Анализа на податочно множество за бруто домашен производ по жител

Матеј Митев – 221039

Вовед

Во оваа семинарска работа ја анализирам глобалната економска состојба преку бруто домашниот производ по жител (GDP per capita) на државите во светот. БДП на една земја се пресметува како цената на сите продукти и сервиси произведени или остварени во рамките на една земја во период од една година. БДП по жител претставува економски индикатор кој се пресметува така што вкупниот годишен БДП на една земја се дели со бројот на жители. На овој начин се добива директна претстава за животниот стандард на човекот во таа земја.

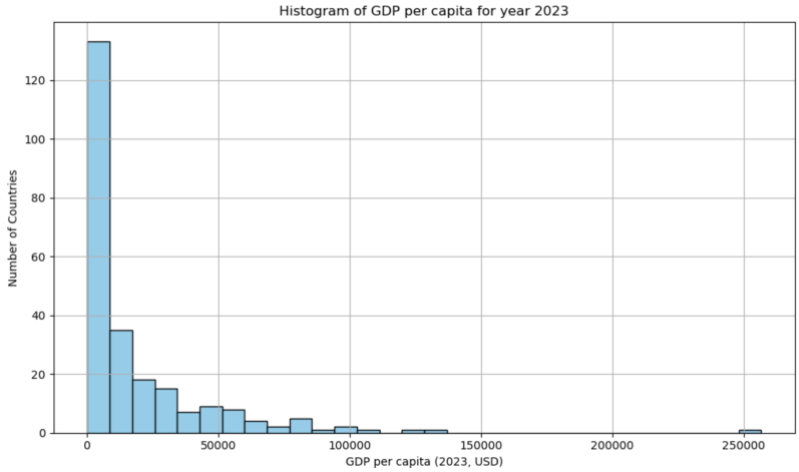
Податочното множество што се користи во оваа анализа не е претходно процесирано од моја страна. Во рамките на семинарската работа се обработени неколку важни точки:

- **Лоренцова крива и Гини коефициент:**
  Со помош на Лоренцовата крива и Гини коефициентот е анализирана економската нееднаквост во распределбата на БДП по жител меѓу земјите.
- **Hill estimator:**
  Понатаму, со примена на Hill estimator-от се анализира „тежината“ на опашката на распределбата, односно колку екстремно се издвојуваат најбогатите земји во однос на останатите богати држави.
- **Chow тест:**
  За САД е направена анализа на БДП по жител со примена на локален Chow тест, кој открива години кога се појавиле значајни структурни промени во трендот или нивото. Ова овозможува откривање на локални економски шокови и рецесии, како што се оние околу 1991 (Gulf војната), 2008 (светска финансиска криза) и 2020 (COVID-19).
- **Generalized Extreme Value (GEV) анализа:**
  На крај, е применета теоријата на екстремни вредности за да се моделираат максималните и минималните стапки на раст на БДП по жител. Со тоа се пресметани веројатности за тоа колку е веројатно во еден двегодишен период максималниот или минималниот раст да надмине или падне под одредено ниво.

Општи визуелизации

Најпрвин прикажан е хистограм на бруто домашниот производ по жител (GDP per capita) за сите земји во светот за 2023 година. На овој хистограм, на х-оската се прикажани директните вредности на БДП по жител, додека на у-оската се прикажува бројот на земји кои се наоѓаат во одреден интервал. Овој график јасно покажува дека повеќето земји имаат релативно ниски вредности на БДП по жител, додека мал број

земји имаат исклучително високи вредности, што ја развлекува распределбата надесно. Всушност, поради постоењето на неколку исклучително богати земји, најголем дел од земјите се концентрирани во првите неколку столбови, што го прави овој хистограм тежок за читање кога би сакале да ги согледаме релативните разлики помеѓу „средно“ и „ниско“ развиените земји.

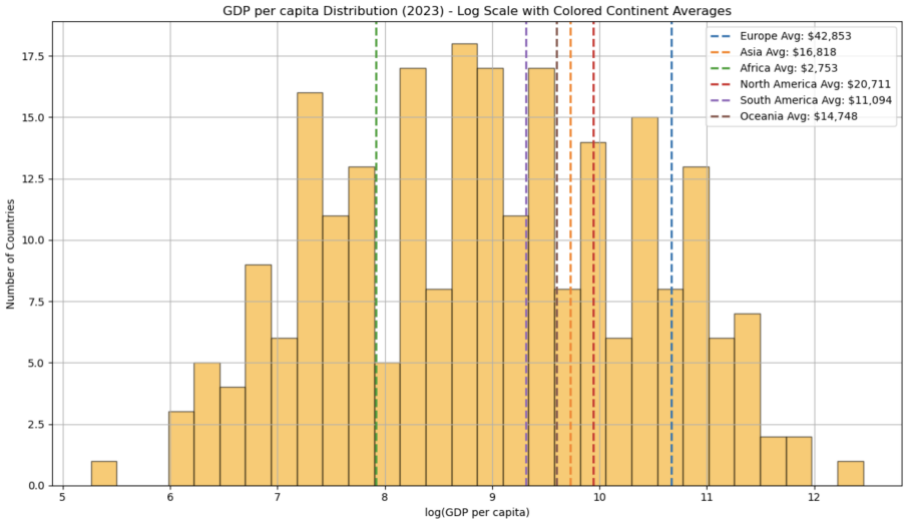


За подобро да се прикаже формата на распределбата и да се избегне доминацијата на неколкуте екстремно богати земји, податоците се логаритмирани (со природен логаротам). Логаритмирањето на податоците овозможува да се прикажат податоците на скала која ги намалува екстремите и ја приближува распределбата кон нормалната распределба, па подобро може да се види каде е центарот и како се распоредени земјите. Можеме да забележиме дека најголем број земји имаат вредност за БДП по жител концентрирани околу $4.000 – $6.000 USD, што кореспондира на логаритамска вредност од околу 8.5 .

