost drugega elementa osnovnega seznama `null`. Za neveljavne poizvedbe pa programski vmesnik vrača seznam z enim elementom, ki vsebuje podatke o napaki poizvedbe (primer 2).

[3] `http://api.worldbank.org/incomeLevels?format=json`
[4] `http://api.worldbank.org/indicators?format=json`
[5] `http://api.worldbank.org/lendingTypes?format=json`
[6] `http://api.worldbank.org/topics`

```
[
    {
        'page': 1,
        'pages': 137,
        'per_page': '50',
        'total': 6831
    },
    [
        <podatki>,
        ...
    ]
]
```

Primer 1: Osnovna oblika odgovora programskega vmesnika Svetovne banke za veljavno poizvedbo indikatorjev.

```
[
    {
        'message': [
            {
                'id': '120',
                'key': 'Parameter \'country\' has an invalid value',
                'value': 'The provided parameter value is not valid'
            }
        ]
    }
]
```

Primer 2: Osnovna oblika odgovora programskega vmesnika Svetovne banke za neveljavne poizvedbe.

2.1.1 Opis seznama indikatorjev

Programski vmesnik Svetovne banke za indikatorje razvoja nam ponuja seznam vseh indikatorjev z imeni, opisi, kodami in drugimi metapodatki (primer 4). Programski vmesnik nam omogoča tudi dostop do podatkov posameznega indikatorja določenega s kodo in presejanje seznama indikatorjev glede na vir podatkov 3. V našem programu smo uporabili le poizvedbo za celoten seznam indikatorjev, da smo omogočili iskanje in presejanje po vseh poljih indikatorjev.

```
1  http://api.worldbank.org/indicators?format=json
2  http://api.worldbank.org/indicators?format=json&source=5
3  http://api.worldbank.org/indicators/A10i?format=json
```

Primer 3: Primeri poizvedb po seznamu indikatorjev. 1) seznam vseh indikatorjev, 2) seznam indikatorjev glede na vir podatkov, 3) podatki indikatorja "A10i"

2.1.2 Opis seznama držav

Seznam držav na programskem vmesniku Svetovne banke vsebuje podatke o imenih, opisih, kodah držav po standardu ISO-3166-1, regijah in druge metapodatke (primer 6). Programski vmesnik nam omogoča tudi presejanje seznama držav po kodi države, regiji, višini dohodka in tipu posojil (primer 5)

Ta seznam ne vsebuje zgolj držav, ampak tudi regije in skupine držav, združenih glede na različne kriterije, kot so višina dohodka, velikost, stopnja razvoja. Poleg tega zgornji seznam vsebuje tudi nekatere izjeme, kot je trenutno Kosovo. V nadaljevanju bomo za vse naštete tipe lokacijskih podatkov uporabljali besedo "države".

```
{
    'id': '1.0.HCount.2.5usd',
    'name': 'Poverty Headcount (\$2.50 a day)',
    'source': {
        'id': '37',
        'value': 'LAC Equity Lab'
    },
    'sourceNote': 'The poverty headcount index measures the
                   proportion of the population with daily per
                   capita income (in  2005 PPP) below the poverty
                   line.',
    'sourceOrganization': 'LAC Equity Lab tabulations of SEDLAC
                           (CEDLAS and the World Bank).',
    'topics': [
        {
            'id': '11',
            'value': 'Poverty '
        }
    ]
}
```

Primer 4: Podatki indikatorja stopnja revščine pri dohodku 2,5 dolarja na dan.

```
http://api.worldbank.org/countries?format=json
http://api.worldbank.org/countries/svn?format=json
http://api.worldbank.org/countries?format=json&incomeLevel=HIC&region↩
    =ECS
```

Primer 5: Primeri poizvedb po seznamu držav. 1) seznam vseh držav, 2) podatki ene države, 3) seznam držav v Evropi in Osrednji Aziji z visoko višino dohodka.

```
{
    'id': 'ABW',
    'iso2Code': 'AW',
    'name': 'Aruba',
    'region': {
        'id': 'LCN',
        'value': 'Latin America & Caribbean '
    },
    'adminregion': {
        'id': '',
        'value': ''
    },
    'incomeLevel': {
        'id': 'HIC',
        'value': 'High income'
    },
    'lendingType': {
        'id': 'LNX',
        'value': 'Not classified'
    },
    'capitalCity': 'Oranjestad',
    'longitude': '-70.0167',
    'latitude': '12.5167'
},
```

Primer 6: Izsek podatkov veljavne poizvedbe držav.

2.1.3 Dostop do podatkov indikatorjev

Za dostop do podatkov posameznega indikatorja potrebujemo kodo indikatorja s seznama vseh indikatorjev in kodo ene ali več držav. Namesto kode ene ali več držav, lahko uporabimo tudi ključno besedo `all`, ki označuje vse kode držav. Pri večjih količinah podatkov lahko z dodatnimi parametri določimo število podatkov na stran in želeno stran podatkov. Primer 7 prikazuje osnovno obliko poizvedbe, kjer so:

`country` s podpičjem ločen seznam kod izbranih držav, ki jih preberemo iz polja `id` ali `iso2Code`, ki sta prikazana v primeru 6, ali pa ključna beseda `all`,

`indicator_id` polje `id` indikatorja ki je prikazano v primeru 4,

`parametri` Dodatni parametri GET

Za poizvedbe do podatkov indikatorjev so poleg osnovnih parametrov GET `per_page`, `page` in `format`, opisanih v poglavju 2.1, na voljo tudi dodatni parametri za presejanje rezultatov poizvedbe:

`mrv` Številska vrednost, ki določi maksimalno število zadnjih meritev, ki jih programski vmesnik vrne. Ko uporabljamo polje `mrv`, bo programski vmesnik izpustil ničelne vrednosti za obdobja v katerih ni meritev.

`gapfill` Zastavica `y` ali `n` za manjkajoče vrednosti meritev. Vrednost `y` v kombinaciji s poljem `mrv` poskrbi, da programski vmesnik ne izpusti nobenega časovnega intervala.

`date` Polje oblike `'leto'` ali `'leto:leto'` ki omeji rezultate poizvedbe na določeno leto ali interval med določenimi leti.

Privzeta vrednost za količino podatkov na stran `per_page` je 50. Zgornja meja pa ni strogo določena, vendar je odvisna od velikosti odgovora. Ugotovili smo, da se zanesljivost programskega vmesnika manjša z večjo količino podatkov na stran. V našem programu smo se omejili na 1000 podatkov na

```
http://api.worldbank.org/en/countries/<country>/indicators/<↩
    indicator_id>?<parametri>
```

Primer 7: Osnovna oblika poizvedbe za podatke enega indikatorja.

stran, kar se je izkazalo za uporabno razmerje med hitrostjo in zanesljivostjo programskega vmesnika. Privzeto bo programski vmesnik vrnil podatke za vse časovne vrednosti. V odgovoru programskega vmesnika dobimo seznam objektov (primer 8) z datumom, indikatorjem, državo in vrednostjo.

```
{
    'indicator': {
        'id': 'SP.POP.TOTL',
        'value': 'Population, total'
    },
    'country': {
        'id': 'IL',
        'value': 'Israel'
    },
    'value': '6289000',
    'decimal': '0',
    'date': '2000'
}
```

Primer 8: Podatki za indikator SP.POP.TOTL (skupno število prebivalcev države) za Izrael leta 2000.

Slabosti programskega vmesnika indikatorjev Svetovne banke za uporabo v namene podatkovnega rudarjenja so v tem, da vmesnik ni namenjen prenosu večje količine podatkov z eno samo poizvedbo. Zaradi paginacije moramo za en sam indikator narediti več poizvedb, da prenesemo podatke z vseh strani. Prav tako podatkovni vmesnik ne podpira poizvedb po več indikatorjih hkrati, kar potrebujemo za iskanje zakonitosti med posameznimi indikatorji.

2.2 Podatki podnebnih meritev

Programski vmesnik Svetovne banke za podnebne podatke omogoča dostop do podatkov napovednih modelov in zgodovinskih meritev meteoroloških postaj. V tej diplomski nalogi smo se odločili uporabiti samo podatke zgodovinskih meritev, saj si s temi podatki lahko uporabnik programa Orange sam sestavi svoje napovedne modele.

Za razliko od uporabe programskega vmesnika indikatorjev, lahko pri tem programskem vmesniku uporabljamo veljavne kode držav ISO 3166-1 alpha-2 ali ISO 3166-1 alpha-3, ali pa številski identifikator vodotočnega območja.

Ta programski vmesnik nam omogoča dostop do podatkov o povprečnih temperaturah in padavinah v časovnih obdobjih enega leta, desetletja ali pa nam omogoča dostop do mesečnih povprečij skozi vsa leta meritev.

