Универзитет у Београду
Математички факултет

Семинарски рад из Истраживања Података 2

Задатак број 6

08.01.2024.

Професор: Ненад Митић

Александра Лабовић и Мила Лукић

6. januar 2024.

Увод

Сврха овог рада је одређивање кретања изолата SARS-CoV-2 коронавируса ка Србији. "Кретање"обухвата упоређивање секвенци и датума појаве одређеног изолата у некој од околних земаља пре појављивања у Србији, као и приказ њиховог груписања. Такође је потребно одредити и проценат разлика одређеног изолата у Србији који је "дошао"у Србију из околних земаља. Техника истраживања података коришћена у овом раду је анализа временских серија.

Скуп података

У скупу података са којима смо радили подаци су у FASTA формату, који се често користи за представљање нуклеотидних или протеинских секвенци. Линија заглавља, која почиње са »", пружа информације о секвенци.
На слици у наставку је један пример инстанце скупа података за државу Србију (са скраћеном нуклеотидном секвенцом).

```
>hCoV-19/Serbia/P1135902-0210/2021|EPI_ISL_10123048|2021-10-02
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA
AATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
ACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTTTTGCAGCCGATCATCAGCACATCTAGGTTT
TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGC
CTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACAT
CTTAAAGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTC
GTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAG
AACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACG
GAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTA
GCACGTGCTGGTAAAGCTTCATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTTGAAATTAAAT
TGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGACGGGCGATTTTGTTAAAG
CCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGAAGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTT
GTTAAAATTTATTGTCCAGCATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGG
```

Slika 1: Једна инстанца скупа података за Србију

Заглавље садржи информације о одређеном случају вируса hCoV-19 (SARS-CoV-2), укључујући детаље као што су:

- Локација (Србија)
- Идентификатор узорка (P1135902-0210)
- Датум прикупљања узорка (2021-10-02)
- Број приступа секвенци (EPI_ISL_10123048).

Сама секвенца представља генетски код вируса, при чему свако слово представља нуклеотид (А, Т, Ц или Г). Секвенца је прилично дуга, тако да је овде приказан само део. Потпуна секвенца садржи информације о генетском саставу вируса, што је кључно за проучавање његове структуре, функције и еволуције.

Државе које смо обрађивали у склопу овог пројекта су Србија, Албанија, Северна Македонија, Босна и Херцеговина, Црна Гора и Мађарска.

Претпроцесирање

Претпроцесирање података подразумева:

1. **Одабир релевантних атрибута за одређени модел:**
   - Одабир само битних атрибута који доприносе моделу и уклањање непотребних.
2. **Преношење једног типа податка у други:**
   - Конвертовање или преношење података из једног формата у други, на пример, из текстуалног у числовни формат.
3. **Чишћење података:**
   - Уклањање или исправљање непрецизних, непотпуних или недостајућих података.
   - Обрада аномалија и уклањање шума из података.
4. **Редукција и трансформација података:**
   - Смањивање димензионалности података како би се олакшала анализа и убрзао рад модела.
   - Примена трансформација како би се подаци адаптирали боље моделу.

Ови кораци нису универзални и зависе од специфичности задатка и типа података. У наставку ће бити извршена индивидуална, специфична предобрада података.

Брисање Дупликата

Први корак при претпроцесирању ових података је писање Python скрипте која ће обрадити FASTA фајлове, идентификовати и уклонити дуплиране секвенце у њима, а затим и сачувати јединствене секвенце у нове фајлове. Скрипта користи "Biopython"библиотеку за рад са биолошким подацима.

Функција "remove_duplicates"

Функција remove_duplicates дефинише се како би обрадила улазни FASTA фајл и генерисала излазни фајл без дупликата.

- Секвенце унутар улазног FASTA фајла се читају коришћењем `SeqIO.parse` функције, а затим се смештају у листу под називом `sequences`.
- Елиминисањем дуплираних секвенци на основу њиховог садржаја, формира се листа `unique_sequences` која садржи јединствене секвенце.
- Те јединствене секвенце затим се записују у излазни FASTA фајл коришћењем `SeqIO.write`.

Ова функција се позива за сваку државу унутар директоријума "Балкан": Албанију, Босну и Херцеговину, Црну Гору, Мађарску, Македонију и Србију.

Резултати брисања дупликата

Tabela 1: Број секвенци на почетку и број јединствених секвенци за сваку државу.

Држава	Број секвенци на почетку	Број јединствених секвенци
Албанија	555	543
БиХ	797	780
ЦГ	449	438
Мађарска	567	485
Македонија	40	39
Србија	691	651

Након позивања функције ‘remove_duplicates‘ укупно је уклоњено 163 секвенце, односно приближно 5.3% података.

Додавање референтног генома у сваку државу

Други корак претпроцесирања података је додавање референтног генома на почетак цильних секвенци у одређеном директоријуму. Python скрипта итерира кроз датотеке у одређеном

директоријуму, проверава одређене услове, и за сваку одабрану датотеку додаје референтни геном на почетак циљне секвенце.

Функција dodaj_referentni_genom

- Функција прима два параметра - `direktorijum` (путања до директоријума са циљним датотекама) и `referentna_datoteka` (путања до датотеке са референтним геномом).
- Итерира се кроз све директоријуме, у потрази за циљним датотекама. Циљну датотеку представља сваки фајл који почиње са "unique"и завршава са ".fasta".
- Отвара се референтна датотека у режиму читања (`"r"`) и чита се целокупна референтна секвенца.
- Отвара се циљна датотека у режиму читања (`"r"`) и чита се циљна секвенца.
- Циљна датотека се поново отвара, овај пут у режиму писања (`"w"`) и у њу се уписује комбинација референтне и циљне секвенце.