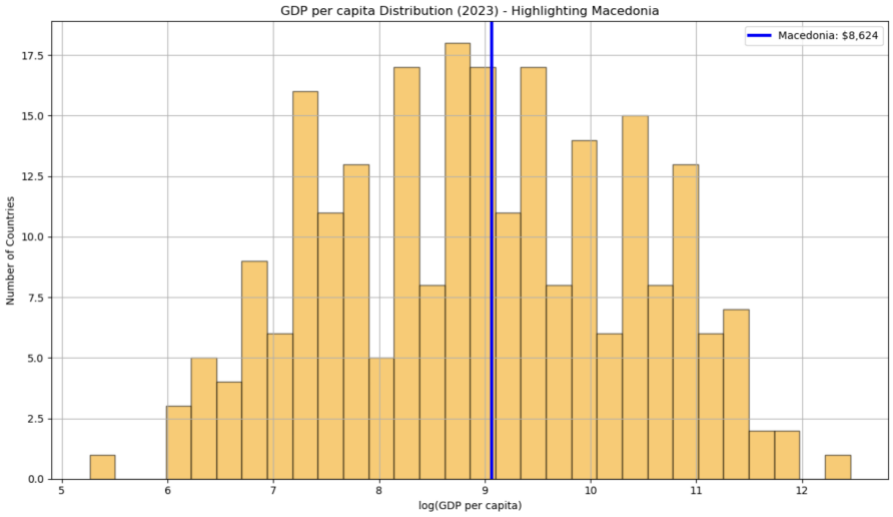
На крај, на истиот хистограм се нанесени и просечните вредности на БДП по жител за секој континент (Европа, Азија, Африка, Северна Америка, Јужна Америка и Океанија). Овие просеци се прикажани со обоени испрекинати вертикални линии. Од графикот може да се заклучи дека:

- Африка има најниска просечна вредност, што значи дека земјите од овој континент се генерално најсиромашни и животниот стандард е најлош.
- Европа има највисока просечна вредност.
- Северна Америка и Азија исто така имаат релативно високи просечни вредности, додека Океанија и Јужна Америка се наоѓаат околу средината.

На овој начин, со помош на визуелизацијата може да се согледа глобалната економска нерамнотежа не само помеѓу индивидуалните земји, туку и помеѓу целите континенти.



На графикот подолу е нанесена вредноста на БДП по жител за Македонија.



Лоренцова крива и гини кеофициент

Лоренцовата крива е графичка алатка која се користи за да се прикаже нееднаквоста во распределбата на некоја економска варијабла. Во нашиот случај, Лоренцовата крива се користи за да се прикаже колку е нееднакво распределен БДП по жител помеѓу државите во светот.

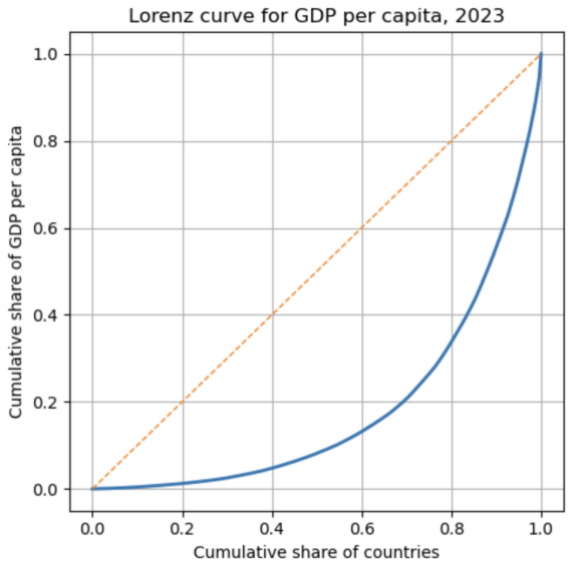
На графикот:

- x-оската претставува кумулативен процент на земјите, подредени од најмала вредност кон најголема вредност на БДП по жител.
- y-оската претставува кумулативен процент на светскиот БДП по жител што го сочинуваат земјите од x-оската.

Ако сите земји имаа ист GDP per capita, Лоренцовата крива ќе биде права линија од (0,0) до (1,1). На пример тоа значи доколку земеме 20% од земјите во светот, сумата на

нивните вредности за БДП по жител ќе има удел точно 20% во сумата на сите држави во светот.