2.2.1 Dostop do podatkov podnebnih meritev

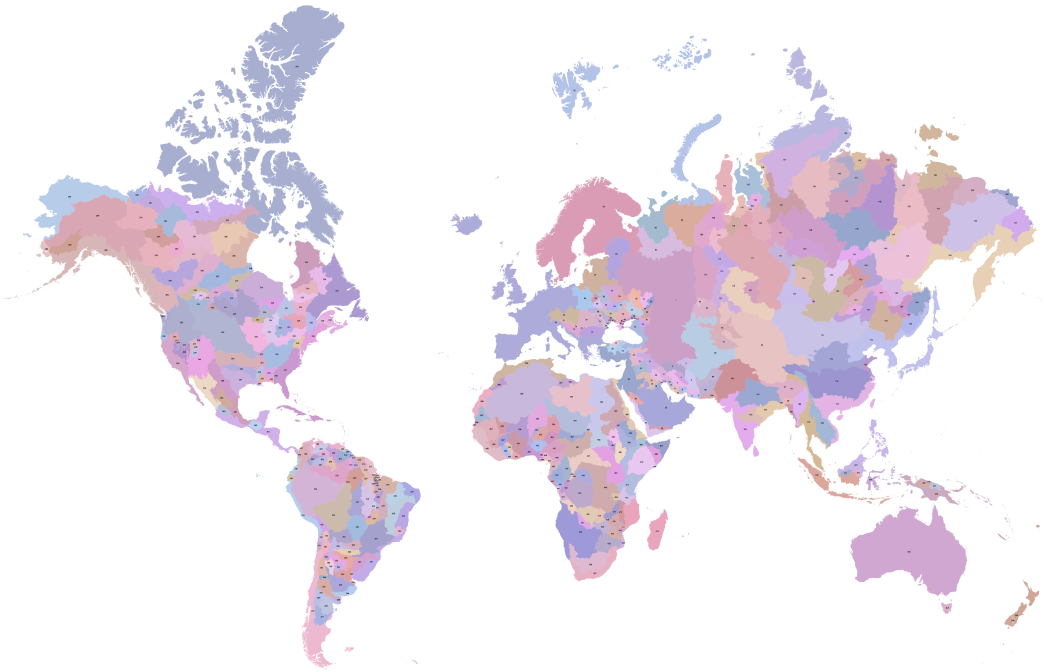
Za dostop do podnebnih podatkov preko programskega vmesnika Svetovne banke potrebujemo kodo države ISO-3166-1 alpha-3 ali številski identifikator vodotočnega območja (slika 2.1). Programski vmesnik nam omogoča dostop do meritev povprečnih količin padavin in temperatur za letno ali desetletno obdobje. Poleg letnega in desetletnega obdobja nam programski vmesnik ponuja tudi povprečno količino padavin in temperatur za posamezne mesece skozi vsa leta meritev. Obliko poizvedbe prikazuje primer 9, kjer je:

`loc_type` vrsta identifikatorja območja (“basin” za vodotočno območje, “country” za države),

`data_type` vrsta meritev (“pr” za padavine, “tas” za temperature),

`interval` vrsta meritvenega obdobja (“month” za mesečno, “year” za letno in “decade” za desetletno),

`location` koda države ali številski identifikator vodotočnega območja.



Slika 2.1: Prikaz vodotočnih območij sveta.

Za razliko od programskega vmesnika indikatorjev, nam programski vmesnik podnebnih meritev z eno poizvedbo omogoča dostop do podatkov le za eno državo. To pomeni, da je količina podatkov dovolj omejena, da nam programski vmesnik vedno vrne vse podatke brez paginacije, kot prikazuje primer 10.

```
1  http://climatedataapi.worldbank.org/climateweb/rest/v1/<loc_type>/cru↩
       /<data_type>/<interval>/<location>
```

Primer 9: Osnovna oblika poizvedbe za podnebne podatke.

```
[
    {
        'month': 0,
        'data': 68.93643
    },
    {
        'month': 1,
        'data': 64.23069
    },
    {
        'month': 2,
        'data': 81.098724
    },
    ...
]
```

Primer 10: Primer odgovora za poizvedbo količine padavin v posameznih mesecih v Sloveniji.

2.3 Težave pri uporabi programskih vmesnikov Svetovne banke

Programski vmesniki Svetovne banke zajemajo podatke iz različnih virov, zato je težko zagotoviti pravilnost in konsistentnost podatkov. Poleg tega pa se programski vmesnik in spletna stran z dokumentacijo občasno spremenita, kar povzroča še dodatne težave pri uporabi. Nekatere težave, ki smo jih opazili so:

- nekaterim delom dokumentacije se je med izdelavo te diplomske naloge spremenil spletni naslov, tako da do tistih delov sedaj nimamo več dostopa,
- polje za datum `date` v odgovoru je opisano, vendar niso dokumentirane vse možne vrednosti (nekaj primerov nedokumentiranih vrednosti: “last known value” “2001 - 2015” “2040”),
- delovanje sita z različnimi kombinacijami polj `mrv`, `gapfill` in `date` ni ustrezno opisano,

- v odgovoru poizvedbe po podatkih indikatorjev ponekod manjkajo vrednosti kot so koda države, ime države ali ime indikatorja,
- zgornja meja števila izbranih lokacij na 250 ni navedena, prav tako pa ni dokumentirana napaka, ki jo v tem primeru vrne programski vmesnik,
- nemogoče je ugotoviti pogostost vzorčenja indikatorja `frequency`.

Poglavje 3

Knjižnica in gradniki za Orange

V okviru diplomske naloge smo razvili tri ločene komponente za programerje in končne uporabnike programa Orange. Prva komponenta je dostopna kot samostojni paket `simple_wbd`[1]. Druga in tretja komponenta pa sta združeni v paketu `orange3-datasets`[2].

V nalogi razvita programska knjižnica `simple_wbd` omogoča enostaven dostop do programskega vmesnika indikatorjev in podnebnih podatkov Svetovne banke. Knjižnico smo implementirali z uporabo čim manjšega števila odvisnosti in je namenjena splošni uporabi v programih napisanih v jeziku Python. Poudarka pri zasnovi knjižnice `simple_wbd` sta predvsem enostavnost razširitve in zanesljivost. Ta cilja dosežemo z mehanizmom za vključevanje lastne kode v komponente knjižnice in mehanizmi za popravljanje ali odstranjevanje okvarjenih podatkov.

Drugi sestavni del je razširitev knjižnice `simple_wbd` s funkcionalnostmi potrebnimi za lažje delo v programu Orange. To predvsem zavzema pretvorbo pridobljenih podatkov v podatkovno tabelo Orange in tabelo numpy. Ta sklop je namenjen skriptnemu delu s programom Orange [12] in je dostopen kot modul `api_wrapper` v programskem jeziku Python znotraj paketa `orangecontrib.wbd`.

[1] `https://pypi.python.org/pypi/simple_wbd/0.5.1`
[2] `https://pypi.python.org/pypi/Orange3-Datasets/0.1.3`

Tretji sestavni del naše rešitve je grafični vmesnik za uporabo `api_wrapper` modula. Namen grafičnega vmesnika je omogočiti dostop do podatkov programskega vmesnika Svetovne banke znotraj programa Orange za namen obdelave, analize in iskanja zakonitosti v podatkih s pomočjo vizualizacij in ostalih grafičnih gradnikov, ki jih Orange ponuja.

3.1 Knjižnica simple_wbd

Knjižnica `simple_wbd` programerjem olajša dostop do podatkov programskega vmesnika Svetovne banke. Glavni namen knjižnice je združevanje večjega števila zahtev po podatkih in enostavna predstavitev prejetih rezultatov. Te rezultate je nato iz več dimenzij možno pretvoriti v dvo-dimenzionalno polje, primerno za uporabo v programu Orange. Glavna razreda te knjižnice sta `IndicatorAPI` in `ClimateAPI`. Prvi omogoča pridobivanje podatkov iz programskega vmesnika indikatorjev, drugi pa s programskega vmesnika podnebnih meritev.

Čeprav za dostop do programskega vmesnika Svetovne banke že obstajajo rešitve, kot sta knjižnici `wbdata`[3] in `wbpy`[4], smo se odločili za lastno implementacijo podobne knjižnice. Glavni razlog za to je, da obstoječe rešitve poskušajo čim bolj natančno predstaviti programski vmesnik Svetovne banke, ne pa olajšati dostop do čim večje količine podatkov.

Za potrebe te knjižnice smo razvili lastno rešitev za predpomnjenje poizvedb, saj so se bolj splošne rešitve, kot na primer `vcrpy`[5] in `requests-cache`[6], izkazale za prepočasne, ko delamo z večjimi količinami podatkov. Naša rešitev za predpomnjenje izkorišča dejstvo, da je vsaka poizvedba določena le z naslovom URL in da so vsi odgovori oblike JSON. Za vsak URL naredimo novo datoteko v sistemskem začasnem imeniku, v kateri hranimo serializirane JSON podatke. Ker se podatki na programskem vmesniku Svetovne banke

[3] `https://pypi.python.org/pypi/wbdata/0.2.7`
[4] `https://pypi.python.org/pypi/wbpy/2.0.1`
[5] `https://pypi.python.org/pypi/vcrpy/1.10.0`
[6] `https://pypi.python.org/pypi/requests-cache/0.4.12`

redko posodabljajo, smo za čas veljavnosti začasnih datotek izbrali en teden.