Поравнања јединствених секвенци

Након брисања дупликата и додавања референтног генома на почетак циљних секвенци за сваку државу, потребно је поравнати секвенце унутар сваког од директоријума. Ово је урађено коришћењем MAFFT програма.

MAFFT (Multiple Alignment using Fast Fourier Transform) је програм за поравнање биолошких секвенци, специфично за поравнање нуклеотидних и протеинских секвенци. Главни циљ MAFFT-a је да пружи ефикасно и прецизно поравнавање секвенци у скупу података који може садржавати велики број секвенци.

MAFFT користи итеративни алгоритам Брзе Фуријеове Трансформације, посебно дизајниран за обраду великог броја секвенци, што му омогућава ефикасност. Може се користити интерактивно, али се често користи и у скриптама за аутоматизацију поравнавања. За сврхе овог пројекта, коришћен је у склопу bash скрипте.

Сви геноми су поравнати у односу на први геном у FASTA фајлу, односно, референтни геном.

Израчунавање позиција мутација

Следећи корак претпроцесирања је проналажење мутација на секвенцама. Мутације представљају било каква одступања од референтног генома. Ово смо радили коришћењем функције pronadji_pozicije_razlika.

Функција pronadji_pozicije_razlika

- Функција прима референтну секвенцу, тренутну секвенцу, име тренутне секвенце и објекат за писање CSV датотеке.
- Користи се `enumerate` да би се итерирало кроз позиције у секвенци, а затим пронашли све позиције на којима се разликују референтна и тренутна секвенца.
- Ако постоје разлике, додаје ред у CSV датотеку са именом секвенце и позицијама разлика.

Мутације се проналазе на сваком поравнатом fasta фајлу, односно за сваку државу у директоријуму "Balkan".
Постоје две верзије ове функције: једна која рачуна карактер "n"(недостајућу вредност) као мутацију, и друга која то не сматра мутацијом.

Повезивање CSV и TSV фајлова

Након израчунавања позиција мутација, повезују се CSV и TSV фајлови за сваку државу.

Спајање Држава

Као последњи корак претпроцесирања, сви спојени CSV и TSV фајлови се спајају у један: *sve_povezane_informacije.csv*.

Алгоритми кластеровања

Кластеровање је техника која се користи како би се груписали слични објекти или подаци на основу њихових карактеристика. Циљ кластеровања је идентификовати природне групе или образце унутар скупа података, при чему су објекти унутар исте групе сличнији једни другима него објектима у другим групама.

У кластеровању, алгоритму нису унапред дефинисане ознаке или категорије. Уместо тога, алгоритам анализира податке и додељује објекте кластерима на основу њихове сличности. Сличност између објеката одређује се узимајући у обзир карактеристике или атрибуте. Често коришћени алгоритми кластеровања укључују K-средина, хијерархијско кластеровање и DBSCAN (кластеровање на основу густине просторних података са шумом).

У овом раду, користили смо алгоритам K-средина и алгоритам хијерархијског кластеровања.

Пре примене алгоритама, било је неопходно сортирати све податке у односу на датум преузимања узорка (“Collection date” колона). Затим, сви датуми се мењају у тип datetime помоћу "pd.to_datetime" функције. NaN вредности унутар података о мутацијама се замењују нулама помоћу "np.nan_to_num" функције.

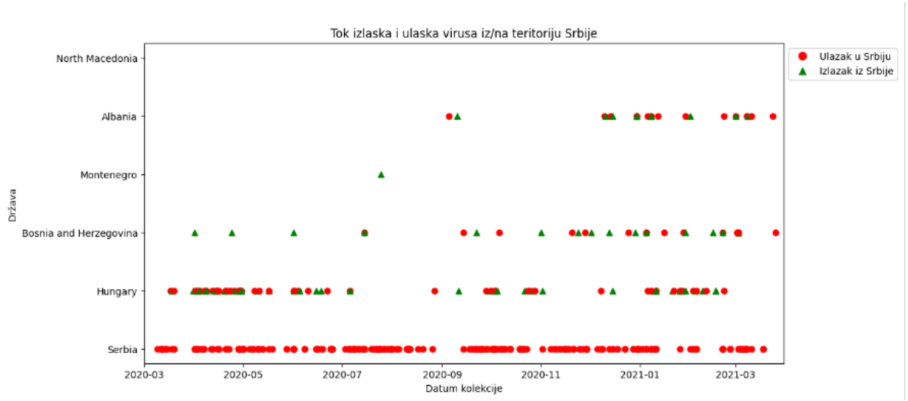
Кластеровање K-средина

K-средина (K-means) је један од најчешће коришћених алгоритама кластеровања у машинском учењу и анализи података. Циљ му је да подели дати скуп података на k кластера, при чему свака тачка података припада кластеру са најближим средњим вредностима (центром). Циљ K-средина алгоритма је да минимизује збир квадратних удаљености између тачака података и њихових одговарајућих кластер центара.

При коришћењу овог алгоритма, морамо унапред задати број кластера које тражимо. Број кластера је постављен на 5.

Формирана је листа "flow_data"која садржи информације о току изласка и уласка вируса из територије Србије и ка територији Србије за сваки кластер. Подаци се прикупљају сортирани према датуму.

Визуелизација и резултати



Slika 2: Визуелизација за алгоритам кластеровања K-средина

- На основу визуелизације видимо да током прве године корона вируса није било никаквих улазака ни излазака између Србије и Северне Македоније.
- Такође, можемо приметити да је у Црну Гору вирус ушао из Србије у августу 2020. године.

