Facultatea de Cibernetică, Statistică și Informatică Economică

PROIECT SERII DE TIMP

Evoluția Companiei NVIDIA la Bursa de Valori
Analiza Sistemului Educațional Românesc

Profesor coordonator: Adriana Davidescu (Alexandru)

JILAVU Alexandru

LIȘIȚĂ Corina

MEDELEANU Raluca - Andra

Grupa 1078, seria A

CUPRINS

APLICAȚIA I

Introducere

Această lucrare își propune să analizeze evoluția companiei NVIDIA într-una dintre cele mai influente perioade de la listarea sa la Bursa de Valori, concentrându-se pe creșterea substanțială a indicelui NVDA, care a fost în mare parte alimentată de adaptarea gigantului tehnologic la trendul dominant din ultimii ani, anume inteligența artificială. Întrebarea principală pe care dorim să o abordăm este dacă o evoluție explozivă similară ar fi putut fi prognozată în principal cu ajutorul unor analize bine cunoscute în domeniul analizei seriilor de timp.

În cadrul studiului de caz asociat lucrării ne vom concentra pe analiza datelor lunare ale acțiunilor NVDA în perioada ianuarie 2017- aprilie 2024. Vom aplica metode și teste specifice pentru ainvestiga caracteristicile cheie ale acestor date, inclusiv:

- *Verificarea staționarității;*
- *Analiza sezonalității;*
- *Aplicarea modelului ARIMA;*

În plus, vom consulta articole și lucrări de specialitate care explorează deciziile financiare și logistice ale companiei NVIDIA și modalitățile de analiză și prognozare a seturilor de date similare în domeniul tehnologiei și al piețelor financiare. Prin integrarea acestor aspecte, ne propunem să furnizăm o analiză vastă și o evaluare a potențialului de prognozare a evoluției indicelui NVDA în contextul trendurilor tehnologice și financiare actuale.

Literatură de specialitate

Un articol care iese în evidență în discuția legată de posibilitatea de prognozare a prețurilor acțiunilor este *„Exploiting Data Science for Measuring the Performance of Technology Stocks”*, lucrare care ia ca reper un set de 4280 de date financiare pentru fiecare din companiile Apple (AAPL), Microsoft Corporation (MSFT), Taiwan Semiconductor Manufacturing Company Limited (TSM), NVIDIA Corporation (NVDA) și Broadcom Inc. (AVGO).

Perioada considerată în acest studiu de caz este 2005-2022, pe care s-a aplicat o varietate de metode de analiză de date precum Random Forest Classifier (RF), XGBoost Classifier (XG), Bagging Classifier (BC), Auto-Regressive Integrated Moving Average (ARIMA) și altele, pentru a determina care dintre acestea poate produce rezultate consistente pentru simularea unei prognoze pe un set de date de test.

Motivația alegerii unei mulțimi vaste de modalități de analiză din spața machine learning sau deep learning se bazează pe posibilitatea acestora de a determina anumite elemente cheie care să fie mai apoi folosite în structurarea unei prognoze. (Citat din articol: *„Machine learning (ML) algorithms, LSTM (deep learning model), ARIMA, and HMM (time series analysis models) of data science are used to infer the mechanics behind the data efficiently. Often, an observer interprets the exact information from the data, but machine learning algorithms can do it. ”*).

Mai jos se observă o figură extrasă ca rezultat al analizei dezvoltate în articolul discutat, unde se poate extrage concluzia că pe departe cele mai consistente metode de prognozare a acestor indici bursieri sunt LSTM, ARIMA și HMM, cu o diferență semnificativă față de metodele concurente.

Acest studiu evidențiază importanța unui model de analiză stabil în vederea determinării unei prognoze financiare pe o durată considerabilă de timp, un motiv în plus pentru care folosirea metodei ARIMA în determinarea cursului indicelui NVDA este o alegere eficientă.

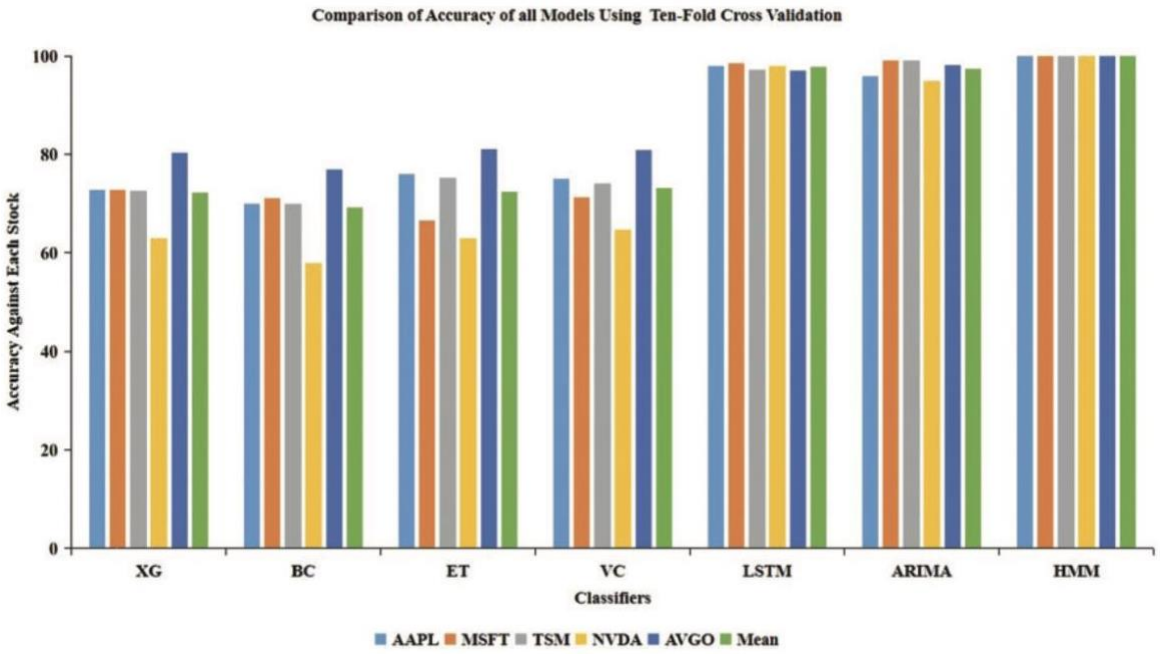


Figura 1: Compararea Acurateței Tuturor Modelelor utilizând Validarea Încrucișată în Zece Etape (extrasă din „Exploiting Data Science for Measuring the Performance of Technology Stocks")

Un alt studiu competent în vederea abordării modalităților de manipulare a datelor indicilor acțiunilor la Bursa de Valori este cel realizat de analiștii de la Universitatea Maharishi de Tehnologia Informației din India. Articolul se bazează pe parcurgerea unui set de date zilnice pentru indicele NVDA pentru anul 2022 prin folosirea limbajului de programare Python, implicit a bibliotecii „Autots", pentru a verifica performanța modalităților de prognozare folosind atât algoritmi de machine learning, cât și de deep learning.

În studiu se atestă faptul că Nvidia este unul din giganții din lumea tech, al cărei preț la Bursa de Valori este direct corelat de cererea produselor pe piață (Citat din studiu: *„NVIDIA Corporation is a leading chipmaker in the world and its stock price is highly influenced by the demand for its products."*). Acesta este considerat ca fiind și unul dintre motivele principale pentru care dezvoltarea unui algoritm de prognozare cu o acuratețe ridicată este un obiectiv principal în zona financiară.

Deși modelele folosite dovedesc progresele realizate în domeniu și acuratețea la care pot ajunge cu un set suficient de semnificativ, sunt evidente și dezavantajele unei astfel de prognoze, printre care se numără incapabilitatea calculării unor evenimente neașteptate (se menționează dezastre naturale, evenimente politice, războaie și recesiuni economice), disponibilitatea redusă a datelor, non-staționaritatea și „zgomotul" din date.

Ca și concluzii extrase din acest studiu, se observă o creștere semnificativă de performanță în cazul în care analiza a fost realizată folosind o tehnică de deep learning

comparativ cu zona de machine learning, însă în ambele cazuri se confirmă o direct proporționalitate între mărimea setului de date și acuratețea crescută o rezultatelor.

Studiul menționat anterior echilibrează rezultatele modelelor LSTM și ARIMA în analiza stocurilor zilnice ale companiei Nvidia. LSTM depășește cel de-al doilea model privind acuratețea predicțiilor, fiind capabil să gestioneze mai ușor natura non-staționară a acțiunilor.

Datele analizate cuprind înregistrări zilnice ale perioadei 2013-2024, surprinzând o creștere constantă începând cu anul 2013 până în a doua jumătate a anului 2021, urmată de o stagnare. Începând cu a doua jumătate a anului 2022 prețurile încep să revină la echilibru, urmând maximul atins în august 2023. Pentru prognoza acestor mișcări modelul ARIMA utilizează cele trei componente ale sale:

- *Autoregresia (AR):* dependența dintre prețul actual și valorile sale anterioare;
- *Diferențierea (I):* transformarea seriilor non-staționare în serii staționare;
- *Media mobilă (MA):* estimarea erorilor;

În continuare, se folosește criteriul informațional AIC și metoda diferențierii în scopul stabilizării curbei temporale. Modelul ARIMA (4,1,2) astfel rezultat arată o potrivire excelentă a acestuia cu datele inițiale de serie temporale.