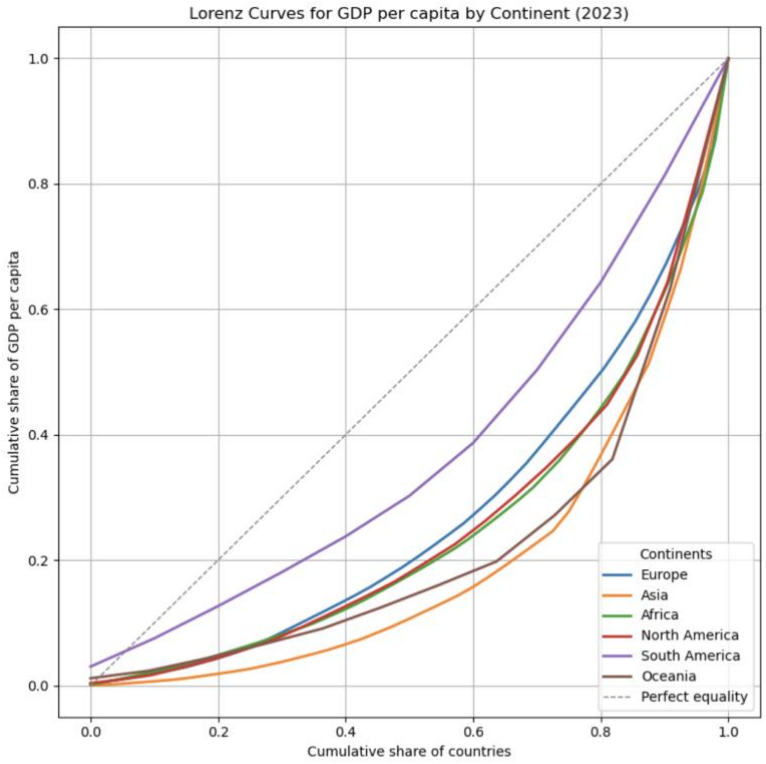
Гини коефициентот се пресметува како: површината на линијата на совржена еднаквост (правата од точката (0,0) до (1,1)) и лоренцовата крива, врз вкупната површина под линијата на совршена еднаквост. Од Лоренцовата крива за 2023 година може да се види дека добиената крива е значително свиткана под линијата на перфектна еднаквост, што покажува дека голем дел од светскиот БДП по жител е концентриран во релативно мал број земји. Поточно околу 80% од земјите придонесуваат со помалку од 40% од глобалниот БДП по жител.



На графикот подолу се прикажани Лоренцовите криви за БДП по жител за годината 2023, поделени по континенти. Лоренцовата крива визуелно го прикажува степенот на економска нееднаквост меѓу државите на секој континент. Од графикот може да се забележат неколку значајни заклучоци:

- Азија има најизразена нееднаквост, со Гини коефициент од 0.594, што укажува дека мал број земји (како Сингапур, Катар, Јапонија) држат голем дел од вкупниот БДП по жител на континентот, додека мнозинството земји имаат значително пониски вредности.
- Океанија има Гини од 0.550, што укажува на доста голема нееднаквост, пред сè поради постоењето на развиени држави како Австралија и Нов Зеланд и малите островски држави со многу понизок БДП по жител.
- Африка исто така покажува релативно висока нееднаквост со Гини коефициент од 0.495, што е очекувано со оглед на економската разлика помеѓу државите на тој континент.
- Северна Америка има Гини од 0.485, што покажува дека САД и Канада со својата висока економска сила значително влијаат врз нееднаквоста на континентот.
- Европа има Гини коефициент од 0.451, што укажува на умерена нееднаквост. Ова е резултат на тоа што Европа има поголем број развиени држави со слично висок БДП по жител, и има понеразвиени држави во источна Европа.

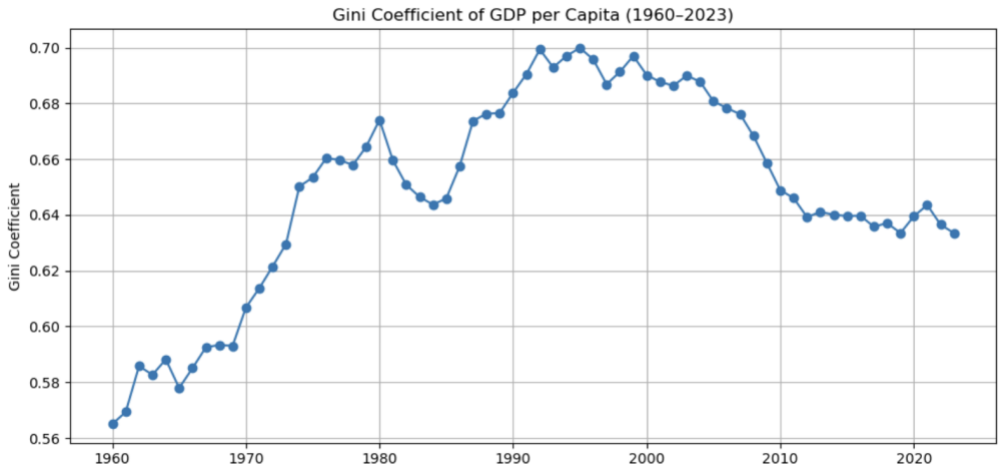
- Најинтересен резултат дава Јужна Америка, со најнизок Гини коефициент од 0.243, што значи дека БДП по жител е релативно рамномерно распределен меѓу државите на овој континент. Тоа не значи дека Јужна Америка е богата, туку животниот стандард е сличен на луѓето во сите држави на тој континент.




```
Europe Gini: 0.451
Asia Gini: 0.594
Africa Gini: 0.495
North America Gini: 0.485
South America Gini: 0.243
Oceania Gini: 0.550
```