- Када је у питању Албанија, током прве половине прве године пандемије није забележен ниједан улазак ни излазак вируса, са првим појављивањем размене у септембру 2020. године. Међутим, између децембра 2020. и марта 2021. године забележено је више улазака и излазака вируса између ове две државе.
- У Босну и Херцеговину је вирус претежно долазио из Србије у првој половини 2020. године, са паузом између Јула и Септембра када није било "размена". Међутим, од Октобра 2020. до Марта 2021. види се значајан пораст и у излазцима вируса из Србије ка Босни и Херцеговини, и обрнуто.
- Мађарска је држава из које је највише инстанци вируса дошло у Србију, поготово током Марта, Априла и Маја 2020. године. Између Јула и Септембра 2020. године је размена стагнирала, али се убрзо вратила и настављена је до Марта 2021. године.

Алгоритам Сакупљајућег Кластеровања

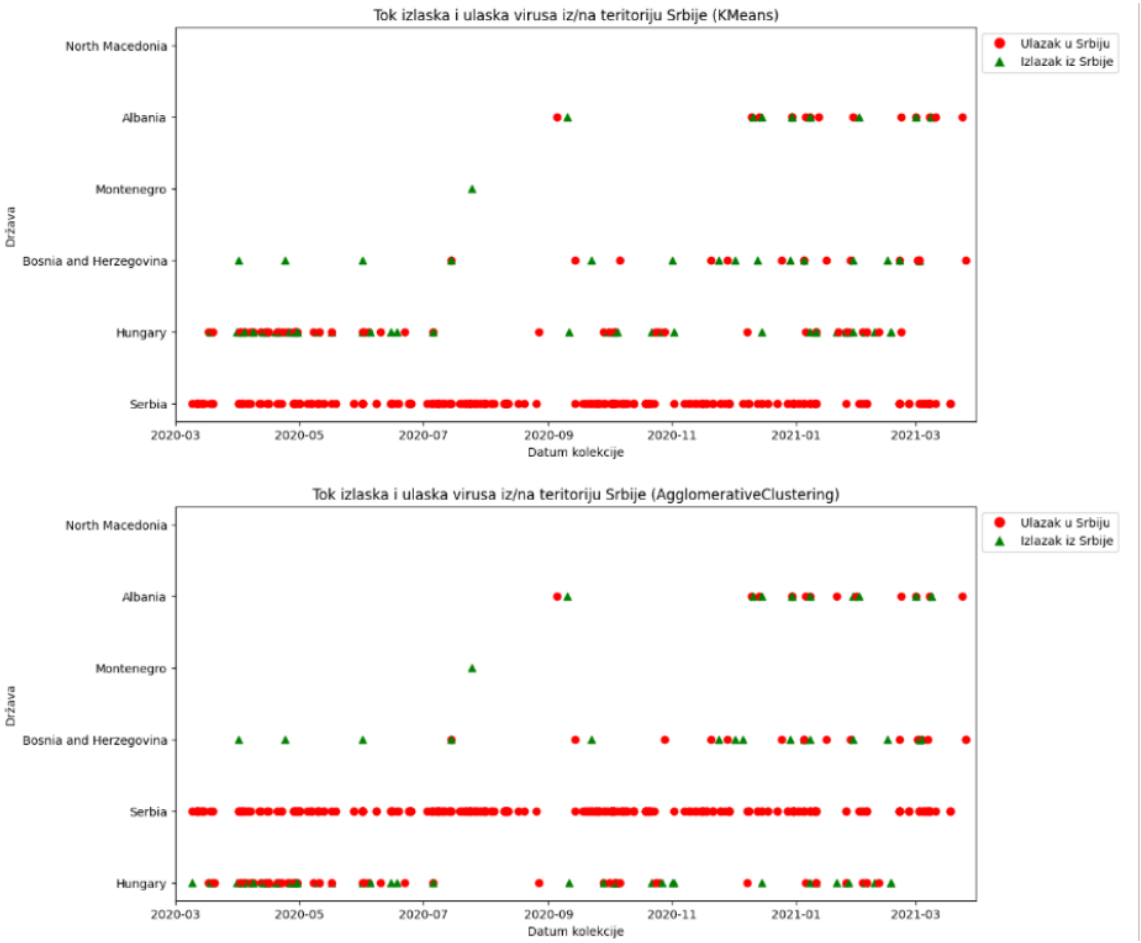
Алгоритам Сакупљајућег Кластеровања (Agglomerative Clustering) је алгоритам хијерархијског кластеровања који се користи за груписање сличних тачака података у кластере.

Ово је приступ одоздо према горе (bottom-up), где свака тачка података почиње као сопствени кластер, а затим се итеративно спајају кластери на основу њихове сличности све док се не достигне жељени број кластера.

Претпроцесирање скупа података и припрема за овај алгоритам идентична је као за алгоритам К-средина.

Визуелизација и резултати

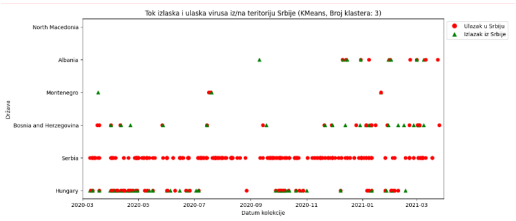
Када је у питању овај алгоритам, са графика не уочавамо велику разлику у односу на алгоритам К-средина. На слици у наставку су приказане визуелизације оба алгоритма, једна до друге.