Pentru construirea modelului LSTM fiecare porțiune temporară selectată primește intrări la momentul actual și ieșirile de la momentul anterior, utilizând formulele specifice de decizie:

- *Porțile de uitare (forget gate);*
- *Porțile de intrare (input gate);*
- *Porțile de ieșire (output gate);*

- Astfel, se realizează prognoza prețurilor care evidențiază modelul de potrivire al datelor și tendințelor de previziune pentru următoarele 10 zile.

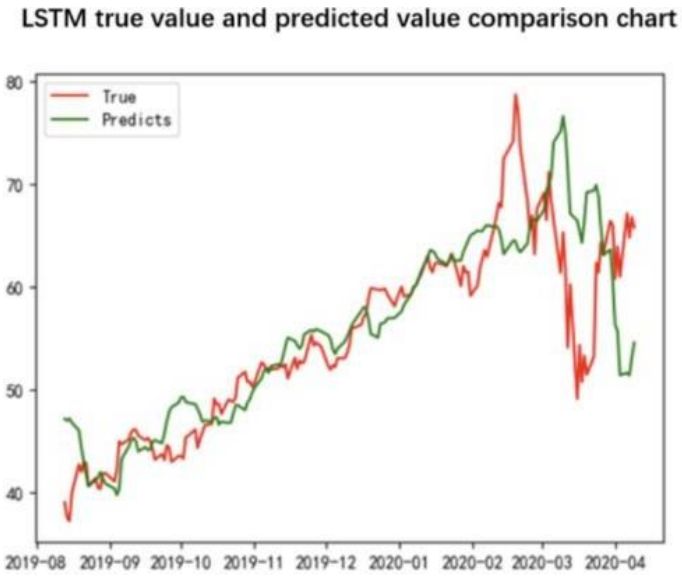


Figura 2: Predicție folosind modelul LSTM

Analiză pe Serie de Timp

Pentru a putea obține o privire de ansamblu asupra evoluției indicelui NVDA până la boom-ul de la începutul anului 2024, se va utiliza în analiză o perioadă de timp de puțin peste 7 ani, întrucât aceasta captează cel mai bine avansurile și începuturile unei creșteri sistematice a valorilor prețului. Această secvență de timp oglindește în mare parte interesul internațional asupra subiectului inteligenței artificiale, întrucât Nvidia este unul din giganții care au încurajat și fondat această trecere de-a lungul anilor, cu investiții substanțiale în crearea de noi procesoare capabile să susțină dezvoltarea de astfel de modele.

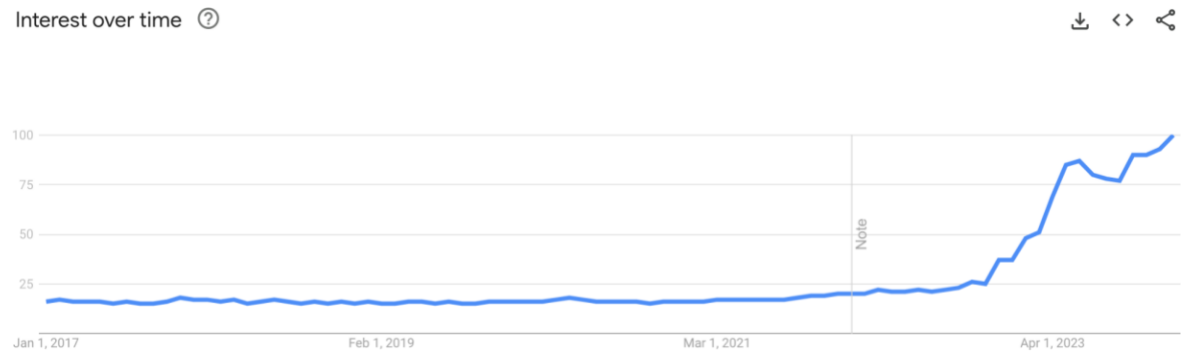


Figura 3: Interesul internațional pe subiectul „AI" pe perioada 01.01.2017 - 01.04.2024 (sursa „Google Trends")

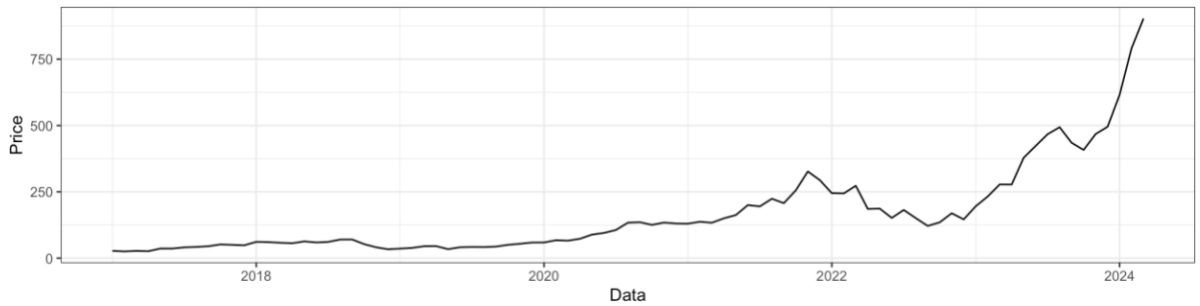


Figura 4: Graficul tip „autoplot()" atașat seriei de timp NVDA pe perioada 01.01.2017 - 01.04.2024 (sursa „R Studio")

Se observă o fluctuație similară în trendul asociat celor două statistici, mai ales pe perioada 01.04.2023 - 01.04.2024. Având această evoluție ca și referință, un punct de interes relevant de la care ar fi relevantă testarea unor modalități de prognozare este luna aprilie a anului 2023. Așadar, se va face împărțirea setului de date în antrenare și testare, luând această dată ca și limită de divizare a celor două serii de timp discutate:

```
training <- window(prices_ts, start=2017, end=c(2023,4))
test <- tail(prices_ts, 10)
```

Figura 5: Comenzi asociate împărțirii setului de date în antrenare și testare (sursa „R Studio”)

```
> training
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2017  27.2950  25.3700  27.2325  26.0750  36.0875  36.1400  40.6275  42.3600  44.6925  51.7025  50.1775  48.3750
2018  61.4500  60.5000  57.8975  56.2250  63.0475  59.2250  61.2150  70.1700  70.2550  52.7075  40.8575  33.3750
2019  35.9375  38.5650  44.8900  45.2500  33.8650  41.0575  42.1800  41.8775  43.5175  50.2550  54.1850  58.8250
2020  59.1075  67.5175  65.9000  73.0700  88.7550  94.9775 106.1475 133.7450 135.3050 125.3400 134.0150 130.5500
2021 129.8975 137.1450 133.4825 150.0950 162.4450 200.0250 194.9900 223.8500 207.1600 255.6700 326.7600 294.1100
2022 244.8600 243.8500 272.8600 185.4700 186.7200 151.5900 181.6300 150.9400 121.3900 134.9700 169.2300 146.1400
2023 195.3700 232.1600 277.7700 277.4900
```

Figura 6: Setul de date de antrenare NVDA (sursa „R Studio”)

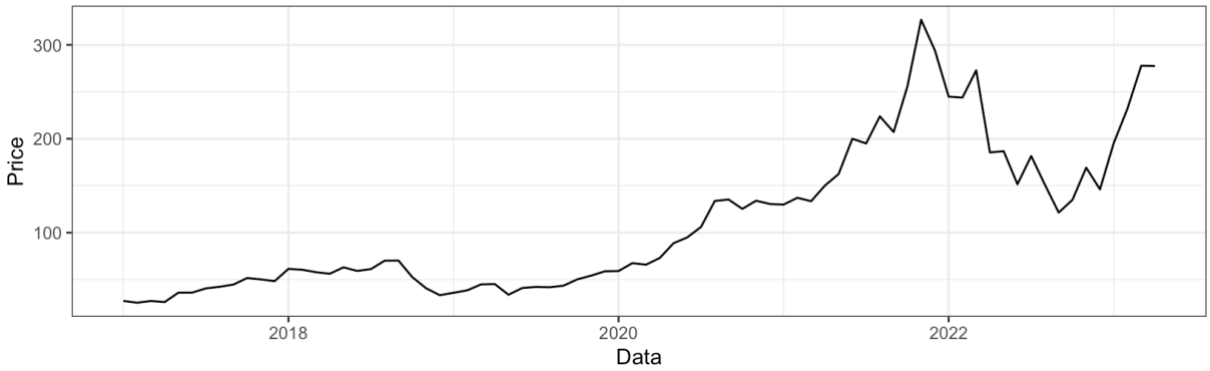


Figura 7: Evoluția setului de date de antrenare NVDA (sursa „R Studio”)

```
> test
        Jan    Feb    Mar Apr May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
2023                                 423.02 467.29 493.55 434.99 407.80 467.70 495.22
2024 615.27 791.12 903.56
```

Figura 8 : Setul de date de testare NVDA (sursa „R Studio”)

Odată ce setul de date este împărțit corespunzător obiectivelor de prognozare, vom continua prin a consulta metodologia Box-Jenkins. Primul pas conform metodei este testarea sezonalității modelului, element care va fi de folos în determinarea analizelor corespunzătoare.