На овој график е прикажано движењето на Гини коефициентот за БДП по жител во светот во периодот од 1960 до 2023 година. Од графикот може да се забележи дека:

- Во периодот 1960–1980, Гини коефициентот растел од околу 0.56 на над 0.67, што укажува на тоа дека во тие децении економската нееднаквост меѓу државите значително се зголемила.
- Во периодот од 1980 до крајот на 90-тите, се одржува на релативно високо ниво со мал пад во почетокот на 80-тите, па потоа дури достигнувајќи локални врвови од 0.70 околу 1995 година, што го покажува највисокото ниво на глобална економска во историјата.
- По 2000 година, трендот се менува и коефициентот започнува постепено да опаѓа. Ова укажува на полека намалување на глобалната економска нееднаквост меѓу државите во последните две децении.



Hill estimator

За да се анализира глобалната економска нееднаквост, користено е Гини коефициент, кој дава слика за тоа колку општо се нееднакво распределени БДП по жител вредностите меѓу сите држави во светот. Сепак, тој коефициент не ја зема во предвид формата на распределбата во нејзините екстреми, односно не покажува дали нееднаквоста потекнува од горната опашка, од средината или од долниот дел на распределбата. Поради тоа, како дополнителна анализа се користи Hill estimator-от, кој ја оценува „тежината“ на опашката на распределбата, односно опишува колку се нееднакви меѓусебно само најбогатите држави(кои се наоѓаат во опашката). На овој начин добиваме појасна слика за тоа дали постои екстремна концентрација на богатството кај мал број држави со исклучително висок БДП по жител.

Математички параметарот кси (ξ) се дефинира како просек на логаритамските разлики на најголемите k вредности од распределбата во однос на (k+1) - та најголема вредност. Параметарот алфа се дефинира како еден врз параметарот кси. Формулата изгледа вака:

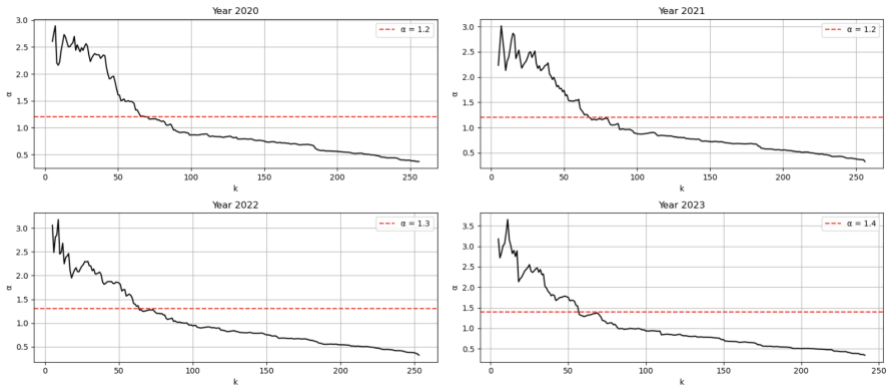
$$\hat{\alpha}^{(H)} = \hat{\alpha}_{k,n}^{(H)} = \frac{1}{\hat{\xi}_{k,n}^{(H)}} = \left(\frac{1}{k} \sum_{j=1}^{k} \log(X_{j,n}) - \log(X_{k,n}) \right)^{-1}$$

Ова значи дека Hill estimator-от ја пресметува просечната „логаритамска оддалеченост“ на најбогатите држави од некој праг, што овозможува да се оцени колку екстремно се одделуваат најбогатите држави од тој праг. Колку оваа просечна логаритамска разлика е поголема, толку кси параметароте поголем, а алфа вредноста е помала, што означува потешка опашка.

Параметарот алфа не ја опишува ширината или дебелината на опашката на визуелно на хистограмот, туку ја мери брзината со која опаѓа веројатноста да се појават се поголеми вредности, кога веќе се наоѓаме во опашката на распределбата. Поедноставено кажано, мало алфа значи дека кога ќе влеземе во опашката, веројатноста да наидеме на уште поголеми вредности опаѓа многу бавно, што укажува на тешка опашка (екстремните вредности се релативно чести). Од друга страна, големо алфа значи дека веднаш штом влеземе во опашката, веројатноста да се појават уште

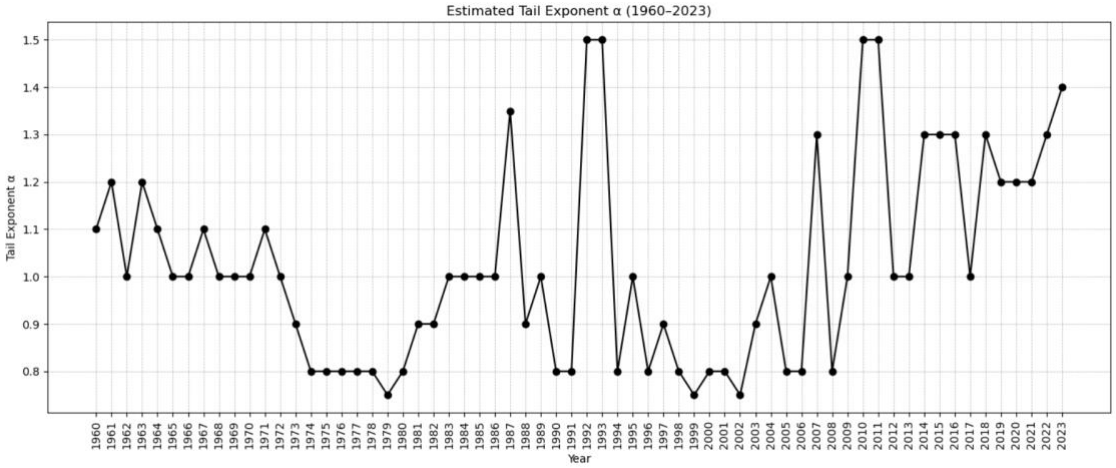
поголеми вредности брзо опаѓа, односно дека опашката е лесна и нема значителни екстреми. Во однос на податоците мало алфа означува дека неколку држави имаат значително поголем БДП по жител во опашката, што упатува на голема економска нееднаквост меѓу најбогатите држави. Од друга страна, поголемо алфа значи дека дури и најбогатите држави не се толку далеку од останатите богати држави.