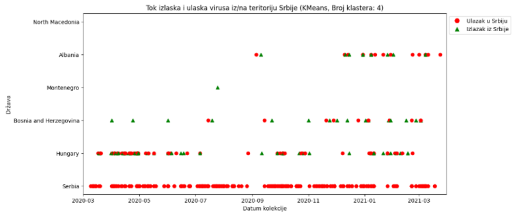
Slika 3: Визуелизација за алгоритам кластеровања К-средина и алгоритам сакупљајућег кластеровања

Приказ резултата у зависности од броја кластера

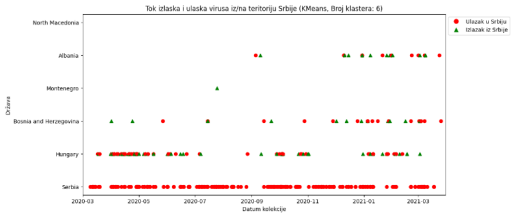
Независно од изабраног броја кластера у оба алгоритма, примећујемо да резултати конвергирају ка конзистентним и сличним информацијама. Ова конзистентност у резултатима може указивати на стабилност и репродуктивност алгоритама класификације у анализи података.
У наставку можемо видети визуелизацију резултата за алгоритам К-средина за који је узето 3, 4, 6 и 7 кластера. Примећујемо да су сви веома слични резултатима за узетих 5 кластера.



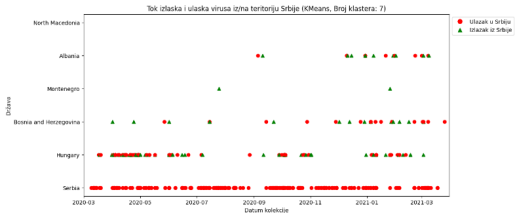
Slika 4: К-средина, 3 кластера



Slika 5: К-средина, 4 кластера



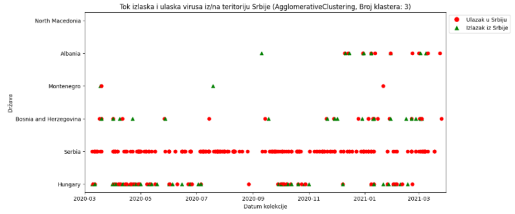
Slika 6: К-средина, 6 кластера



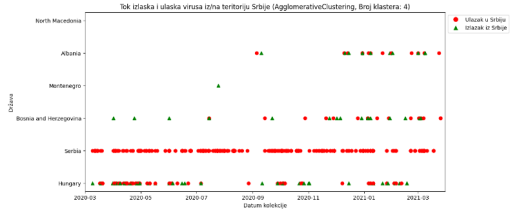
Slika 7: К-средина, 7 кластера

Slika 8: Поређење резултата алгоритма К-средина за 3, 4, 6 и 7 кластера

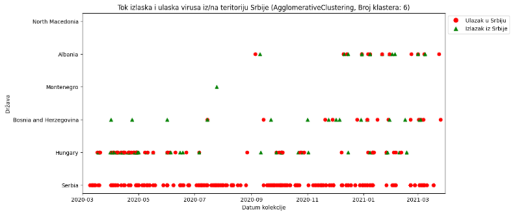
Исту појаву можемо приметити код алгоритма скупљајућег кластеровања за који је узето 3, 4, 6 и 7 кластера. Примећујемо да су сви веома слични резултатима за узетих 5 кластера.



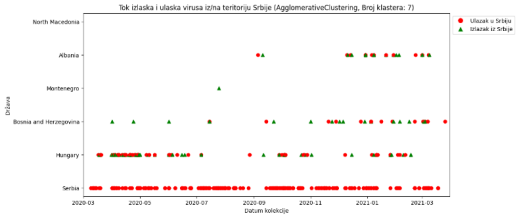
Slika 9: Скупљајуће кластеровање, 3 кластера