Mai întâi, se va verifica rădăcina unitare sezonieră prin utilizarea testelor Hylleberg, Engle, Granger și Yoo (HEGY). Acest test implică două ipoteze fundamentale:

- *H_0 implică faptul că seria prezintă rădăcină unitară sezonieră*
- *H_1 contrariază prima ipoteză prin faptul că seria nu prezintă o astfel de rădăcină unitară sezonieră*

Prin aplicarea formulei *hegy.test(training)* obținem un p-value mult mai ridicat decât limita acceptată ($0.94 > 0.1$), așa că acceptăm ipoteza nulă, deci seria prezintă rădăcină unitate sezonieră și este nevoie să diferențiem sezonier.

Odată ce se aplică testul pe prima diferență a setului de antrenare, se obține o valoare concordanță cu ipoteza H_1, unde seria nu prezintă rădăcină unitate sezonieră.

```
        HEGY test for unit roots

data:  training

        statistic p-value
t_1        0.0247    0.94
t_2       -1.8378  0.0426 *
F_3:4      7.1175  0.0014 **
F_5:6      3.9478  0.0208 *
F_7:8      1.0312  0.3463
F_9:10      3.593  0.0288 *
F_11:12    6.9987  0.0015 **
F_2:12    48.4736       0 ***
F_1:12      44.48       0 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Deterministic terms: constant
Lag selection criterion and order: fixed, 0
P-values: based on response surface regressions
```

```
        HEGY test for unit roots

data:  diff(training)

        statistic p-value
t_1       -3.3294  0.0079 **
t_2       -1.8365  0.0425 *
F_3:4      7.3336  0.0012 **
F_5:6      3.8441   0.023 *
F_7:8      1.0932  0.3249
F_9:10     3.2206  0.0409 *
F_11:12    6.4926  0.0023 **
F_2:12     6.1857       0 ***
F_1:12      6.159   1e-04 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Deterministic terms: constant
Lag selection criterion and order: fixed, 0
P-values: based on response surface regressions
```

Figura 9: Rezultatele testului HEGY pentru setul de antrenare și prima diferență a acestuia (sursa „R Studio”)

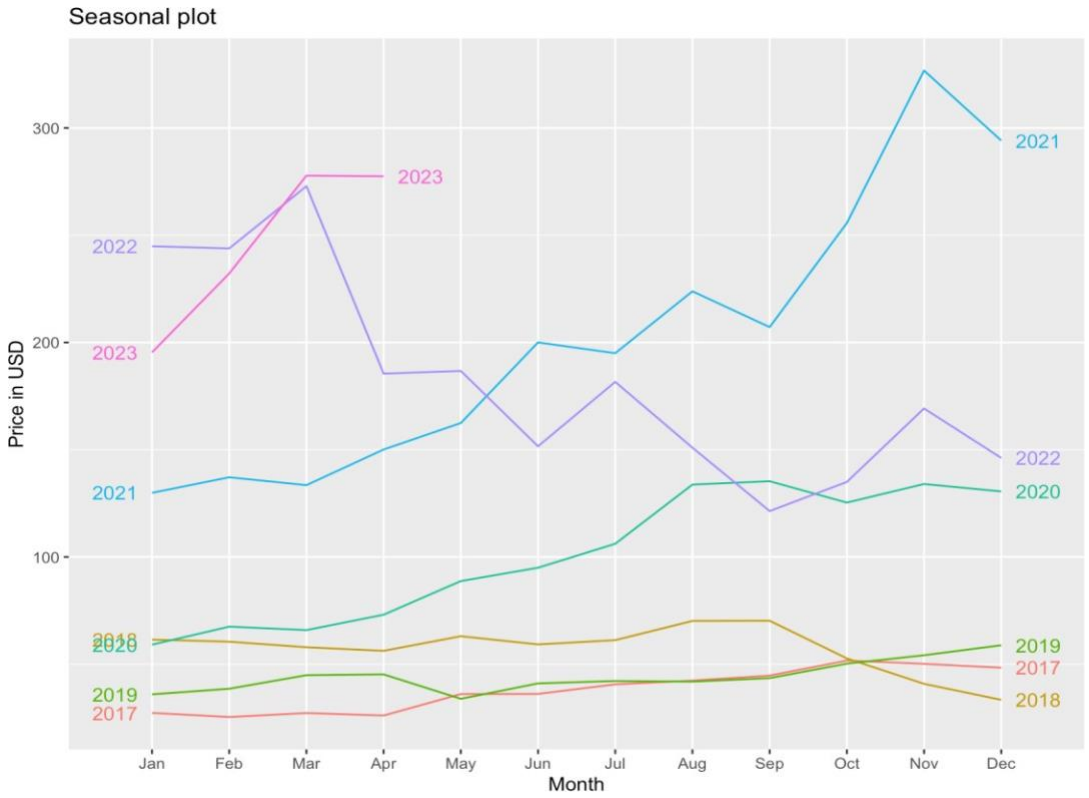


Figura 10: Evoluția anuală sezonieră a setului de antrenare (sursa „R Studio”)

Pentru a confirma existența sezonalității în setul de antrenare, vom rula formula *ggsubseriesplot(training)* pentru a observa clar modelul sezonier.

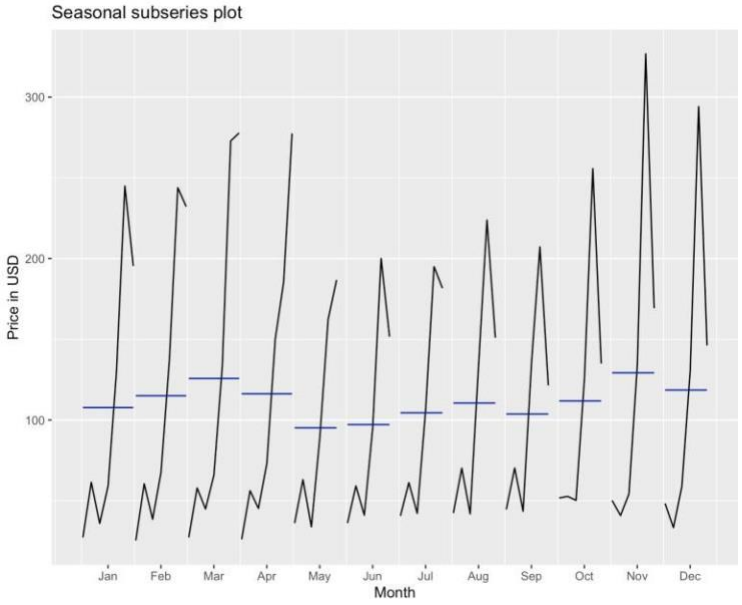


Figura 11: Componenta sezonieră structurată lunar pentru setul de antrenare (sursa „R Studio”)

Odată confirmată sezonalitatea seriei de timp, vom continua analiza prin a aborda cele două metode sezoniere principale Holt-Winters, anume HW aditiv și HW multiplicativ. Cele două vor fi apoi comparate cu evoluția reală a seriei, pentru a determina acuratețea unui astfel de model în prognozarea unei evoluții spontane de acest gen pentru indicatorul NVDA.

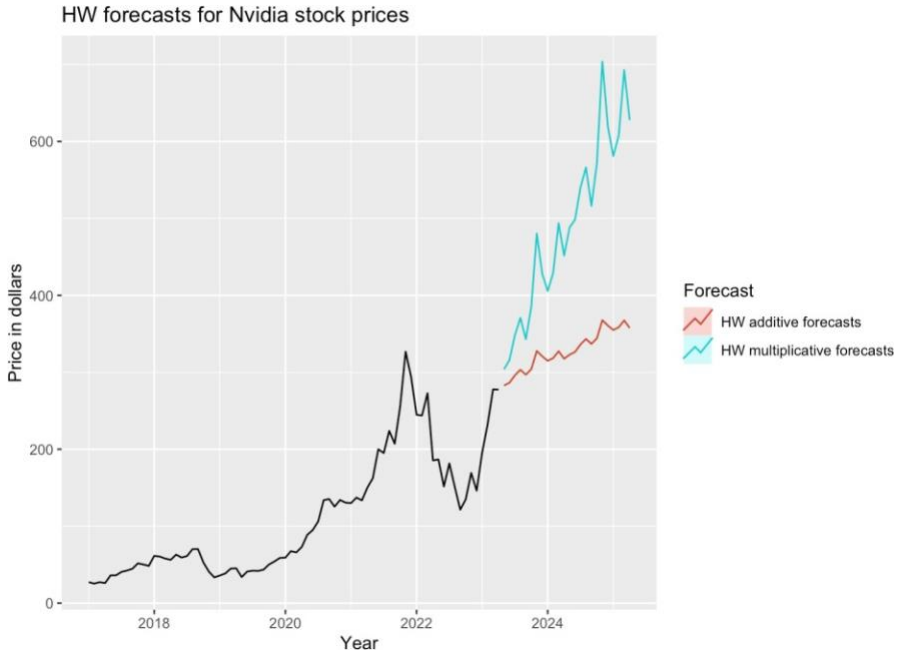


Figura 12: Prognoză suprapusă HW Aditiv & Multiplicativ (sursa „R Studio”)

Ca urmare a acestei analize, putem determina că Holt-Winters Multiplicativ a determinat cu o acuratețe semnificativă trendul dezvoltat în lunile următoare, pe când HW aditiv a confirmat existența unui trend ascendent, însă mult mai puțin accentuat. Putem confirma această diferență dintre cele două prognoze analizând indicatorii erorilor:

```
> round(accuracy(fit1),2)
                ME  RMSE   MAE   MPE  MAPE MASE ACF1
Training set -0.01 20.58 14.74 -2.86 15.62 0.26 0.05
> round(accuracy(fit2),2)
                ME  RMSE   MAE  MPE  MAPE MASE ACF1
Training set 1.02 17.81 12.33 0.08 12.42 0.22 0.14
```

Figura 13: Verificarea acurateții modelelor HW Aditiv & Multiplicativ (sursa „R Studio")

Se observă faptul că 5 din cele 7 erori considerate prezintă o valoare mai mică în modul pentru HW multiplicativ, ceea ce confirmă ipoteza că aceasta este varianta cu o acuratețe sporită în acest caz.