За да се анализира однесувањето на опашката на распределбата на БДП по жител низ времето, пресметана е вредноста на параметарот алфа за секоја година во периодот од 1960 до 2023. За секоја година е конструиран Hill график (Hill plot), на кој на хоризонталната оска е прикажан бројот на најбогатите k држави што се користат во пресметката на алфа параметарот, а на вертикалната оска е прикажана вредноста на параметарот. Со црвена испрекината линија на графиците е означено визуелно определената вредност за алфа параметарот за одредена година.



На следниот график прикажани се проценетите вредности на параметарот алфа за секоја година во периодот од 1960 до 2023, добиена со Hill estimator-от. Од графикот се гледа дека во 1970-тите и почетокот на 1980-тите, алфа има најниски вредности (околу 0.8), што укажува на тоа дека тогаш најбогатите држави биле значително побогати од останатите во опашката, со многу поголем БДП по жител. Во 1990-тите и 2000-тите, се забележува нестабилно движење со периодични „скокови“ на алфа до околу 1.5, што сугерира дека во одредени години концентрираноста на богатството била нешто помала. Во последната декада (2010–2023) се гледа умерен раст на алфа, со вредности околу 1.2 до 1.4, што значи дека опашката и понатаму останува тешка, но сепак нееднаквоста во горниот дел на распределбата е малку помала во споредба со најекстремните години од минатото.

Ова покажува дека, генерално, распределбата на БДП по жител низ државите константно има тешка опашка, што значи дека и денес постои група држави кои се значително побогати од останатите богати држави во опашката. Сепак, благо зголемување на алфа во последните години упатува на тоа дека концентрацијата на екстремно високи БДП по жител вредности е за нијанса намалена.



Chow test

Chow тестот претставува класичен статистички метод кој се користи за да се провери дали во некоја временска серија постои структурна промена. На пример, во анализа на БДП по жител низ годините, Chow тестот може да помогне да се открие дали во некоја година се појавил различен тренд на раст или пад. Основната идеја на тестот е да се споредат два модели:

- Еден модел кој ги опишува сите податоци со една единствена линеарна регресија.
- Модел со две посебни линеарни регресии. Една регресија за периодот пред моментот за кој проверуваме дали има некоја структурна промена и една за податоците после тој момент.

Chow тестот го гледа намалувањето на квадратната грешка (RSS) при користење на две линеарни регресии, во споредба со тоа кога се користи една линеарна регресија. Ако користењето на две линеарни регресии доведе до значајно намалување на грешката, тогаш се добива голема F-статистика и мала p-вредност, што значи дека треба да се отфрли нултата хипотеза (дека нема структурна промена) и да се прифати дека постои структурна промена. Во спротивно, ако p-вредноста е голема, се задржува нултата хипотеза.

Поефикасен се покажа локалниот Chow тест, наместо класичниот Chow тест. Со локалниот Chow тест се анализираат само неколку години пред и по секој потенцијален момент на структурна промена (4 години пред и по), со што се овозможува да се детектираат локални нагли промени во трендот или нивото на БДП по жител. Забележува локални структурни промени, што класичниот Chow тест не би можел да ги забележи. БДП по жител податоците имаат изразен долгорочен растечки тренд, кој ако се подели на два долги периода, класичниот Chow тест тешко би открил структурни промени.

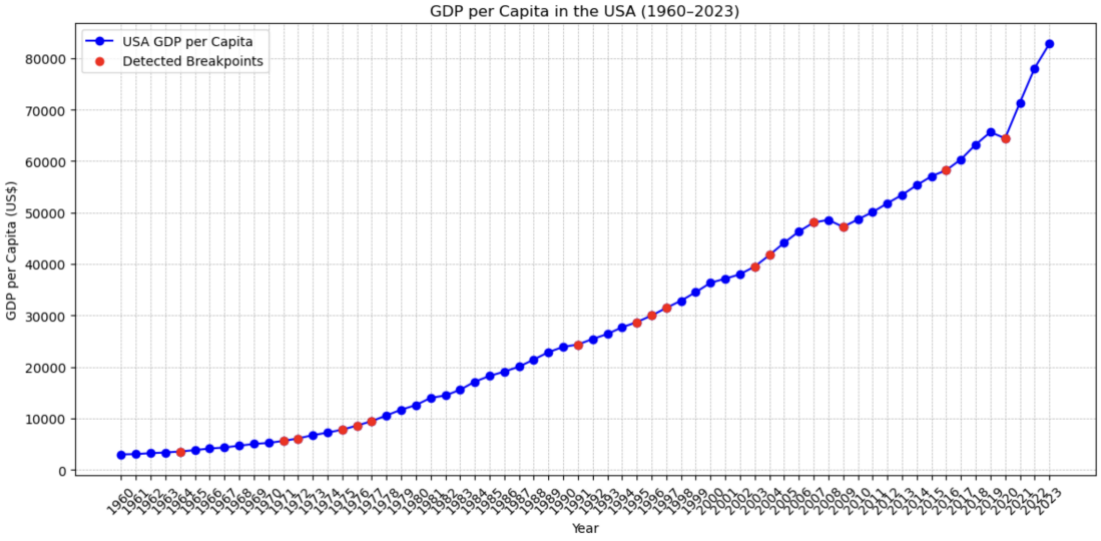
Во тестот се користи следната формула за Ф статистиката:

$$CHOW = \frac{(RSS_p - (RSS_1 + RSS_2))/k}{(RSS_1 + RSS_2)/(N_1 + N_2 - 2k)},$$

- Значењето на записот во броителот е колку во просек се променува подобрувањето во фитување на една регресијата во споредба со две регресии, во однос на бројот на параметри во регресијата(во нашиот случај к=2).
- Именителот е стандардната формула за пресметување на варијанса на резидуалите кај линеарна регресија, само тука за две линеарни регресии.

Кога ја пресметуваме целата статистика всушност се проверува дали подобрувањето што се добива ако се поделат податоците во два сегмента е доволно големо во однос на тоа колку се расфрлани податоците? Па Ако F-статистиката е голема, тоа значи дека намалувањето на грешките при фитување на две линеарни регресии е значително поголемо од варијансата на податоците, па може да се заклучи дека постои структурна промена. Ако F-статистиката е мала, тоа значи дека подобрувањето не е доволно значајно споредено со варијанса, па нема доказ за структурена промена.

Иако на линискиот дијаграм подолу со БДП по жител за САД во периодот 1960–2023 визуелно не се забележува некој јасен структурен прекин во некои од годините детектирани со Chow тестот, сепак математичката анализа со овој тест открива дека таму постои значајна промена. Тоа е токму поради тоа што F-статистиката ја зема во предвид варијансата на податоците околу фитуваните линии. Во делови од податоците каде варијансата е многу мала, дури и релативно мали отстапувања се доволни за да се добие висока F-статистика и да се заклучи дека има структурен прекин.



Ако се споредат овие години со познати економски настани, може да се види дека:

- Во 1991 година е детектирана промена што се совпаѓа со војната во таа година (Gulf war),
- Во 1995–1997 година се забележува промена што може да се поврзе со забрзаната експанзија на технолошкиот сектор,

- Во 2007, 2008 и 2009 година се потврдени структурни промени како резултат на светската финансиска криза,
- Додека во 2020 година јасно се регистрира влијанието од пандемијата од COVID-19.

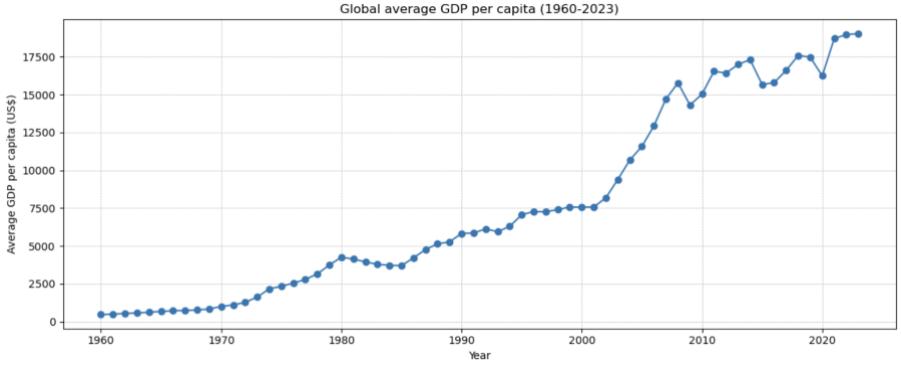
Исто така има и други години со значајни локални промени кои не секогаш се поклопуваат со глобално познати кризи, што е очекувано бидејќи БДП по жител е чувствителен на многу економски фактори - како глобална трговија, цени на суровини, локални политики итн.

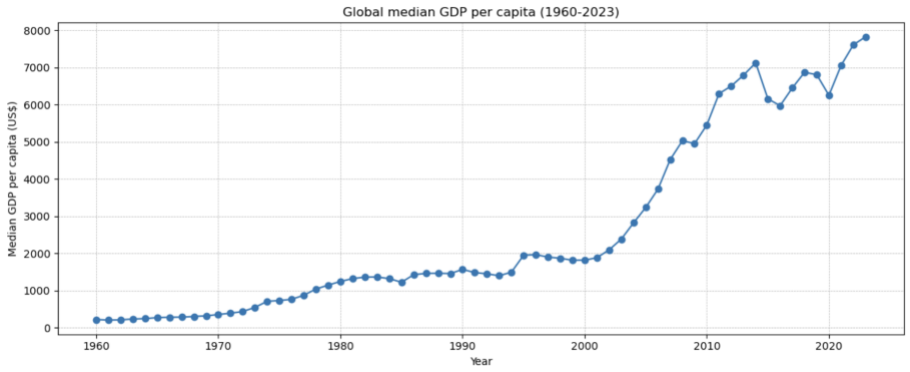
Генерализирана распределба на екстремни вредности