Slika 10: Скупљајуће кластеровање, 4 кластера



Slika 11: Скупљајуће кластеровање, 6 кластера

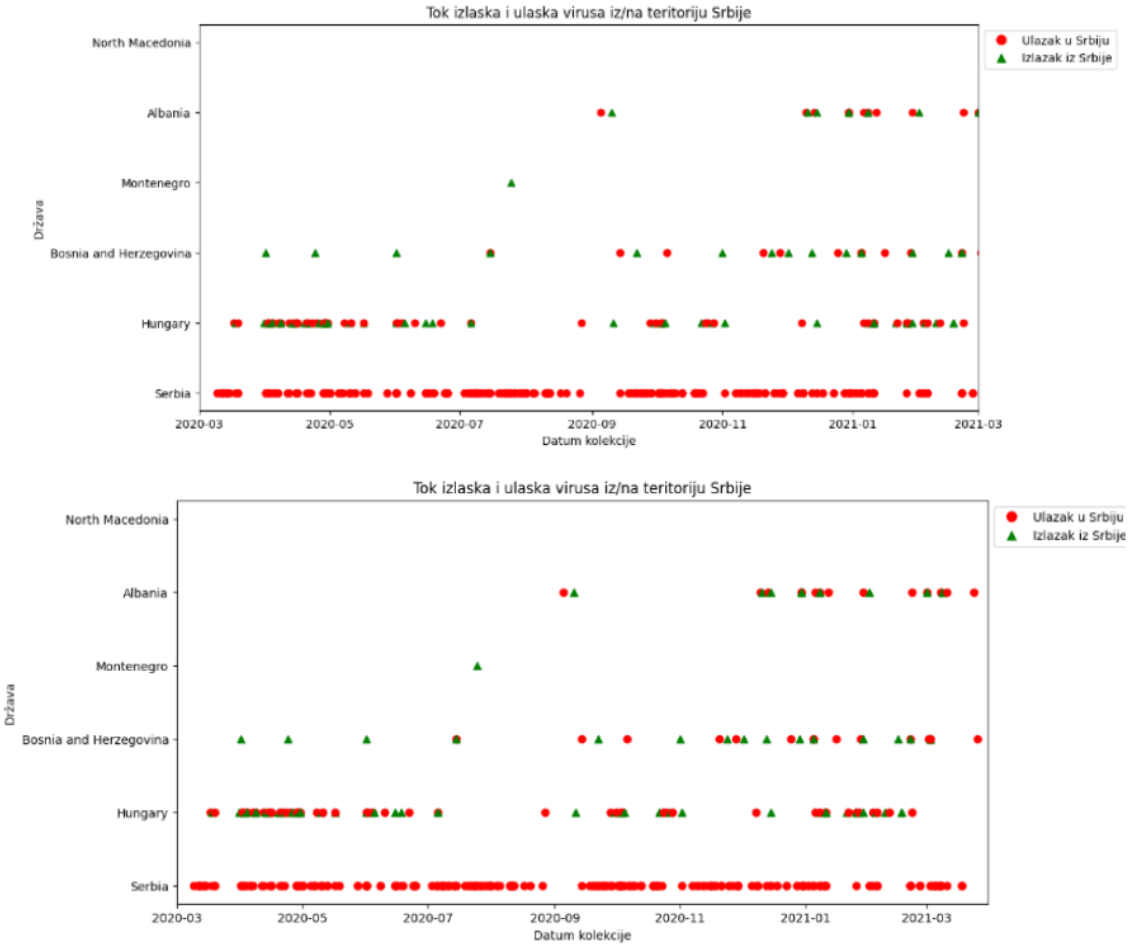


Slika 12: Скупљајуће кластеровање, 7 кластера

Slika 13: Поређење резултата алгоритма K-средина за 3, 4, 6 и 7 кластера

Приказ резултата у зависности од рачунања недостајућих вредности

Иако смо применили алгоритме на два различита CSV фајла - један у коме су недостајуће вредности третиране као мутације, и један у коме нису, нисмо увидели разлику између резултата извршених алгоритама.



Slika 14: Визуелизација за алгоритам кластеровања K-средина са два различита начина претпроцесирања недостајућих вредности

Визуелизација промена мутација кроз време

Последњи део пројекта заснива се на одређивању периода током којих је примећено највише мутација,

Промена броја мутација на нивоу појединачних држава кроз време

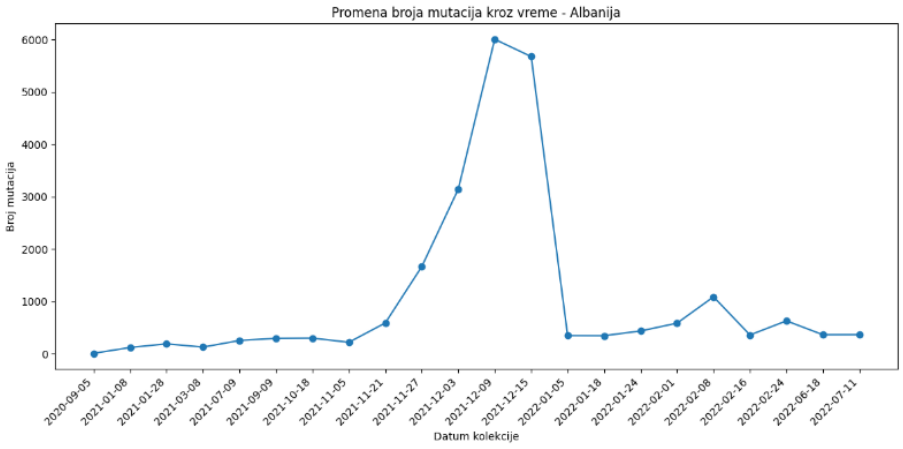
Први корак при истраживању промена мутација је анализа и визуализација промене броја мутација кроз време на основу података из CSV фајлова.

Дефинисана је функција "analiziraj_promene_mutacija"која прима путању до CSV фајла и име поддиректоријума. Функција врши анализу података из CSV фајла и затим врши визуализацију промене броја мутација кроз време. Ова функција биће позвана за сваки поддиректоријум директоријума "Балкан".

Промене мутација кроз време за сваку државу можемо видети у наставку текста.

Албанија

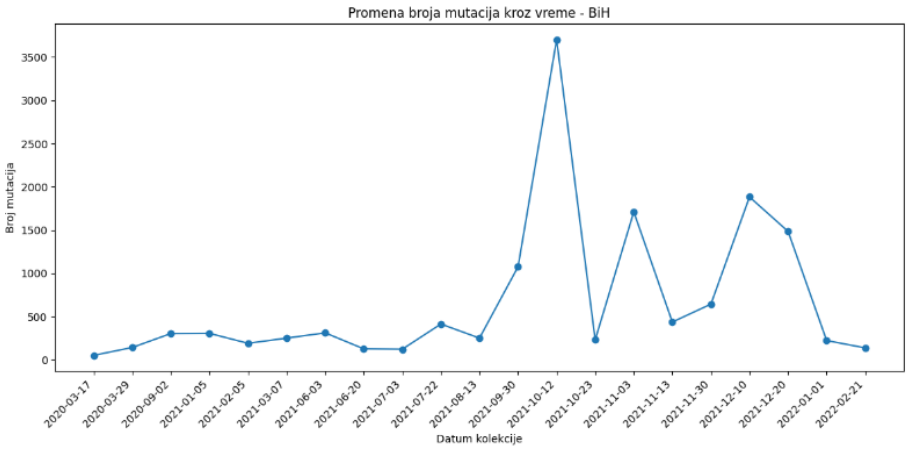
На основу графика промена броја мутација кроз време можемо приметити да је вирус изненада почео да мутира крајем Новембра 2021. године. Највећи број мутација у овој држави примећен је прве недеље Децембра 2021. године - чак 6000 мутација, након чега је значајно опао током друге недеље Децембра - на испод 1000 мутација.