Mai departe, vom aborda o altă variantă de prognozare, anume modelul ETS (Exponential Smoothing with State Space). Acesta este valoros pentru prognoza seriilor de timp datorită flexibilității sale de a modela tendințe și sezonalități, abilitatea de a descompune seria de timp în componente cheie (nivel, tendință, sezonalitate), și capacitatea de a gestiona incertitudinea în prognoze. Vom folosi comanda *ets(training)* pentru a genera modelul discutat, iar cu ajutorul *summary( )* vom confirma acuratețea acestuia.

```
> summary(fit_ets) # acuratetea modelului
ETS(M,A,N)

Call:
 ets(y = training)

  Smoothing parameters:
    alpha = 0.9952
    beta  = 1e-04

  Initial states:
    l = 21.2056
    b = 2.6846

  sigma:  0.1417

     AIC     AICc      BIC
716.9905 717.8477 728.6442

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 0.692439 21.81291 13.58404 -1.634487 11.28279 0.2398809 -0.01186327
```

Figura 14: Statistici descriptive pentru modelul ETS (sursa „R Studio")

Pe lângă indicatori cunoscuți precum deviația standard mică și valorile erorilor care tind majoritar spre 0, factori care decid acuratețea sporită a modelului, putem observa o valoare pentru alpha aproximativ unitară, fapt care indică o reacție rapidă la schimbările recente în date, ceea ce poate fi benefic pentru o serie de timp volatilă ca cea discutată.

În continuare, vom folosi codul atașat pentru a genera prognoza modelului, putând mai apoi să îi testăm acuratețea.

```
fit_ets %>% forecast::forecast(h=5) %>% # prognoza modelului
  autoplot() +
  ylab("Price in dollars")
```

Figura 15: Metodă de generare a prognozei ETS (sursa „R Studio”)

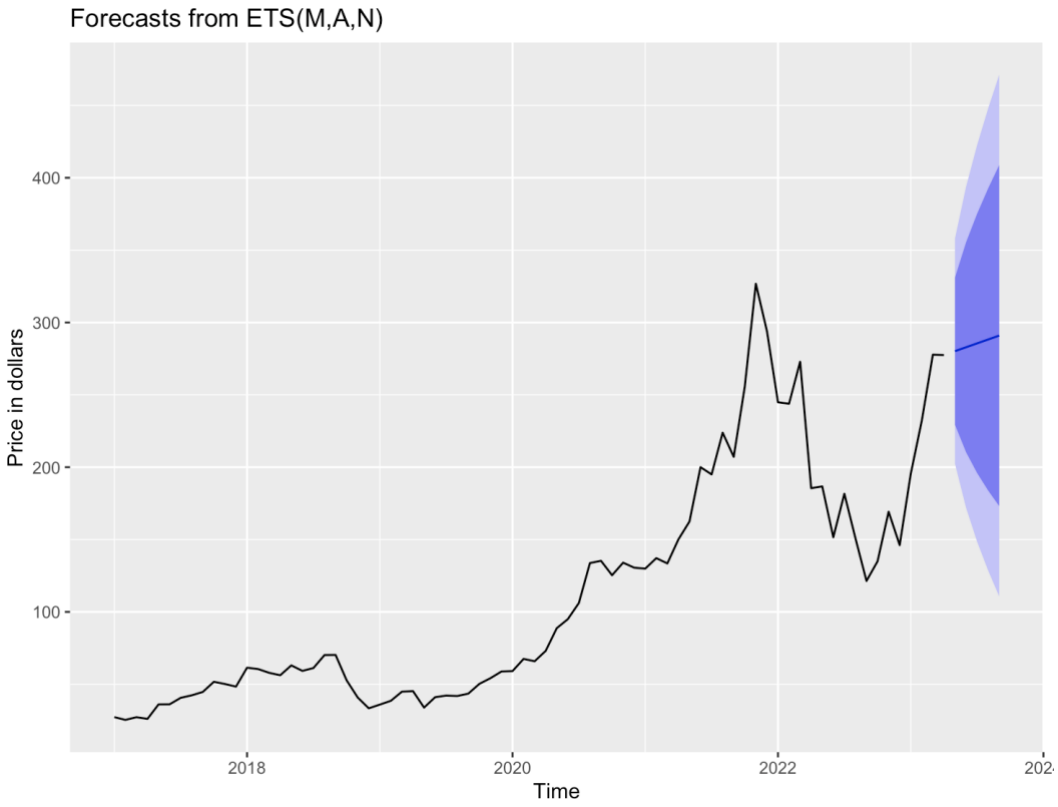


Figura 16: Prognoză ETS (sursa „R Studio”)

Se observă o tendință pozitivă și în acest caz, cu o zonă de prognozare care acoperă și valorile reale căutate. Putem acum să comparăm acuratețea modelului cu cel discutat anterior:

```
> round(accuracy(fit2),2)
                ME  RMSE   MAE  MPE  MAPE MASE ACF1
Training set 1.02 17.81 12.33 0.08 12.42 0.22 0.14
> round(accuracy(fit_ets),2)
                ME  RMSE   MAE   MPE  MAPE MASE  ACF1
Training set 0.69 21.81 13.58 -1.63 11.28 0.24 -0.01
```

Figura 17: Comparație erori HWM ETS (sursa „R Studio”)

Așadar, putem confirma faptul că modelul HW Multiplicativ produce un rezultat cu o acuratețe puțin mai ridicată decât ETS, întrucât atinge valori în modul mai mici la 4 dintre erorile consultate, deși ambele pot fi considerate satisfăcătoare datorită valorilor mici obținute.

Un alt mod de a realiza o comparație între nivelul de acuratețe al fiecărui model discutat este testul Diebold Mariano. Acesta funcționează pe două ipoteze:

- *H_0 implică faptul că prognozele au aceeași acuratețe*
- *H_1 contrariază prima ipoteză (prognozele au acuratețe diferită)*

Metodele se vor grupa două câte două pentru a realiza analiza:

```
> dm.test(residuals(fit1),residuals(fit2))# deoarece p < 0.1 respingem H0

        Diebold-Mariano Test

data:  residuals(fit1)residuals(fit2)
DM = 4.2303, Forecast horizon = 1, Loss function power = 2, p-value = 6.523e-05
alternative hypothesis: two.sided
```

Figura 18: Diebold Mariano pentru HWA & HWM (sursa „R Studio")

În cazul metodelor Holt-Winters inițial abordate, observăm un p-value mai mic de pragul de 0.1 (6.523e-05), ceea ce indică respingerea ipotezei nule și acceptarea ipotezei H_1.

```
> dm.test(residuals(fit1),residuals(fit_ets))# deoarece p < 0.1 respingem H0

        Diebold-Mariano Test

data:  residuals(fit1)residuals(fit_ets)
DM = 4.2304, Forecast horizon = 1, Loss function power = 2, p-value = 6.521e-05
alternative hypothesis: two.sided
```

Figura 19: Diebold Mariano pentru HWA & ETS (sursa „R Studio")

De asemenea, dacă se compară HW aditiv cu modelul ETS se observă un p-value mai mic de pragul de 0.1 (6.521e-05), ceea ce indică respingerea ipotezei nule și acceptarea ipotezei H_1.

```
> dm.test(residuals(fit2),residuals(fit_ets)) # deoarece p > 0.05 acceptam H0 pentru un prag de semnificatie de 5%

        Diebold-Mariano Test

data:  residuals(fit2)residuals(fit_ets)
DM = 1.676, Forecast horizon = 1, Loss function power = 2, p-value = 0.09791
alternative hypothesis: two.sided
```

Figura 20: Diebold Mariano pentru HWM & ETS (sursa „R Studio")

De această dată, dacă asociem HW multiplicativ cu ETS, ajungem la aceeași concluzie ca la analiza anterioară, întrucât cele două au o acuratețe similară.