3.1.1 Razred IndicatorAPI

`IndicatorAPI` je razred namenjen pridobivanju podatkov indikatorjev razvoja držav. Ker ima programski vmesnik Svetovne banke omejitev koliko podatkov lahko prenesemo z eno poizvedbo in nam dovoli tvoriti poizvedbe le za en indikator hkrati smo napisali razred, ki v ozadju tvori in izvede poizvedbe za vse strani vseh zahtevanih indikatorjev. Po prvi poizvedbi za en indikator se naša rešitev sprehodi čez število preostalih strani , ki so na voljo, in pridobljene podatke več strani združi in predstavi kot rezultat ene same poizvedbe. Ta postopek ponovi za vse zahtevane indikatorje in njihove rezultate vrne v obliki slovarja, ki ima za ključ kodo indikatorja posamezne zahteve. Poleg tega, da skrbi za prenos vseh strani podatkov, tudi beleži število izvedenih in število potrebnih poizvedb za celoten prenos. Ta števila se lahko uporablja za prikaz napredka prenosa podatkov.

Za namene uporabe v razredu `IndicatorAPI` smo v knjižnici `simple_wbd` razvili mehanizme za odpravo nekaterih napak omenjenih v poglavju 2.3.

Pri manjkajočih vrednostih držav v poizvedbah za podatke indikatorjev poskušamo določiti pravilne vrednosti. To naredimo s pomočjo dveh slovarjev: prvi slika kode držav v imena, drugi pa imena držav v kode. V primeru manjkajoče vrednosti kode ali imena, poskušamo to prebrati iz enega od naštetih slovarjev. Če nam ne uspe ugotoviti manjkajočih vrednosti, trenutni vnos odstranimo iz rezultata poizvedbe.

Drugi tip napak, ki ga lahko delno popravimo, so napačne vrednosti v polju `date` v poizvedbah za podatke indikatorjev. Ker lahko v temu polju pričakujemo poljubno besedilo, dela naš pretvornik za polje `date` v datum, tako da poskuša v datum pretvoriti čim daljšo predpono besedila. Naprimer, v besedilu “2005Q1 - 2006Q2” je najdaljša predpona, ki še označuje veljaven datum opisan v dokumentaciji polja `date`, predpona “2005Q1“. Če nam ne uspe besedila pretvoriti v veljaven datum, trenutni vnos odstranimo iz rezultata poizvedbe.

Glavne metode ki jih ponuja razred IndicatorAPI so:

`get_indicators` za pridobivanje seznama indikatorjev s kodami, imeni in opisi,

`get_countries` za pridobivanje seznama držav z metapodatki,

`get_dataset` za pridobivanje instance razreda `IndicatorDataset`, ki vsebuje podatke indikatorjev.

Ena izmed lastnosti razreda `IndicatorAPI` je, da mu lahko ob inicializaciji podamo razred v katerem želimo prejeti rezultat poizvedbe. Ta razred mora dedovati od osnovnega razreda `IndicatorDataset`. Na ta način lahko enostavno razširimo funkcionalnost `simple_wbd` knjižnice. V primeru 11 vidimo en način za razširitev razreda `IndicatorDataset`, tako da uporabniku razreda `MyIndicatorAPI` ni potrebno izrecno podati razreda `MyIndicatorDataset` v konstruktor.

```
1  class MyIndicatorDataset(simple_wbd.IndicatorDataset):
2
3      def as_numpy(self):
4          raise NotImplemented()
5
6      def as_orange_table(self):
7          raise NotImplemented()
8
9  class MyIndicatorAPI(simple_wbd.IndicatorAPI):
10
11     def __init__(self):
12         super().__init__(MyIndicatorDataset)
```

Primer 11: Primer razširitve osnovnega razreda rezultatov poizvedb.

Razred IndicatorDataset

Razred `IndicatorDataset` je osnovni razred v katerem dobimo zahtevane podatke indikatorjev. Ta razred vsebuje vse potrebne metode in podatke za

predstavitev rezultatov programskega vmesnika na dva načina: kot slovar rezultatov poizvedb za posamezen indikator in dvo dimenzionalen seznam. Posamezna vrednost v teh podatkih je določena z državo, časovno komponento in kodo indikatorja.

Podatke lahko predstavimo kot dvodimenzionalno polje v dveh oblikah: kot časovne vrste ali kot podatki držav. Obliko predstavitve izberemo s parametrom `time_series` metode `as_list`. Za predstavitev obeh oblik je prva vrstica polja uporabljena kot naslovna vrstica, ki opisuje podatke v stolpcih.

Ko uporabljamo obliko časovnih vrst, so elementi prve vrstice kartezični produkt kod indikatorjev in držav. V prvem stolpcu polja pa imamo časovno komponento podatkov. Na ta način so vsi ostali elementi polja določeni s časovno komponento, državo in kodo indikatorja.

Ko dostopamo do dvodimezionalnega polja, ki predstavlja podatke držav, pa je v prvi vrstici kartezični produkt kod indikatorjev in časovne komponente. Prvi stolpec v tej predstavitvi vsebuje imena držav. Za razliko od predstavitve v obliki časovnih vrst, v to polje vstavimo še dodatne stolpce, ki vsebujejo metapodatke držav iz primera 6: regija `region`, administrativna regija `adminregion`, višina dohodka `incomeLevel`, vrsta posojil `lendingType`, geografska širina `latitude`, geografska dolžina `longitude`. Tudi tukaj so vsi ostali elementi določeni s časovno komponento, državo in kodo indikatorja.

3.1.2 Razred ClimateAPI

Razred `ClimateAPI` olajša dostop do podnebnih podatkov programskega vmesnika Svetovne banke. Ta programski vmesnik dovoli poizvedbe po podatkih le ene vrste meritev za eno vrsto meritvenega obdobja in eno državo. Naš razred naredi kartezični produkt med vsemi zahtevanimi vrstami meritev, vrstami meritvenih obdobij in državami. Nato iz tega zgradi in izvede vse poizvedbe in predstavi podatke kot enotni odgovor. V razredu `ClimateAPI` hranimo tudi število vseh potrebnih poizvedb in število že izvedenih poizvedb, kar lahko uporabimo za prikaz napredka prenosa podatkov.

Razred ClimateDataset

Razred `ClimateDataset` je osnovni razred v katerem dobimo zahtevane podatke podnebnih meritev. Vsebuje vse potrebne metode in podatke za predstavitev rezultatov programskega vmesnika na dva glavna načina: kot gnezden slovar in dvodimenzionalen seznam. Posamezna vrednost v teh podatkih je določena z državo, vrsto podatkov in časovno komponento. Poleg omenjenih načinov predstavitve podatkov lahko dostopamo tudi do neobdelanih podatkov prejetih iz programskega vmesnika za vsako poizvedbo posebej.

Časovno komponento rezultata sestavljata vrsta meritvenega obdobja in začetek obdobja meritve. Sestavljeno časovno komponento uporabljamo, da se izognemo dvoumnim primerom vrednosti začetka obdobja za letni in desetletni interval meritev. Primera takih dveh časovnih obdobij sta `'decade - 1990'` in `'year - 1990'`.

Do podatkov predstavljenih z gnezdenim slovarjem lahko dostopamo preko funkcije `as_dict`. V tej funkciji združimo podatke poizvedb programskega vmesnika v gnezden slovar s štirimi nivoji gnezdenja: država, vrsta meritev, vrsta meritvenega obdobja in obdobje meritve. Zadnji nivo gnezdenja vsebuje vrednosti podnebnih meritev.

Pri predstavitvi podatkov kot dvodimenzionalno polje moramo dve od treh komponent podatkov (država `'country'`, vrsta podatkov `'type'`, in časovna komponenta `'interval'`) združiti in ju skupaj prikazati v vrsticah ali stolpcih. Za razliko od razreda `IndicatorDataset`, ki podpira le dve obliki prikaza, lahko v razredu `ClimateDataset` sami določimo katere komponente bodo v stolpcih in katere v vrsticah. Primer 12 prikazuje različne možnosti izborov komponent. Spremenljivki `list1` ind `list2` iz prejšnjega primera prikazujeta privzeto konfiguracijo, kjer imamo v stolpcih kartezični produkt vrst meritev in vrst meritvenih obdobij, v vrsticah pa podatke države. Spremenljivka `list4` prikazuje konfiguracijo za predstavitev v obliki časovnih vrst.

```
1  import simple_wbd
2
3  api = simple_wbd.ClimateAPI()
4  climate_dataset = api.get_instrumental(['svn', 'usa', 'aus'])
5
6  list1 = ds.as_list()
7  list2 = ds.as_list(columns=['type', 'interval'])  # default  value
8  list3 = ds.as_list(columns=['type'])
9  list4 = ds.as_list(columns=['type', 'country'])
10 list5 = ds.as_list(columns=['country'])
```

Primer 12: Prikaz nekaj možnih oblik dvodimezionalnega polja vrednosti.