Во овој дел од анализата фокусот е ставен на екстремните вредности во распределбата на GDP по жител. Додека претходните методи како Гини коефициентот, Лоренцовата крива и Hill estimator-от додоа увид во општата нееднаквост и „тежината“ на опашката на распределбата, теоријата на екстремни вредности (Extreme Value Theory, EVT) се занимава специјално со моделирање на однесувањето на најголемите (или најмалите) набљудувани вредности. EVT овозможува статистички да се опишат и квантифицираат веројатностите за појава на невообичаено големи или мали вредности.

GEV (Generalized Extreme Value) распределбата произлегува од теоремата на Fisher–Tippett–Gnedenko, која се смета за еквивалент на Централната Гранична Теорема, но за максимуми (или минимуми). Оваа теорема вели дека ако ги разгледуваме максималните или минималните вредности во доволно големи блокови од некоја основна распределба, тогаш по соодветно нормализирање (центрирање и скалирање) нивната распределба ќе конвергира кон една од трите форми: Gumbel, Fréchet или Weibull. GEV распределбата ги обединува овие три случаи во едно унифицирано семејство, дефинирано преку параметарот за кси (алфа = 1/кси).

На почетокот фитувана е GEV распределбата на блок максимумите на БДП по жител вредностите. Сепак, резултатите од ова фитување испаднаа многу лоши. Главната причина е тоа што БДП по жител има постојан раст со текот на времето. Со други зборови, вредности кои во 1980 година биле сметани за екстремни, денес се сосема нормални нивоа на БДП по жител вредност. Ова ја нарушува основната претпоставка на EVT дека податоците се генерирани од стационарна распределба, односно распределба која не се менува со текот на времето. Па наместо тоа, се фитува GEV распределбата на годишните стапки на раст на БДП по жител. Со анализа на екстремните стапки може да се добијат поинформативни сознанија за необичен економски раст или пад.



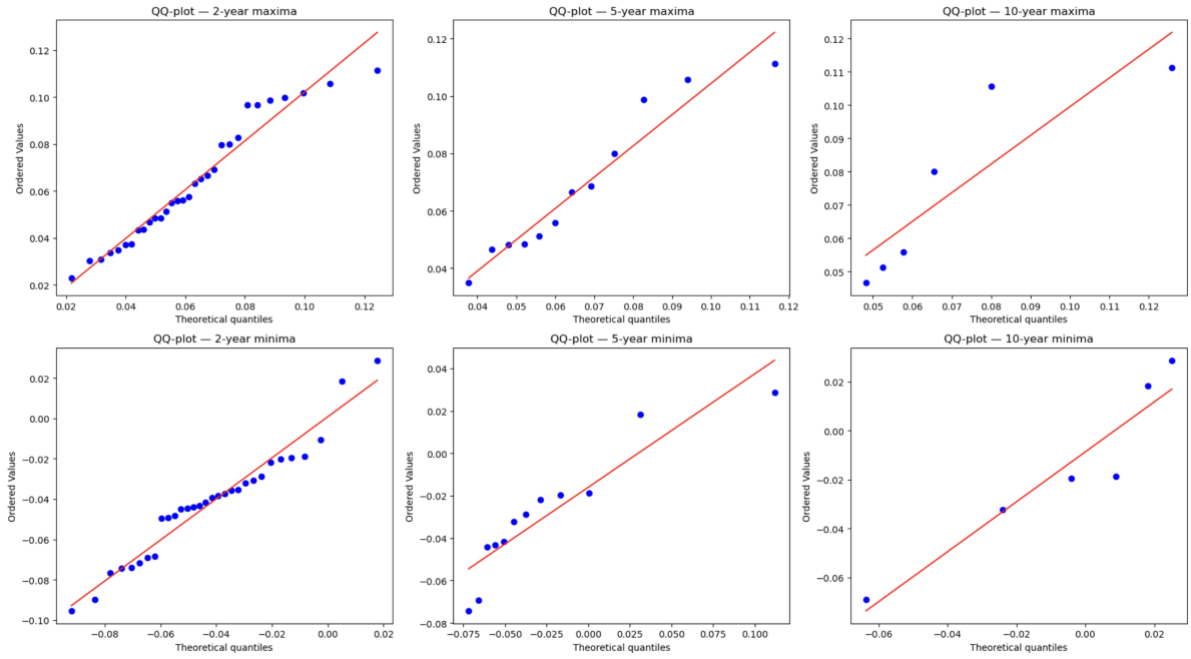


QQ плот (Quantile-Quantile plot) е графички метод кој служи за да се спореди распределбата на податоците со некоја теоретска распределба. Со други зборови, QQ plot ни покажува дали нашите податоци потекнуваат од (или барем личат на) распределбата која сме ја претпоставиле. На x оската се претставени теоретските квантили на дистрибуцијата, а на y оската се претставени емпириските квантили од податоците. Ако точките лежат на права линија, тоа значи дека нашите податоци многу добро се совпаѓаат со теоретската распределба.