După aplicarea modelului, este necesară verificarea existenței autocorelării în reziduuri.

```
res_hw_mlp <- residuals(fit2)
autoplot(res_hw_mlp) + xlab("Day") + ylab("") +
  ggtitle("Residuals from HW multiplicative")

gghistogram(res_hw_mlp) + ggtitle("Histogram of residuals")
```

Figura 21: Cod diagnostic pe reziduuri (sursa „R Studio")

Graficele tip folosite pentru depistarea si diagnosticare reziduurilor au fost folosite în analiză dar vor fi reluate într-o secțiune următoare. În urma parcurgerii datelor s-a determinat lipsa autocorelației în reziduuri, fapt care se poate observa și din figura următoare:

```
ggAcf(res_hw_mlp) + ggtitle("ACF of residuals")
```

Figura 22: Cod funcție de autocorelație a reziduurilor (sursa „R Studio")

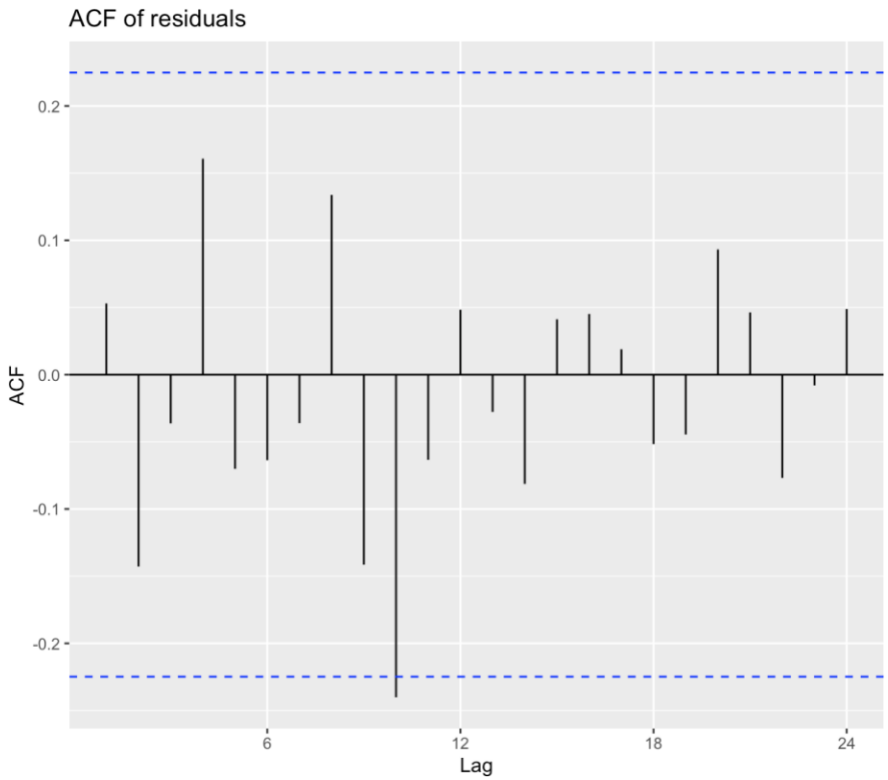


Figura 23: Autocorelație a reziduurilor (Lipsă) (sursa „R Studio")

De asemenea, folosim testul Jarque-Berra pentru a determina existența normalității reziduurilor. Acesta utilizează două ipoteze principale:

- H_0 *implică faptul că seria este normal distribuită*
- H_1 *contrariază prima ipoteză (seria nu este normal distribuită)*

```
> jarque.bera.test(res_hw_mlp) # deoarece p-value > 0.05, seria este distribuita normal

        Jarque Bera Test

data:  res_hw_mlp
X-squared = 1.2736, df = 2, p-value = 0.529
```

Figura 24: Statistici test Jarque-Bera (sursa „R Studio")

Se obține un p-value peste limita minimă pentru prima ipoteză, așadar acceptăm H_0 iar seria este normal distribuită.

Un alt test folosit pentru determinarea existenței autocorelației în reziduuri este Box Pierce. Pentru acest test, ipotezele sunt următoarele:

- H_0 implică faptul că seria reziduurilor nu prezintă autocorelare
- H_1 contrariază prima ipoteză (seria reziduurilor prezintă autocorelare)

În acest caz, s-au testat de la primul până la lag-ul de rangul 10, toate fără autocorelare în reziduuri.

```
> Box.test(res_hw_mlp, lag=1)

        Box-Pierce test

data:  res_hw_mlp
X-squared = 0.21359, df = 1, p-value = 0.644

> Box.test(res_hw_mlp, lag=3)

        Box-Pierce test

data:  res_hw_mlp
X-squared = 1.8639, df = 3, p-value = 0.6011

> Box.test(res_hw_mlp, lag=10) #  p-value > 0.1 => seria reziduurilor nu prezinta autocorelare

        Box-Pierce test

data:  res_hw_mlp
X-squared = 11.874, df = 10, p-value = 0.2936
```

Figura 25 – Testul Box-Pierce

Rezultate similare s-au obținut și în abordarea testului Ljung-Box pentru determinarea autocorelării în reziduuri. În final, se vor genera reziduurile pentru analiza Holt-Winters multiplicativ, întrucât oferă o privire de ansamblu asupra statisticilor discutate.

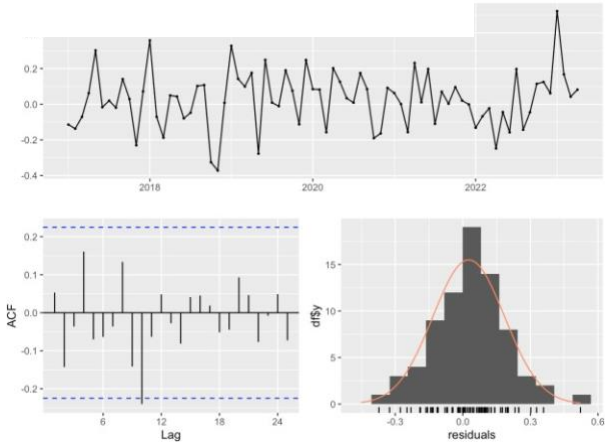


Figura 26: Statistici Reziduuri Holt-Winters multiplicativ (sursa „R Studio")

Staționaritatea este o proprietate fundamentală în analiza datelor care rezidă din prezența unei medii constante a seriei, deci lipsa unui trend global, dispersie și autocovarianță constante. Aceasta poate fi identificată atât prin metoda grafică, urmărind trendul seriei sau cu ajutorul *funcției de autocorelație ACF*. Pentru a obține o serie staționară este necesară diferențierea datelor de training pentru a stabiliza varianța și elimina trendul evident în primul output.

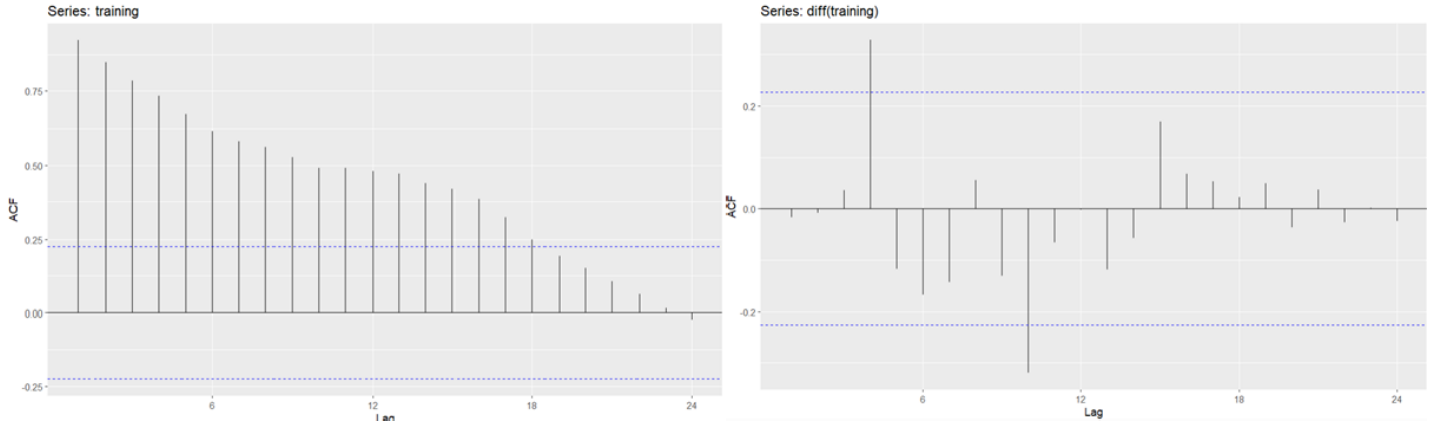


Figura 27 – Funcția de autocorelație ACF

La nivelul celui de-al doilea grafic avem o autocorelație aproape de 0 pentru majoritatea lagurilor, fără un model clar sau un trend persistent. De asemenea, valorile de autocorelație sunt minime la laguri mari, lucru care confirmă staționaritatea seriei.

```
> tseries::adf.test(training) #stationara

        Augmented Dickey-Fuller Test

data:  training
Dickey-Fuller = -3.5463, Lag order = 4, p-value = 0.04385
alternative hypothesis: stationary

> tseries::adf.test(training, k=1) # nestationara

        Augmented Dickey-Fuller Test

data:  training
Dickey-Fuller = -2.0826, Lag order = 1, p-value = 0.5418
alternative hypothesis: stationary

> tseries::adf.test(training, k=2) # nestationara

        Augmented Dickey-Fuller Test

data:  training
Dickey-Fuller = -2.1066, Lag order = 2, p-value = 0.532
alternative hypothesis: stationary

> tseries::adf.test(training, k=3) # nestationara

        Augmented Dickey-Fuller Test

data:  training
Dickey-Fuller = -2.2294, Lag order = 3, p-value = 0.4819
alternative hypothesis: stationary
```