3.2 Modul api_wrapper

Znotraj paketa `orangecontrib.wbd` smo razvili modul `api_wrapper` v katerem smo razširili razreda `IndicatorDataset` in `ClimateDataset` na način, ki je prikazan v primeru 11. Naša razširitev obema razredoma doda metodi za pretvorbo podatkov v podatkovno tabelo Orange in tabelo numpy.

3.2.1 Razširitev razreda IndicatorDataset

Glavne funkcionalnosti, za uporabo programskega vmesnika indikatorjev, so vključene v naši razširitvi razreda `IndicatorDataset`. To je na prvem mestu metoda `as_numpy_array`, ki rezultat metode `as_list` opisane v poglavju 3.1.1, spremeni v polje numpy in odstrani vse stolpce, ki ne vsebujejo niti ene veljavne vrednosti. Druga metoda pa je `to_orange_table`, ki podatke dobljene iz metode `as_numpy_array`, pretvori v podatkovno tabelo Orange. To tabelo lahko oblikuje kot časovno vrsto ali pa kot seznam držav. Obliko tabele Orange, ki jo želimo izbrati, določimo s parametrom `time_series`. Ta metoda tudi poskrbi za pravilno nastavljeno domeno[7] podatkov.

[7]Domena "Domain" je razred v orodju Orange, ki določa tipe in imena značilk in ciljnih razredov.

Slika 3.1: Skupina gradnikov Data Sets, ki smo jih razvili v pričujoči nalogi.

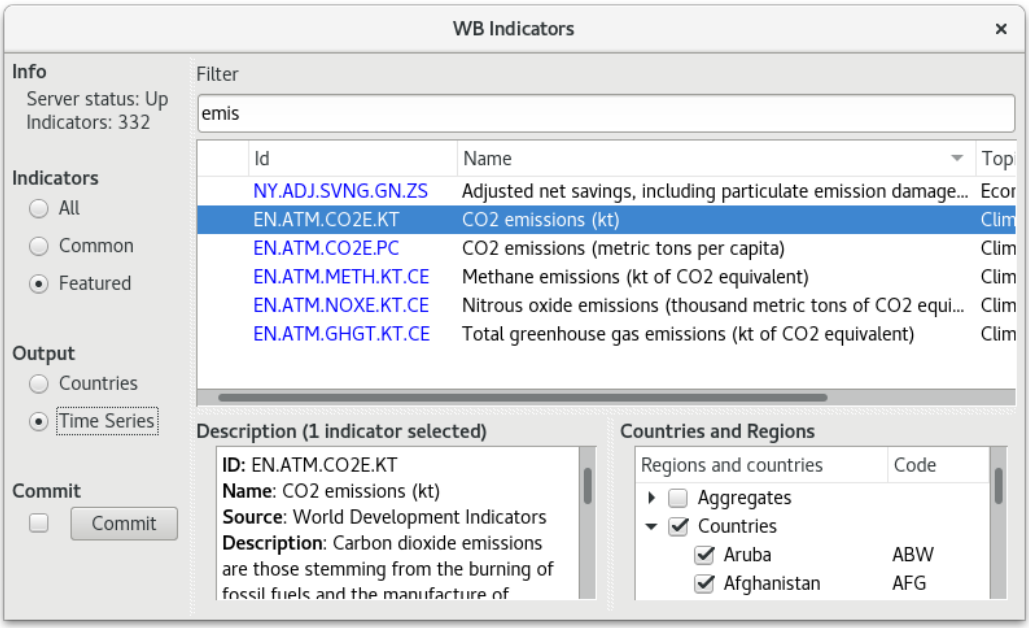
3.2.2 Razširitev razreda ClimateDataset

Prav tako kot razširitev razreda `IndicatorDatasets`, tudi ta razširitev doda metodi `as_numpy_array` in `to_orange_table`. Prav tako kot v razširitvi razreda `IndicatorDatasets`, lahko tudi tukaj s parametrom `time_series` izberemo obliko tabele Orange. Pri vrednosti parametra `time_series = False` se nastavi privzeta oblika tabele, prikazana kot `list1`, sicer pa kot `list3`, iz primera 12. S tem parametrom pa izgubimo možnost poljubne oblike tabele Orange.

3.3 Grafični vmesnik

Programski razširitvi `Orange3-DataSets` za grafični vmesnik programa Orange smo sedaj dodali novo skupino gradnikov imenovano "Data Sets" (slika 3.1). V okviru te naloge smo za skupino "Data Sets" izdelali dva ločena gradnika. Prvi gradnik se imenuje "WB Climate" (slika 3.3) in nam preko grafičnega vmesnika omogoča dostop do podnebnih podatkov Svetovne banke, drugi gradnik pa se imenuje "WB Indicators" (slika 3.2) in nam preko grafičnega vmesnika omogoča dostop do podatkov indikatorjev razvoja.

Oba grafična vmesnika sta narejena skladno z vodili grafičnih vmesnikov programa Orange. To smo dosegli tako, da smo za večino elementov grafičnega vmesnika uporabili predpripravljene gradnike v paketu `Orange.gui`. Pri gradnji vmesnikov smo bili pozorni na odzivnost grafičnega vmesnika. Počasne operacije branja podatkov z interneta smo zato prestavili v ločeno nit.



Slika 3.2: Gradnik WB Indicators.