Slika 15: Промене броја мутација кроз време - Албанија

Босна и Херцеговина

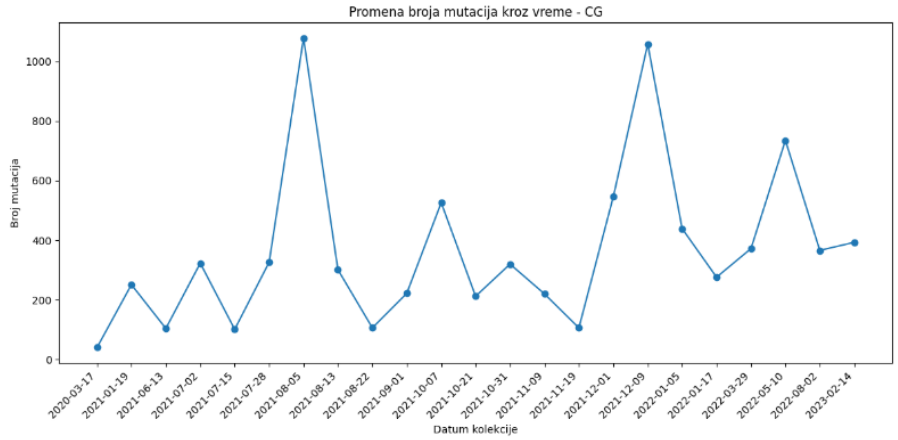
На основу графика промена броја мутација кроз време можемо приметити да је вирус у Босни и Херцеговини мутирао мање него у Албанији. Највећи број мутација у овој држави примећен је средином Октобра 2021. године - са 3500 мутација, скоро дупло мање од Албаније при врхунцу броја мутација. Након овог врхунца, убрзо је уследио и значајан пад, и током следећих неколико месеци нису забележене инстанце са преко 2000 мутација.



Slika 16: Промене броја мутација кроз време - Босна и Херцеговина

Црна Гора

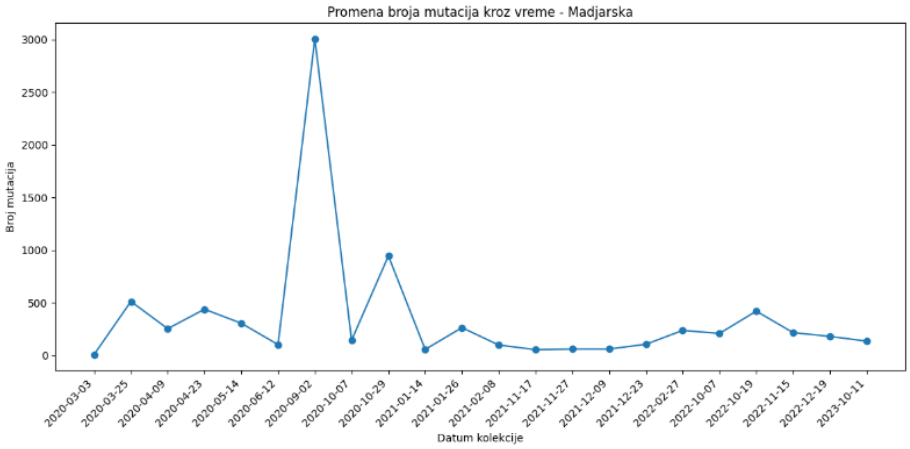
У Црној Гори забележено је мање мутација у односу на Босну и Херцеговину и Албанију. Максимум од 1000 мутација достигнут је почетком Маја 2021. године, као и почетком Децембра исте године.



Slika 17: Промене броја мутација кроз време - Црна Гора

Мађарска

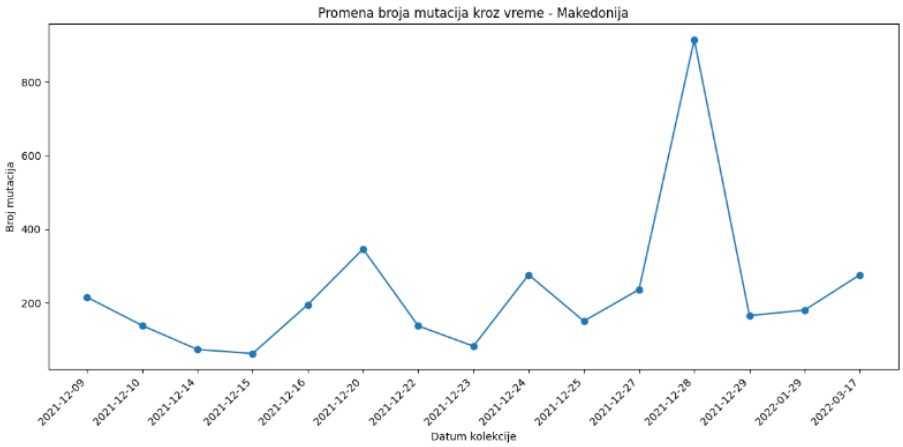
На основу графика промена броја мутација кроз време примећујемо да је Мађарска већину периода од Марта 2020. до Децембра 2022. године имала веома мали број мутација вируса (испод 500), са изузетком прве недеље Септембра 2020. године, када је забележено чак 3000 мутација вируса.