Figura 28 - Testul Dickey-Fuller

În plus, pentru veridicitatea concluziei menționate anterior este util să folosim teste statistice privind staționaritatea seriilor temporale. *Testul Dickey-Fuller (ADF)* oferă o imagine clară în acest sens prin considerarea celor două ipoteze ale sale:

- *H0: seria admite o rădăcină unitară și este nestaționară*
- *H1: seria nu admite o rădăcină unitară și este staționară;*

Valoarea testului statistic la ordinul implicit al lag-ului (k=0) poate indica o respingere a ipotezei nule H0 prin valoarea mică de -3,5463. Totodată, p-value 0, 04385 < 0,05 forțează respingerea ipotezei nule deci concluzionăm staționaritatea seriei pentru acest lag order. În continuare, odată cu creșterea k-ului rezultatele testului ADF se apropie de 0 iar p-value > 0,05, respingem H1 și acceptăm H0, deci seria devine nestaționară.

Pentru interpretarea corectă a testului și îmbunătățirea acurateții abordăm cele 3 opțiuni pentru modelarea tendinței în testul ADF:

1. *none:* testează dacă seria de timp are o rădăcină unitară fără a considera prezența unei constante sau a unei tendințe;
2. *drift:* adaugă o componentă constantă (derivă) în model pentru a testa dacă există o tendință constantă în seria de timp, în plus față de rădăcina unitară;
3. *trend:* adaugă atât o constantă, cât și o componentă de tendință liniară în model pentru a testa existența unei constante și a unei tendințe liniare în seria de timp.

OPȚIUNE	**NONE**	**DRIFT**	**TREND**
z.lag.1	0.448	0.599	0.0409 *
value of test-statistic	0.7635	0.9956	2.4404
critical values 1pct	-2.6	-3.51	-4.04
critical values 5pct	-1.95	-2.89	-3.45
critical values 10pct	-1.61	-2.58	-3.15
CONCLUZII	|0.7635| > |-2.6| / |-1.95| / |-1.61| (F) => nestaționară 0.448 < 0.1 (F) => nestaționară	|0.9956| > |-3.51| / |-2.89| / |-2.58| (F) => nestaționară 0.599 < 0.1 (F) => nestaționară	|2.4404| > |-4.04| / |-3.45| / |-3.15| (F) => nestaționară 0.0409 < 0.1 (A) => staționară

Pentru a obține serii staționare alegem să diferențiem, rezultând următorul output:

OPȚIUNE	**NONE**	**DRIFT**	**TREND**
z.lag.1	1.81e-07***	9.46e-08***	9.05e-08 ***
value of test-statistic	-5.7847	17.7292	17.8543
critical values 1pct	-2.6	-3.51	-4.04
critical values 5pct	-1.95	-2.89	-3.45
critical values 10pct	-1.61	-2.58	-3.15
CONCLUZII	|-5.7847| > |-2.6| / |-1.95| / |-1.61| (A) => staționară 1.81e-07 < 0.1 (A) => staționară	|17.7292| > |-3.51| / |-2.89| / |-2.58| (A) => staționară 9.46e-08 < 0.1 (A) => staționară	|17.8543| > |-4.04| / |-3.45| / |-3.15| (A) => staționară 9.46e-08 < 0.1 (A) => staționară

Un alt test de verificare a staționarității este KPSS (Kwiatkowski - Phillips - Schmidt - Shin) care se bazează de asemenea pe 2 ipoteze: *H0 - seria este staționară* și *H1 - seria este nestaționară*.

```
######################
# KPSS Unit Root Test #
######################

Test is of type: mu with 3 lags.

Value of test-statistic is: 1.5873

Critical value for a significance level of:
                10pct  5pct 2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Figura 29 - KPSS

| 1.5873 | > | 0.347 |/| 0.463 |/| 0.574 |/| 0.739 | A

Respingem ipoteza nulă H0, acceptăm H1 => seria este nestaționară. De aceea, diferențiem și obținem:

| 0.1022 | > | 0.347 |/| 0.463 |/| 0.574 |/| 0.739 | F

Acceptăm ipoteza nulă, respingem H1 => seria este staționară.

Pentru a identifica dacă există o rădăcină unitate sau nu aplicăm testul Phillips-Perron bazat pe două ipoteze: *H0 - seria admite o rădăcină unitate* și *H1 - seria nu admite o rădăcină unitate*.

```
> PP.test(training) # serie nestationara ; p-value > 0.1

        Phillips-Perron Unit Root Test

data:  training
Dickey-Fuller = -2.2195, Truncation lag parameter = 3, p-value = 0.4859

> PP.test(diff(training)) # serie stationara; p-value <0.1

        Phillips-Perron Unit Root Test

data:  diff(training)
Dickey-Fuller = -8.6268, Truncation lag parameter = 3, p-value = 0.01
```

Figura 30 - Testul PP

În primul caz p-value 0.4859 > 0.05, deci respingem H0 și acceptăm H1 => seria nu admite o rădăcină unitate, deci este nestaționară. În urma diferențierii obținem un p-value < 0.05, implicit o serie staționară.

În analiza seriilor de timp, unul dintre modele de prognoza des întâlnite este modelul autoregresiv medie mobilă de parametri p și q, notat ARMA(p,q). Pentru a putea pune în practică acest model este nevoie ca seria de date să fie staționară, fără sezonalitate și să prezinte autocorelația în date.

Deși s-a demonstrat mai sus că seria diferențiată *diff(training)* îndeplinește criteriile de staționaritate și sezonalitate, aceasta nu prezintă autocorelație la aproape niciun lag, fapt confirmat de graficele ACF și PACF de mai jos și testul Box - Pierce

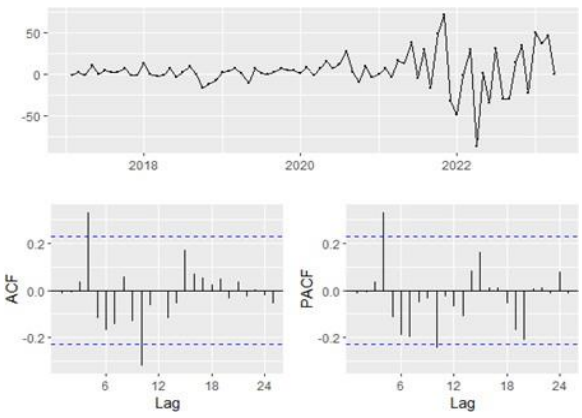


Figura 31 - Grafice pentru seria diferențiată

```
> Box.test(diff(training), lag=1)

        Box-Pierce test

data:  diff(training)
X-squared = 0.019729, df = 1, p-value = 0.8883

> Box.test(diff(training), lag=2)

        Box-Pierce test

data:  diff(training)
X-squared = 0.025037, df = 2, p-value = 0.9876

> Box.test(diff(training), lag=3)

        Box-Pierce test

data:  diff(training)
X-squared = 0.12309, df = 3, p-value = 0.9889
```

Figura 32 - Testul ADF

Unicul model al ecuației de medie acceptat este ARIMA(0,0,0). Acest model ARIMA este, practic, un model cu medie mobilă constantă și indică că nu există componente autoregresive sau componente integrate în model. Cu alte cuvinte, nu există trend sau sezonalitate în date, iar fiecare punct de date este independent de celelalte. Acest model este confirmat și de funcția *auto.arima* din R Studio.