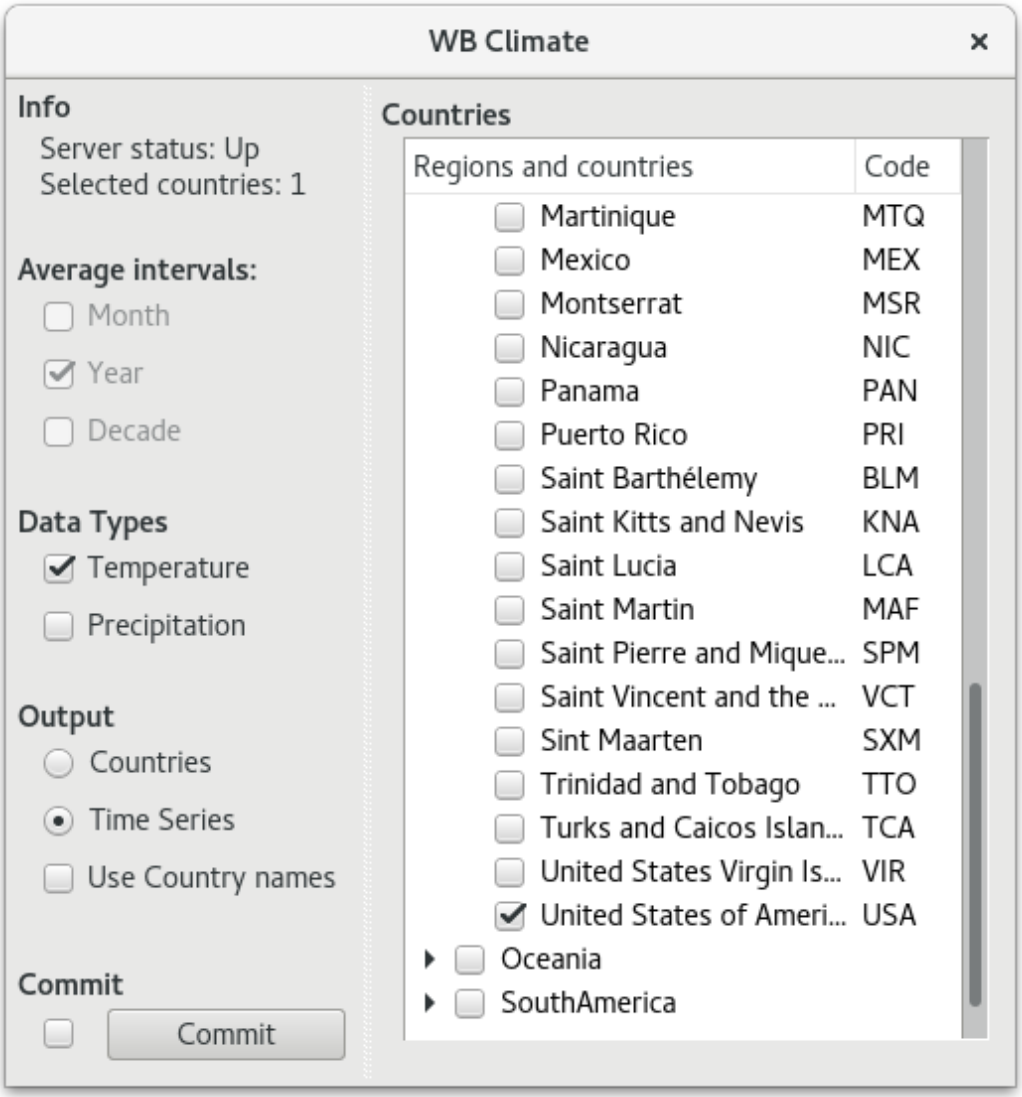
3.3.1 Gradnik WB Indicators

`WB Indicators` je gradnik programa Orange za dostop do podatkov programskega vmesnika indikatorjev. Omogoča nam enostavno izbiro enega ali več indikatorjev in ene ali več držav, za katere želimo dobiti podatke izbranih indikatorjev. Za lažje iskanje indikatorjev smo v grafičnem vmesniku dodali dve možnosti presejanja seznama indikatorjev. Pri prvem situ si lahko izberemo prikaz vseh indikatorjev, pogosto uporabljenih indikatorjev[8] ali pa izpostavljenih indikatorjev[9]. Drugo sito pa je tekstovno presejanje po poljih: koda `id`, ime `name`, teme `topics` in viri `sources`. V grafičnem vmesniku si lahko tudi izberemo eno izmed oblik izhodnih podatkov kot časovne vrste (angl. *Time Series*) ali podatke držav (angl. *Countries*), kot smo ju opisali v razdelku 3.1.1. Implementirali pa smo tudi prikaz napredka prenosa podatkov s števili vseh in že izvedenih poizvedb, omenjenih v razdelku 3.1.1.

[8]Seznam je na voljo na strani `http://data.worldbank.org/indicator?tab=all`
[9]Seznam je na voljo na strani `http://data.worldbank.org/indicator?tab=featured`

3.3.2 Gradnik WB Climate

Gradnik za izbiro podnebnih podatkov podatkovnega vmesnika Svetovne banke nam ponuja možnosti izbire držav, vrste podatkov in vrste meritvenega obdobja. Prav tako kot v gradniku WB Indicators, lahko tudi tukaj izberemo obliko izhodnih podatkov. Možni izbiri oblike izhodnih podatkov sta časovne vrste in podatki držav, kot smo opisali v poglavju 3.2.2. Kot dodatno možnost pa imamo v tem grafičnem vmesniku tudi zastavico, ki določa ali bomo za države izpisovali imena ali pa kode. Tudi temu grafičnemu vmesniku smo dodali prikaz napredka prenosa podatkov.



Slika 3.3: Gradnik WB Climate.

Poglavje 4

Primeri uporabe

4.1 Uporaba modula api_wrapper

Enostavno uporabo modula `api_wrapper` s skriptnim delom programa Orange prikazuje primer 13. V temu primeru pogledamo, kako učinkovito lahko napovemo smrtnost otrok iz raznih indikatorjev zdravja, okolja in infrastrukture. V vrsticah 5 do 15 naredimo poizvedbe po potrebnih podatkih s programskega vmesnika Svetovne banke. Nato v vrsticah 18 do 27 odstranimo vrstice iz tabele, ki nimajo ciljne vrednosti in naredimo novo tabelo z razredom, ki ga želimo napovedovati. Vrednosti, ki jih želimo napovedovati, se nahajajo v stolpcu 55 v tabeli `class_data`. Ta stolpec vsebuje podatke o smrtnosti otrok mlajših od enega leta za leto 2015. V naslednjih vrsticah pa zgradimo tri napovedne modele: naključni gozd z regresijskimi drevesi `rf`, linearna regresija z regularizacijo `ridge` in srednja vrednost `mean`. Za ocene napovednih modelov smo uporabili oceni $RMSE$[1] in R^2 [2]. Iz rezultatov (tabela 4.1) je razvidno, da med izbranimi napovednimi modeli samo naključni gozdovi dajo rezultate, ki so boljši od naklučja.

[1] `https://en.wikipedia.org/wiki/Root-mean-square_deviation`
[2] `https://en.wikipedia.org/wiki/Coefficient_of_determination`

```
import Orange
import numpy as np
from orangecontrib.wbd import api_wrapper

api = api_wrapper.IndicatorAPI()

test_data = api.get_dataset([
    "SH.H2O.SAFE.ZS",  # Improved water source (% of population with access)
    "SH.MED.BEDS.ZS",  # Hospital beds (per 1,000 people)
    "SH.IMM.IDPT",     # Immunization, DPT (% of children ages 12-23 months)
]).as_orange_table()

class_data = api.get_dataset(
    "SP.DYN.IMRT.IN",  # Mortality rate, infant (per 1,000 live births)
).as_orange_table()

# lines with valid class values (not nan)
good_lines = ~np.isnan(np.array(class_data[:,55]))[:,0]

domain = Orange.data.Domain(
    test_data.domain.attributes, class_vars=class_data.domain[55])

data = Orange.data.Table(
    domain,
    np.array(test_data)[good_lines,:],
    np.array(class_data)[good_lines,55]
)

rf = Orange.regression.random_forest.RandomForestRegressionLearner()
ridge = Orange.regression.RidgeRegressionLearner()
mean = Orange.regression.MeanLearner()

learners = [rf, ridge, mean]

res = Orange.evaluation.CrossValidation(data, learners, k=10)
rmse = Orange.evaluation.RMSE(res)
r2 = Orange.evaluation.R2(res)

print("{:25} {:7} {:7}".format("Learner", "RMSE", "R2"))
for i in range(len(learners)):
    print("{:25} {:5.2f} {:6.2f}".format(learners[i].name, rmse[i], r2[i]))
```

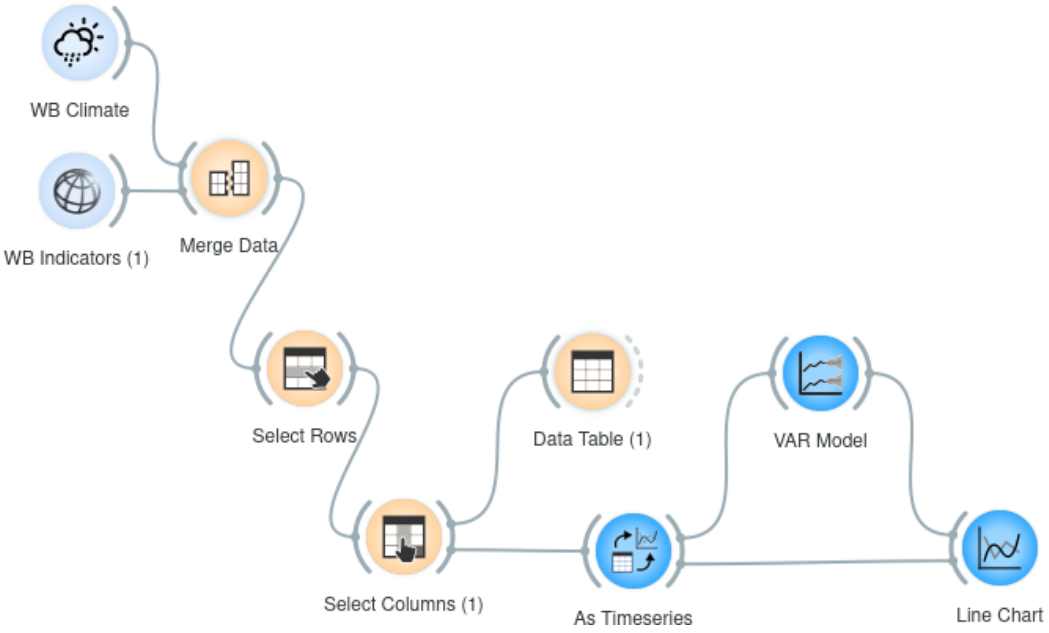
Primer 13: Napovedovanje smrtnosti otrok do enega leta iz podatkov o dostopnosti čiste vode, številu bolniških postelj na 1000 prebivalcev in odstotku cepljenih otrok do drugega leta starosti.