Slika 18: Промене броја мутација кроз време - Мађарска

Македонија

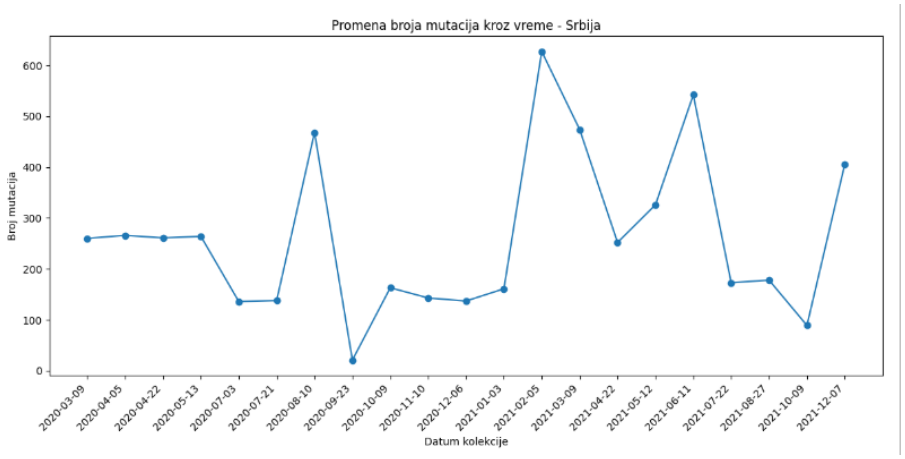
На основу графика промена броја мутација кроз време примећујемо да је Македонија већину периода од Марта 2020. до Децембра 2022. године имала мали број мутација вируса (испод 800). Врхунац броја мутација достигнут је крајем Децембра 2021. године, када је забележено 800 мутација.



Slika 19: Промене броја мутација кроз време - Македонија

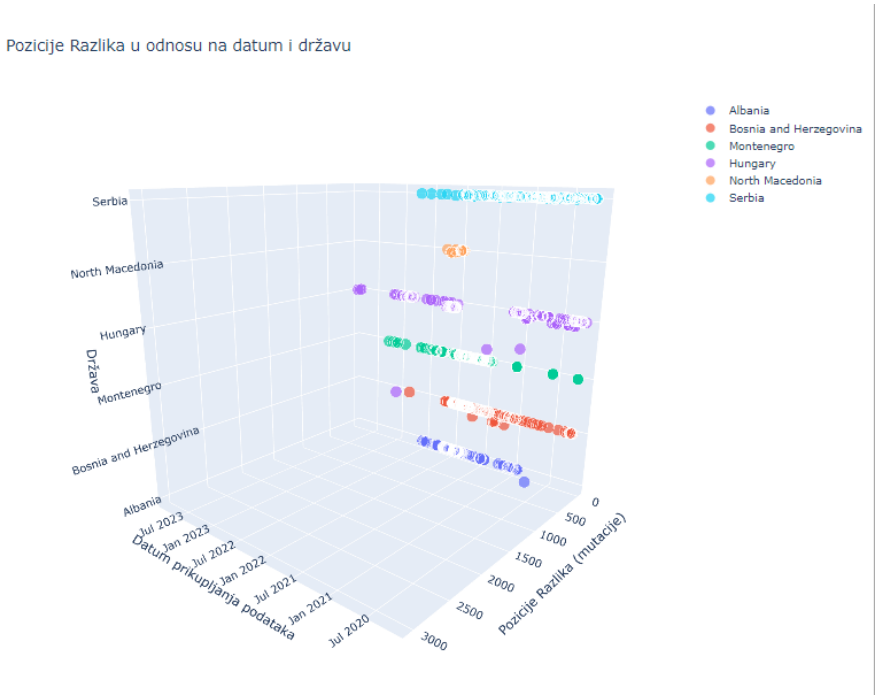
Србија

На основу графика промена броја мутација кроз време примећујемо да је Србија већину периода од Марта 2020. до Децембра 2021. године имала мали број мутација вируса (испод 600). Ово је држава из нашег скупа података за коју је забележено најмање мутација вируса. Врхунац броја мутација достигнут је у Фебруару 2021. године, када је забележено 600 мутација. Поред тога, значајан скок се појављује и у Јуну 2021, када је забележено 500 мутација.