```
> model_arima <- auto.arima(diff(training))
> model_arima
Series: diff(training)
ARIMA(0,0,0) with zero mean

sigma^2 = 492.7:  log likelihood = -338.92
AIC=679.84   AICc=679.89   BIC=682.15
```

Figura 33 - Model ARIMA

Cu alte cuvinte, pentru seria analizată, modelul ARIMA nu este potrivit pentru prognoză. De aceea, vom testa modelul GARCH pentru a captura și a prezice varianța sau volatilitatea condiționată a seriei de rentabilități lunare a acțiunilor NVDA pentru perioada ianuarie 2017 – martie 2024. Rentabilitățile au fost calculate după formula $rt = ln\left(\frac{p\,(t)}{p\,(t-1)}\right)$.

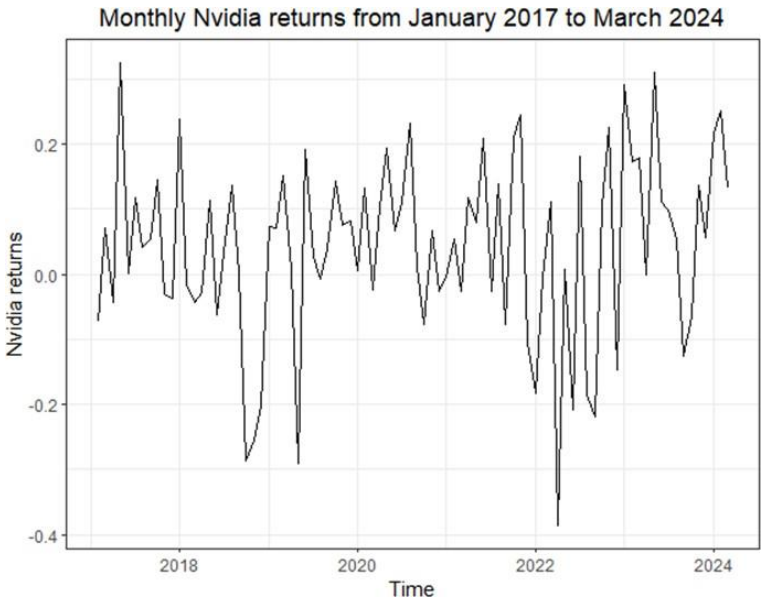


Figura 34 - Evoluția rentabilităților acțiunilor NVDA

Din imaginea de mai sus, observăm că seria rentabilităților este o serie cunoscută sub denumirea de *„zgomot alb”* (white noise), adică nu are nici trend, nici sezonalitate. Pentru a putea merge mai departe vom testa dacă seria este staționară cu ajutorul testului ADF.

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39851 -0.04263  0.03949  0.13412  0.32082

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
z.lag.1     -0.7552     0.1475  -5.121 1.97e-06 ***
z.diff.lag  -0.1293     0.1114  -1.161    0.249
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.148 on 82 degrees of freedom
Multiple R-squared:  0.4427,    Adjusted R-squared:  0.4291
F-statistic: 32.57 on 2 and 82 DF,  p-value: 3.894e-11


Value of test-statistic is: -5.1214

Critical values for test statistics:
     1pct  5pct 10pct
tau1 -2.6 -1.95 -1.61
```

Figura 35 - Testul ADF

Observăm ca seria este staționară, deci putem continua cu estimarea ecuației mediei.

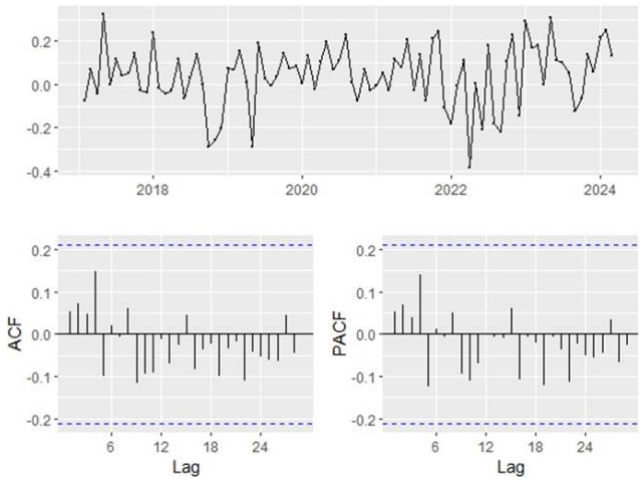


Figura 36 - Graficele rentabilităților

Din graficul ACF și PACF observăm că lagul maximal pentru componentele AR și MA este 0, deci, din nou, obținem modelul ARIMA(0,0,0). Vom continua totuși cu diagnosticul pe reziduuri, în special testul *ARCH LM* ce are ipoteza nulă că seria reziduurilor nu are efecte ARCH și cea alternativă – seria are efecte ARCH.

```
> ArchTest(residuals(arima000), lag = 1) # p > 0.1 => nu avem efecte ARCH

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  residuals(arima000)
Chi-squared = 0.16949, df = 1, p-value = 0.6806
```

Figura 37 - Testul ARCH LM

Deoarece p-value > 0.1, înseamnă că nu avem efecte ARCH și nu putem continua modelul GARCH. Observăm acest lucru și din graficul PACF al seriei reziduurilor ridicate la pătrat.

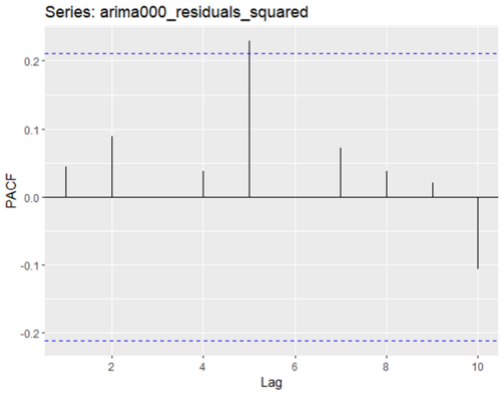


Figura 38 - Graficul PACF pentru reziduurile modelului ARIMA

Concluzie

În cadrul acestui proiect, ne-am concentrat pe modelarea econometrică a seriei de prețuri a acțiunilor companiei Nvidia, urmărind frecvența lunară, pentru o perioadă cuprinsă între ianuarie 2017 și martie 2024. Scopul nostru a fost de a anticipa evoluția prețului acțiunilor pentru perioadele următoare. În acest sens, am explorat mai multe modele de previziune pentru serii de timp, printre care Holt-Winters aditiv și multiplicativ, extrapolarea exponențială cu spațiu de stare (ETS), modelul de medie mobilă integrată autoregresiv (ARIMA) și heteroscedasticitatea condiționată autoregresiv generalizată (GARCH).

Din analiza noastră, am constatat că modelul Holt-Winters multiplicativ se potrivește cel mai bine seriei noastre de timp. Acest model este preferat, deoarece poate captura și modela eficient tendința și sezonalitatea, care se schimbă în mod exponențial în cazul seriei noastre de prețuri. Alegerea acestui model a fost susținută de evaluarea acurateței și diagnosticul bun al reziduurilor.

APLICAȚIA 2

Introducere

Este de necontestat faptul că România a dobândit o creștere semnificativă a nivelului de trai și al bogăției naționale în ultimii 30 de ani, fapt garantat de trendul progresiv al valorii Produsului Intern Brut Național (cu o valoare înregistrată de 132.78 miliarde RON în 1991 și 1.383,78 miliarde RON în anul 2022). Această schimbare exponențială ar putea fi interpretată ca o amplificare a calității pe toate planurile, însă realitatea își spune cuvântul în multe domenii, mai ales în cel academic. Următorul capitol are ca scop abordarea acestui subiect controversat, prin evidențierea investițiilor făcute de statul român pe parcursul ultimelor trei decenii în ramuri ale societății precum unitățile de învățământ și bibliotecile naționale, și în speță modul prin care aceste oscilații la nivelul infrastructurii influențează sau nu numărul de absolvenți ai ciclul academic (preuniversitar și universitar). Se vor folosi tehnici precum cointegrarea pentru a determina în ce măsură sunt încurajați elevii și studenții să-și termine studiile prin nivelul sporit de oportunități de învățare, sau vice versa, în ce măsură este statul îndemnat să investească în aceste aspecte ale vieții academice ca urmare a rezultatelor bune, sau mai puțin bune, ale sistemului educațional.

În studiul de caz atașat analizei vom avea în vedere date obținute de la Institutul Național de Statistică, pentru perioada 1991-2021:

- *Absolvenți pe niveluri de educație, sexe, macroregiuni, regiuni de dezvoltare și județe* (Totalul la nivel național pentru toate nivelurile de educație și ambele sexe)
- *Unitățile de învățământ, pe categorii, forme de proprietate, macroregiuni, regiuni de dezvoltare și județe* (Totalul la nivel național pentru toate categoriile de unități de învățământ și toate formele de proprietate)
- *Biblioteci pe categorii de biblioteci, județe și localități* (Totalul la nivel național pentru toate categoriile de biblioteci)

Literatură de specialitate

Un considerent eficient în determinarea evoluției economice a unui stat, în speță a evoluției României, este prin excelență Produsul Intern Brut Național. Referitor la mențiunea realizată inițial în introducere, este evident faptul că evoluția statului român din acest considerent a fost și este în continuare una foarte favorabilă, întrucât ultimii 30 de ani dezvoltă un trend prioritar ascendent. Această observație poate fi dublată de informațiile oficiale oferite de The World Bank (2022), unde este ilustrată creșterea de aproximativ 1000% a PIB-ului național în perioada menționată.