Learner	RMSE	R2
rf	9.74	0.79
ridge	17.76	0.31
mean	21.35	-0.00

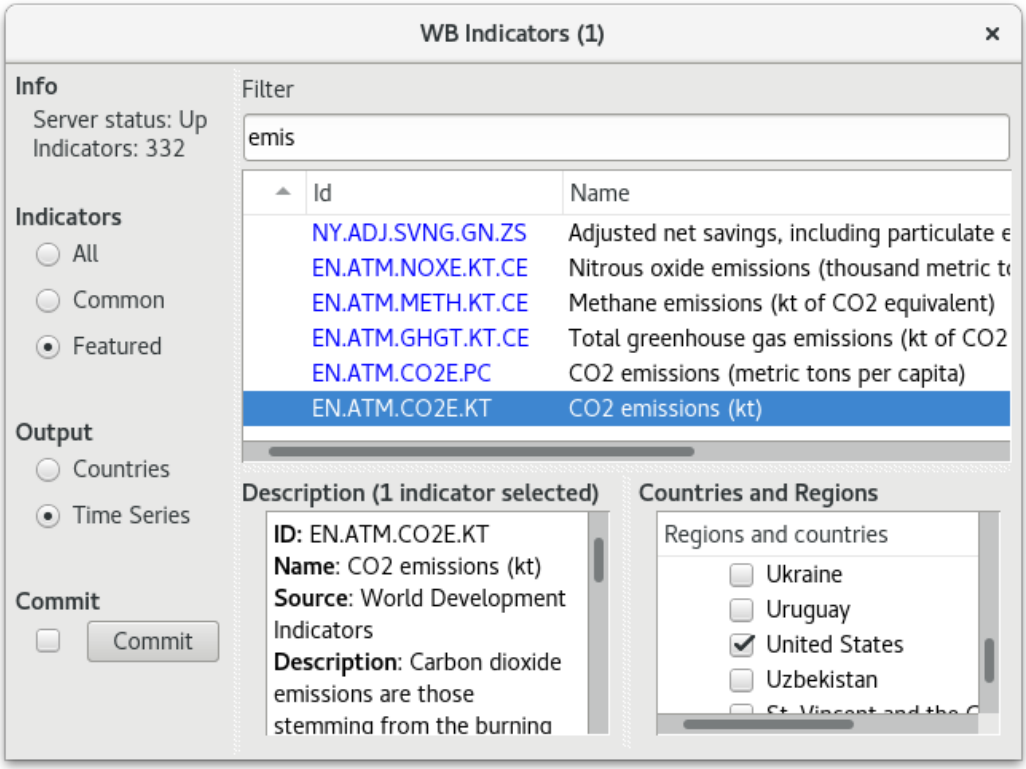
Tabela 4.1: Rezultati napovedi smrtnosti otrok do enega leta starosti.

4.2 Napoved temperature s pomočjo CO_2 izpustov v ZDA

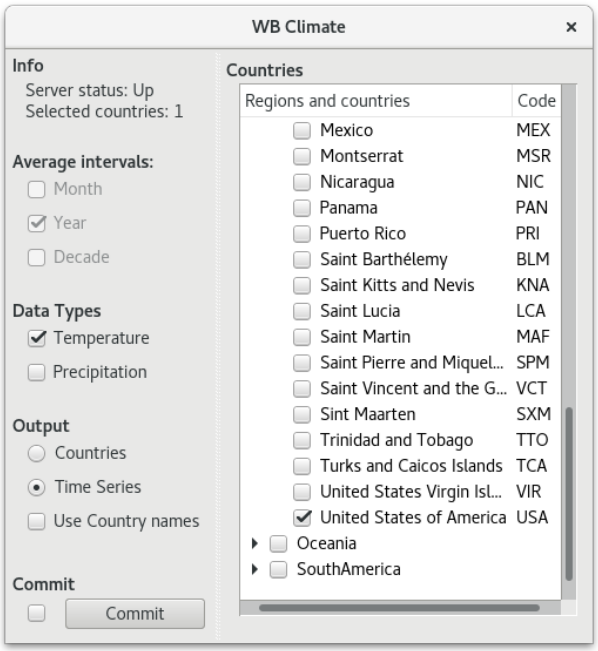
Podatke svetovne banke lahko uporabimo tudi kot časovne vrste z uporabo posebnih gradnikov za delo s časovnimi vrstami [13]. Tukaj si bomo ogledali enostaven primer napovedi temperature v ZDA s pomočjo podatkov o izpustih CO_2. V tej napovedi smo uporabili podatke tako z gradnika WB Indicators (Slika 4.2) kot tudi z gradnika WB Climate (Slika 4.3). Podatke obeh gradnikov smo združili z gradnikom "Merge Data" po obeh časovnih komponentah. Nato smo odstranili vnose časovnih obdobij za katere nimamo na voljo vseh podatkov. Sestavljeno tabelo prikazuje slika 4.4. Iz teh podatkov nato zgradimo časovno vrsto in s pomočjo modela vektorske autoregresije VAR [14] napovemo podatke za povprečno letno temperaturo za naslednjih nekaj let, kar je prikazano na sliki 4.5.



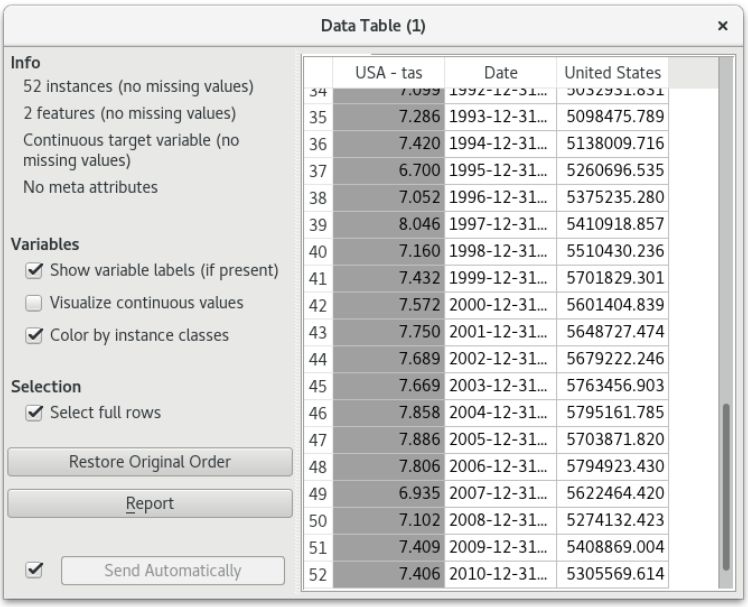
Slika 4.1: Prikaz povezave gradnikov za napoved temperature.



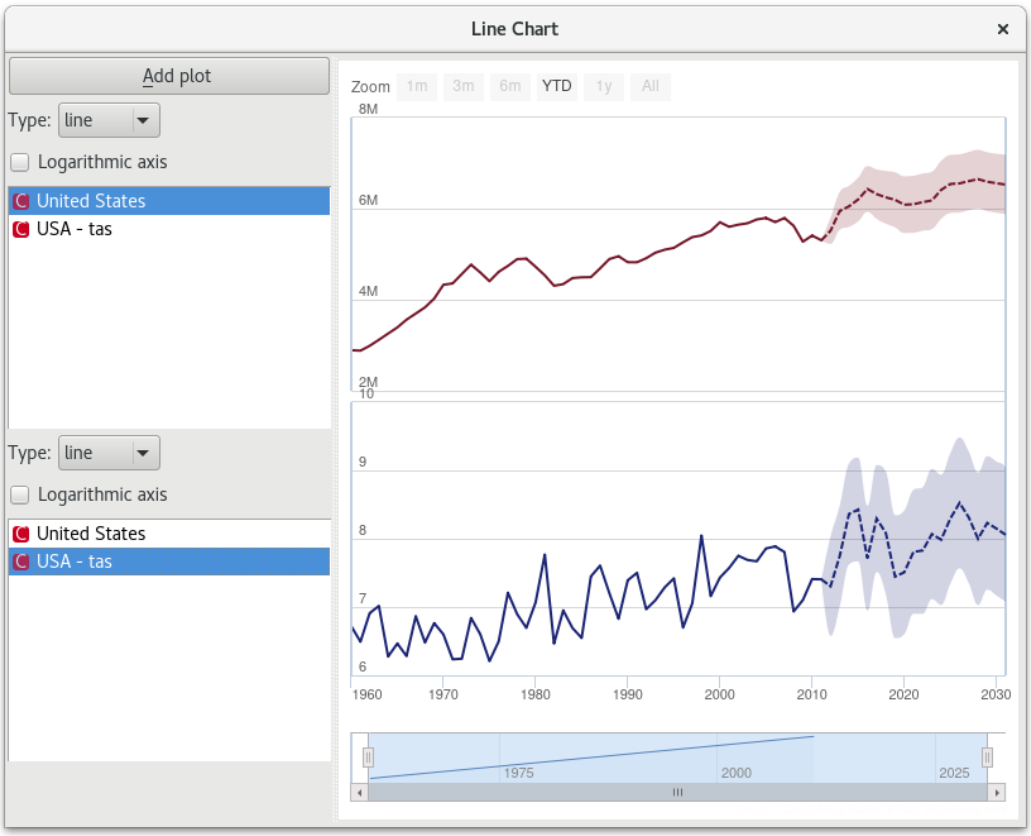
Slika 4.2: Izbor indikatorja CO_2 izpustov v ZDA.



Slika 4.3: Izbor podatkov povprečnih letnih temperatur v ZDA.



Slika 4.4: Podatkovna tabela s ciljnim razredom, in dvema poljema.

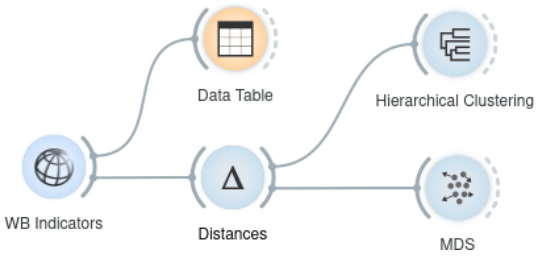


Slika 4.5: Prikaz napovedi gibanja povprečnih letnih temperatur "USA - tas" in CO_2 izpustov "United States".