Во следната анализа пресметани се стапките на раст на БДП по жител за САД:

1961	1.98%	1976	10.14%	1993	3.81%	2009	-2.83%
1962	5.78%	1977	10.01%	1994	4.96%	2010	3.08%
1963	4.03%	1978	11.77%	1995	3.60%	2011	2.91%
1964	5.91%	1979	10.50%	1996	4.45%	2012	3.43%
1965	7.10%	1980	7.71%	1997	4.98%	2013	3.14%
1966	8.33%	1981	11.14%	1998	4.43%	2014	3.55%
1967	4.59%	1982	3.27%	1999	5.06%	2015	3.14%
1968	8.29%	1983	7.69%	2000	5.26%	2016	2.04%
1969	7.16%	1984	10.15%	2001	2.21%	2017	3.63%
1970	4.27%	1985	6.52%	2002	2.33%	2018	4.77%
1971	7.17%	1986	4.58%	2003	3.93%	2019	3.80%
1972	8.64%	1987	5.07%	2004	5.66%	2020	-1.82%
1973	10.38%	1988	6.88%	2005	5.75%	2021	10.72%
1974	7.42%	1989	6.72%	2006	4.94%	2022	9.42%
1975	7.97%	1990	4.51%	2007	3.78%	2023	6.07%
		1991	1.90%	2008	1.08%		
		1992	4.42%				

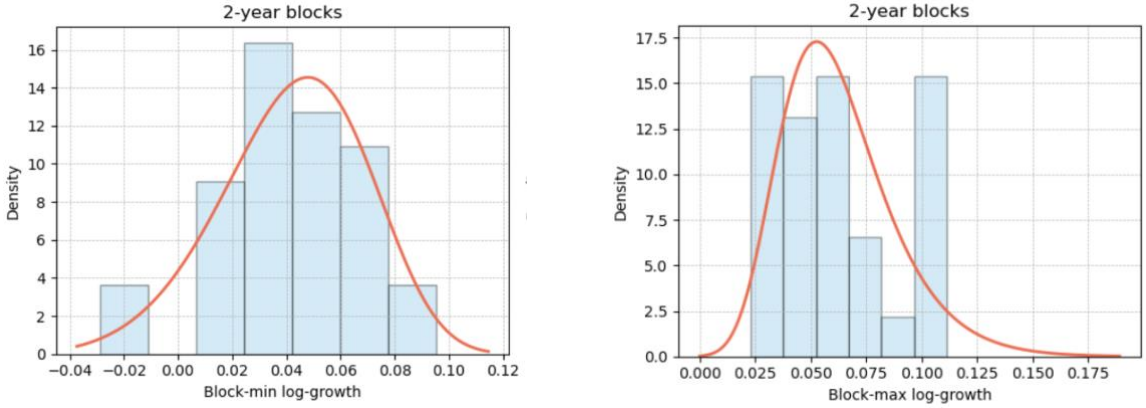
Потоа поделени се податоците на блокови и од секој блок се земени максималната и минималната стапка на раст. На тој начин добиени се податоци врз кои може да се фитува GEV дистрибуција, за да се моделираат екстремните вредности на растот и падот. Испробав различни големини на блокови, од 2 години, од 5 години и од 10 години. Од резултатите се гледа дека и во двете сценарија дистрибуцијата добиена со 2-годишен блок најдобро одговара за нашите податоци.



Гледајќи го графикот подолу, може да се заклучи дека фитуваната распределба за минималните вредности по блоковите е одлична, додека за максималните не е совршена, но е задоволителна. Хистограмите ја прикажуваат емпириската распределба на максималните/минималните растови по блокови од 2 години, односно реалните податоци. Црвената крива ја прикажува функцијата на густина на фитуваната распределба, добиена со пресметаните параметри.

Параметрите за дистрибуцијата за блок минима се: $\xi=-0.181$, $\mu=0.0530$, $\sigma=0.0257$

Параметрите за дистрибуцијата за блок максима се: $\xi=-0.057$, $\mu=0.0514$, $\sigma=0.0213$



Некои прашања кои можеме да се одговорат откако се фитувани дистрибуциите:

1. Која е веројатноста дека максимумот на 2-годишен интервал ќе надмине одредена вредност?

	Procent na rast	Verojatnost deka procentot na rast ke ja nadmine taa vrednost vo 2-godisen blok
0	4	0.8275
1	6	0.5150
2	8	0.2513
3	10	0.1061
4	12	0.0411
5	14	0.0150
6	16	0.0052
7	18	0.0017

2. Која е веројатноста дека минимумот на 2-годичен интервал ќе биде помал од одредена вредност?

	Procent na rast	Verojatnost deka procentot na rast ke bide pomal od taa vrednost vo 2-godisen blok
0	-1	0.0386
1	0	0.0733
2	4	0.4348
3	6	0.7062
4	8	0.9063
5	10	0.9853
6	12	0.9991

Дополнителна доработка

Следни чекори во доработка на проектот:

- Анализирање на земји аутлаери, земји со необичен БДП по жител патерн. Би користел Isolation forest, Local outlier factor, Mahalanobi растојание.
- Предвидување со регресија или со временски серии (Facebook Prophet, ARIMA, LSTM)
- Кластерирање на земјите според параметри како БДП по жител, раст во изминатите години и така натаму.
- Transfer учење – може да се кластерираат земјите според поранешен БДП по жител раст, па тие сличности да се искористат за да се направи предвидување на помалку развиени земји