Slika 20: Промене броја мутација кроз време - Србија

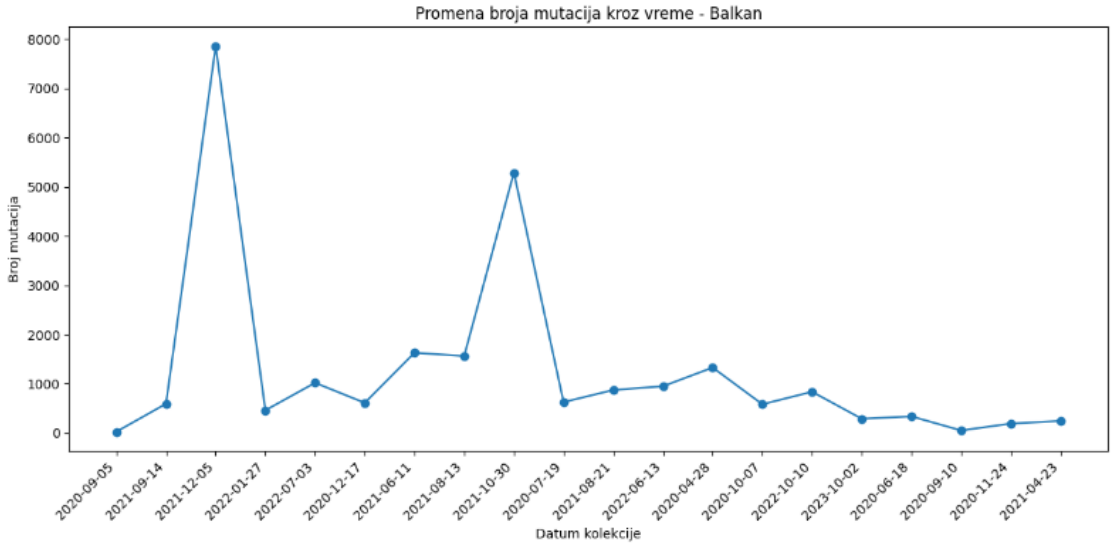
Промена броја мутација на нивоу појединачних држава кроз време - тродимензионални приказ



Slika 21: Промене броја мутација кроз време - Балкан

Промена броја мутација на нивоу Балкана кроз време

На основу графика о промени броја мутација на нивоу Балкана кроз време примећујемо да је највећи број мутација забележен у првој недељи Децембра 2021. године са скоро 8000 промена. Након тога, највећи број мутација забележен је крајем Октобра 2021. Током остатка периода пандемије, у просеку је било испод 2000 мутација сваког месеца.

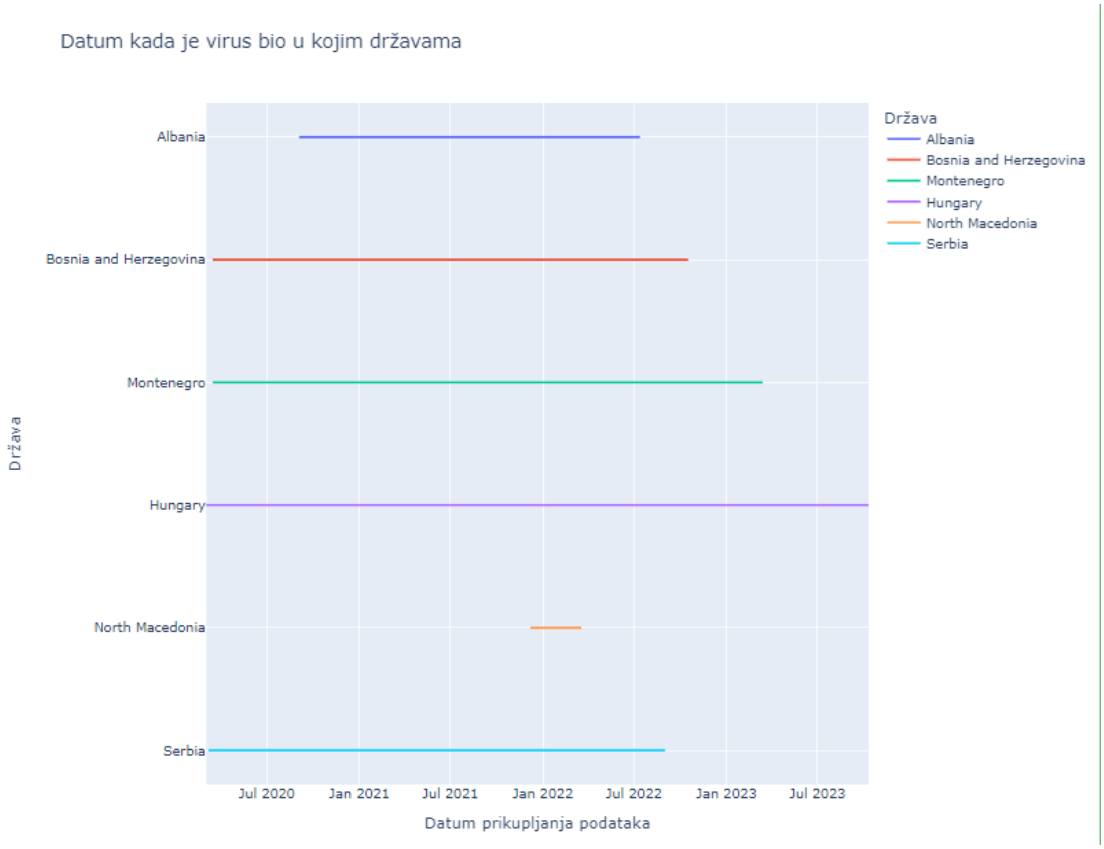


Slika 22: Промене броја мутација кроз време - Балкан

Визуелизација локација и датума прикупљања података о вирусу

Након анализирања промена мутација на нивоу појединачних држава и Балкана, испитали смо датуме када су прикупљене инстанце вируса, као и локације на којима се у том тренутку нашао. Ово нам је омогућило да прикажемо где је вирус био присутан у одређеним периодима.

Са графика примећујемо да је вирус био присутан у Мађарској од Јула 2020. до Јула 2023, док је у Македонији био присутан само током прве половине 2022. године. У Албанији се појављује између Јануара 2021. и Јула 2022. У Босни и Херцеговини га налазимо између Јула 2020. и Јула 2022, а у Црној Гори између Јула 2020. и Јануара 2023. У Србији је вирус, наизглед, присутан од Јула 2020. до Августа 2022.



Slika 23: Датуми када је вирус био у свакој од држава