4.3 Gručenje držav

Podatke, ki jih dobimo z našim dodatkom, lahko v programu Orange uporabimo tudi za grafični prikaz statistik in povezav med državami. Kot možen primer uporabe (Slika 4.6) smo prikazali gručenje držav svetovnih regij glede na naslednje indikatorje (Slika 4.7):

- odstotek ljudi ki živijo v urbanem okolju (angl. *Urban population (% of total)*),
- smrtnost na 1000 živorojenih otrok (angl. *Mortality rate, infant (per*

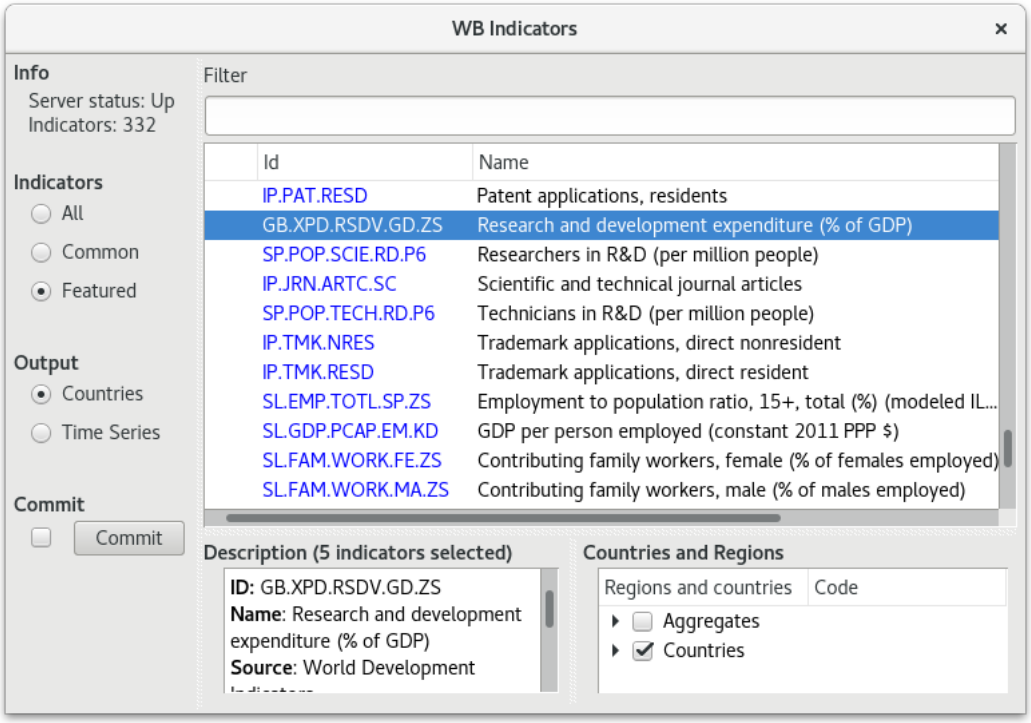


Slika 4.6: Postavitev okolja za prikaz gručenja.

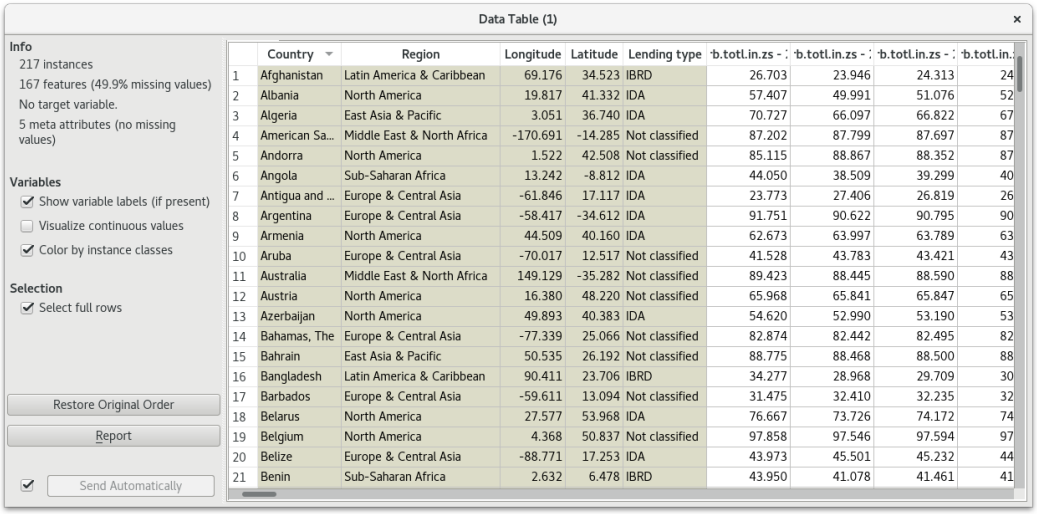
1,000 live births)),

- število bolniških postelj na 1000 prebivalcev (angl. *Hospital beds (per 1,000 people)*),
- delež BDP izdatkov za raziskave in razvoj (angl. *Research and development expenditure (% of GDP)*),
- število prebivalstva pod pragom revščine pri meji 3.10 dolarjev na dan (angl. *Poverty gap at* \$3.10 *a day (2011 PPP) (%)*).

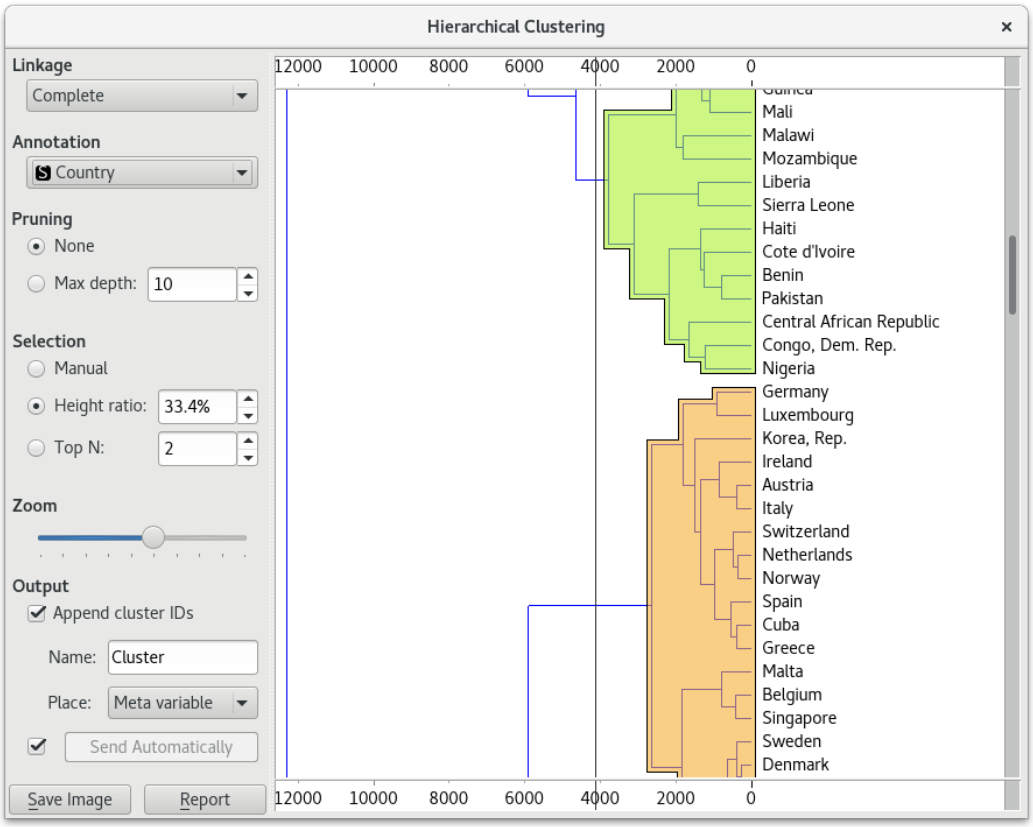
Med temi podatki teh indikatorjev (slika 4.8) smo izračunali evklidsko razdaljo in za prikaz uporabili že obstoječa gradnika programa Orange "MDS" (slika 4.10) in "Hierarchical Clustering" (slika 4.9).



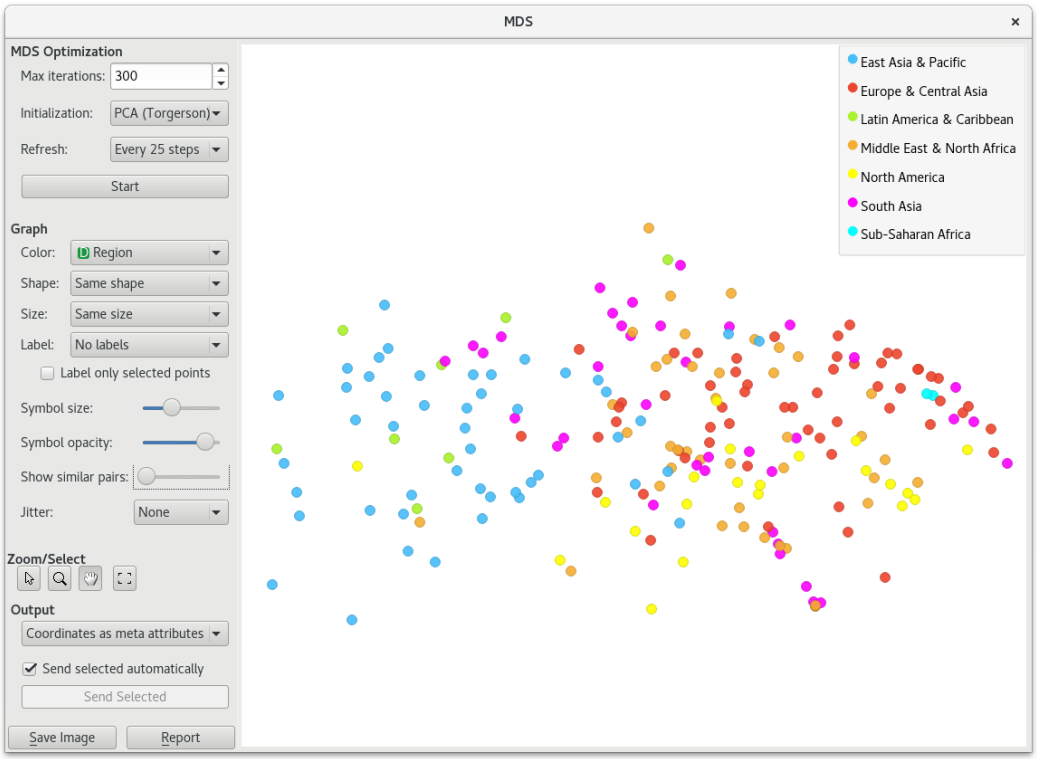
Slika 4.7: Izbor indikatorjev za gručenje.



Slika 4.8: Podatki izbranih indikatorjev.



Slika 4.9: Prikaz hierarhičnega gručenja držav.



Slika 4.10: Prikaz gručenja MDS.

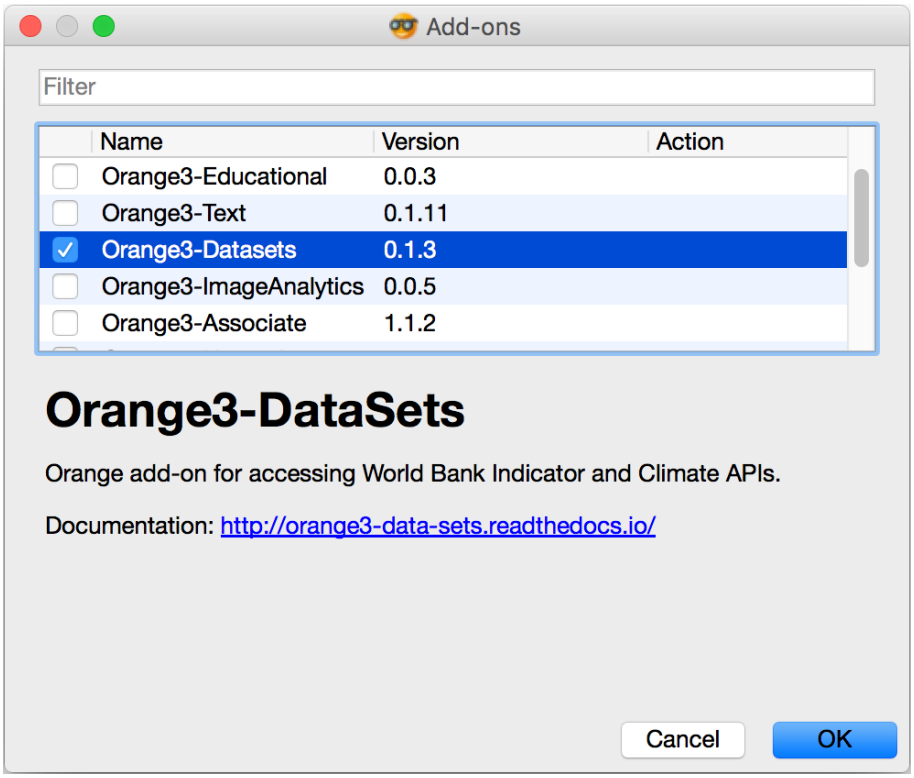
Poglavje 5

Sklepne ugotovitve

Z izdelavo dodatka za program Orange smo zaključili delo na diplomski nalogi. Vsa koda se nahaja na prosto dostopnem repozitoriju GIT na naslovih `https://github.com/zidarsk8/simple_wbd` in `https://github.com/zidarsk8/orange3-data-sets`. Dodatek je že dostopen uporabnikom sistema Orange in ga lahko namestimo s standardnim vmesnikom za delo z dodatki (slika 5.1).

Z razvitim dodatkom smo omogočili dostop do podatkov programskega vmesnika Svetovne banke tako v grafičnem vmesniku kot v skriptnem delu programa Orange. Poleg tega smo z našim vmesnikom tudi poenotili način dostopa do podatkov Svetovne banke v programu Orange in s tem olajšali vzdrževanje in posodabljanje kode v primeru spremembe programskega vmesnika Svetovne banke.

Naš grafični dodatek za dostop do podatkov indikatorjev lahko nadgradimo tako, da uporabnikom grafičnega vmesnika omogočimo večjo izbiro oblik izhodnih podatkov in natančnejše presejanje rezultatov. Dodamo lahko tudi več metapodatkov na posamezne stolpce tabele Orange, ki nam omogočijo boljšo predstavnost v ostalih gradnikih Orange. V grafični vmesnik za dostop do podnebnih podatkov lahko dodamo še možnost izbire vodotočnih območij meritev. Za boljšo predstavo bi lahko postopek izbire držav, regij in vodotočnih območij (slika 2.1) omogočili preko interaktivnega zemljevida sveta.



Slika 5.1: Standardni vmesnik za delo z dodatki sistema Orange, ki že prikaže dodatek, ki je bil razvit v okviru pričujoče naloge.

Literatura

[1] World Development Indicators, The World Bank, (August 2016)
URL: http://data.worldbank.org/data-catalog/world-development-indicators

[2] Data source: Global Financial Development Database (GFDD), The World Bank. Methodology citation: Martin Čihák, Aslı Demirgüç-Kunt, Erik Feyen, and Ross Levine, 2012. "Benchmarking Financial Systems Around the World." World Bank Policy Research Working Paper 6175, World Bank, Washington, D.C. (Junij 2016)
http://data.worldbank.org/data-catalog/global-financial-development

[3] Africa Development Indicators, The World Bank (Februar 2013)
http://data.worldbank.org/data-catalog/africa-development-indicators

[4] Doing Business, The World Bank (http://www.doingbusiness.org) (Julij 2016)
http://data.worldbank.org/data-catalog/doing-business-database

[5] Enterprise Surveys, The World Bank (Julij 2016)
http://data.worldbank.org/data-catalog/enterprise-surveys

[6] Millennium Development Goals, The World Bank (Julij 2016)
http://data.worldbank.org/data-catalog/
millennium-development-indicators

[7] World Bank EdStats (Junij 2016)
http://data.worldbank.org/data-catalog/ed-stats

[8] Gender Statistics, The World Bank (Julij 2016)
http://data.worldbank.org/data-catalog/gender-statistics

[9] HealthStats, World Bank Group (Julij 2016)
http://data.worldbank.org/data-catalog/
health-nutrition-and-population-statistics

[10] IDA Results Measurement System, the World Bank (Julij 2016)
http://data.worldbank.org/data-catalog/
IDA-results-measurement

[11] Climatic Research Unit, University of East Anglia
http://www.cru.uea.ac.uk/data

[12] Janez Demšar and Tomaž Curk and Aleš Erjavec and Črt Gorup and Tomaž Hočevar and Mitar Milutinovič and Martin Možina and Matija Polajnar and Marko Toplak and Anže Starič and Miha Štajdohar and Lan Umek and Lan Žagar and Jure Žbontar and Marinka Žitnik and Blaž Zupan, "Orange: Data Mining Toolbox in Python," Journal of Machine Learning Research, vol. 14, pp. 2349-2353, 2013.

[13] Jernej Kernc, "Orodje za interaktivno analizo časovnih vrst," UL FRI, diplomska naloga, 2016

[14] Eric Zivot, Jiahui Wang, "Vector Autoregressive Models for Multivariate Time Series" Modeling Financial Time Series with S-PLUS, pp. 385-429, 2006.

[15] Jure Dimec (2002), Medjezično iskanje dokumentov
`http://clir.craynaud.com/clir/MEDJEZICNOISKANJEDOKUMENTOV.pdf`