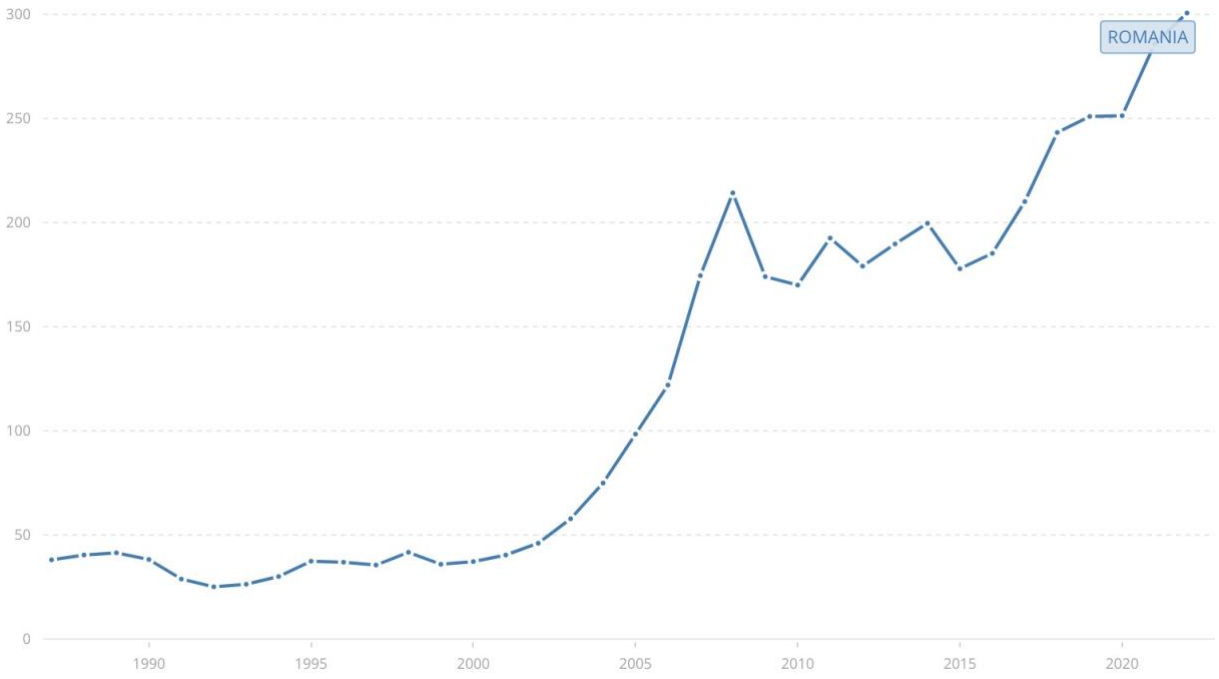


Figura 39 - GDP (current US$) - Romania (Sursă: The World Bank, 2022)

Această realizare, care în teorie ar echivala o evoluție considerabilă pe toate planurile, ridică anumite semne de întrebare, întrucât realitatea statistică prezintă anumite probleme semnificative dintr-o multitudine de puncte de vedere, în speță infrastructura și investițiile acordate sistemului educațional și domeniului academic, deloc în conformitate cu trendul aparent ascendent al economiei.

Poate cele mai importante întrebări relevante în subiectul abordat sunt „De ce?” sau „Cum?” pe subiectul evoluției deloc proporționale a PIB-ului cu alte aspecte de natură economică ale statului român. Pentru a găsi o soluționare la această dilemă se va consulta articolul „Evoluţia Produsului Intern Brut al României”, semnat de profesorul universitar al Academiei de Studii Economice București, domnul Doctor Constantin Anghelace, unde este discutată structura de alcătuire a acestui indicator pe perioada 2001-2013. Studiul urmărește o anumită „privatizare” a Produsului Intern Brut pe parcursul perioadei analizate, urmând un trend care ar putea fi cu ușurință considerat constant și în anii următori.

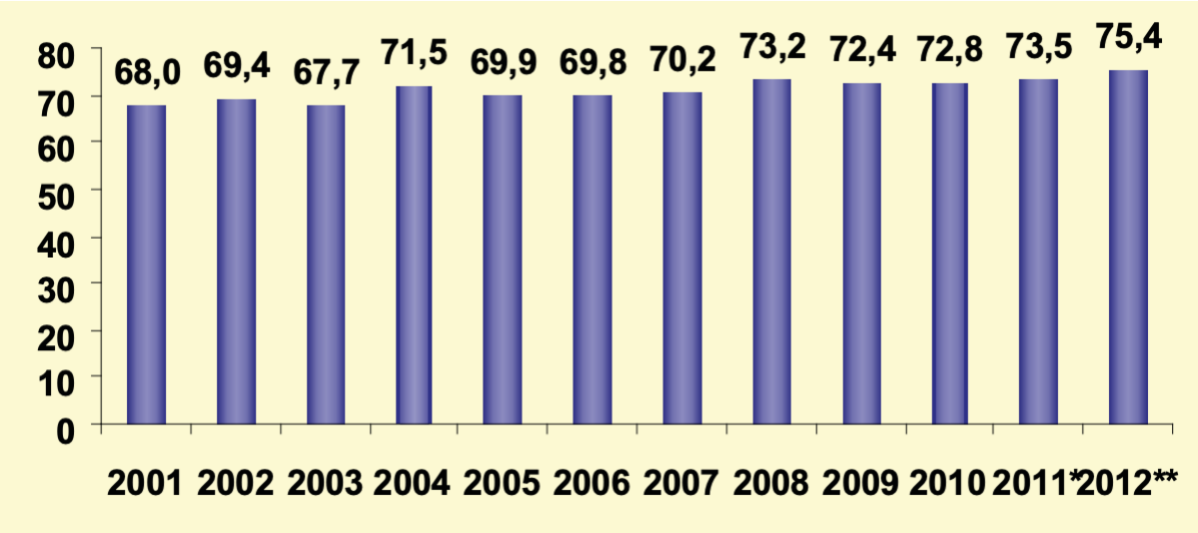


Figura 40 - Produsul Intern Brut, ponderea sectorului privat între anii 2001 – 2012 în procente (Sursă: Prof. univ. dr. Constantin ANGHELACHE, 2013)

Conform analizei consultate, se observă o intenție la nivel național de pivotare a acțiunilor economice în direcția privatizării, cu accent pe amplificarea domeniilor care sunt considerate cele mai capabile să genereze rezultate financiare considerabile.

Această atitudine aparent progresistă asupra încurajării dezvoltării naționale ar putea fi considerată un pas înainte, dacă nu ar fi însoțită de o scădere drastică a infrastructurii în alte domenii considerate chiar mai relevante decât lumea privată, anume sistemul educațional public și deținut de stat.

Date oficiale procurate din baza de date a Institutului Național de Statistică arată o scădere alarmantă atât a numărului de unități de învățământ, cât și a bibliotecilor naționale, în ultimii 30 de ani, fapt care trădează un dezinteres total asupra garantării unui spațiu eficient, nou și prielnic pentru noile generații.

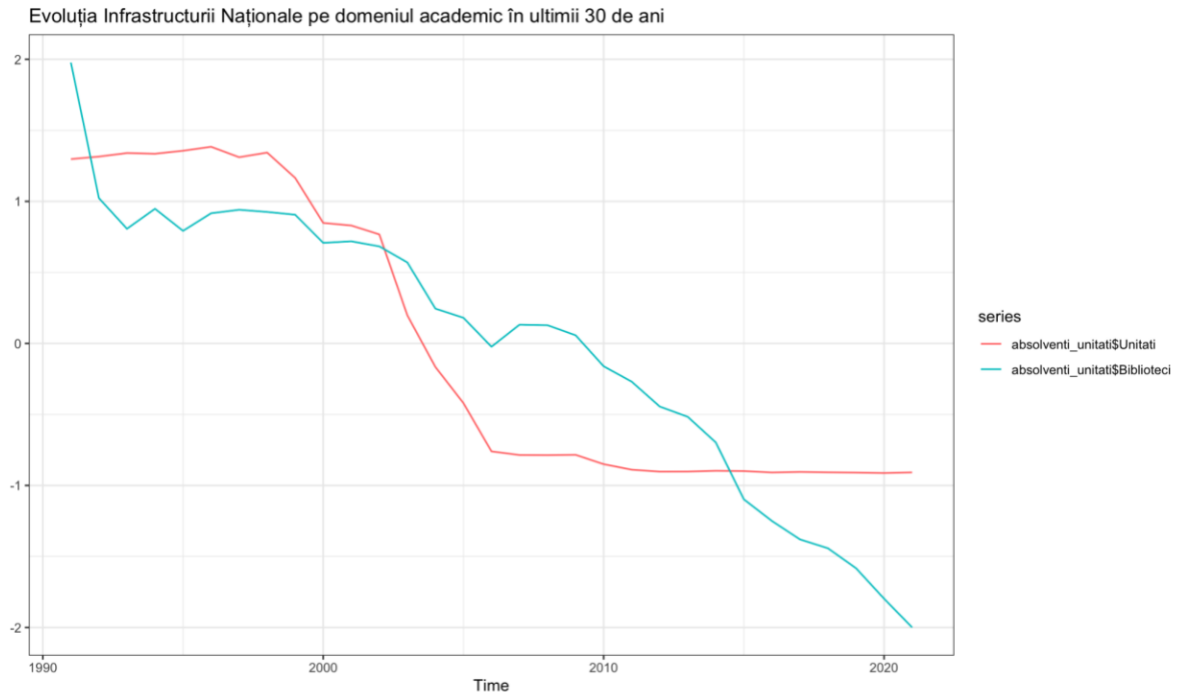


Figura 41 -Evoluția infrastructurii naționale în domeniul academic în ultimii 30 de ani

Conform reprezentării grafice în limbajul de programare R generată mai sus folosind datele oficiale INS, se constată o scădere semnificativă, de la 28951 de unități de învățământ pentru anul 1991, la doar 7015 la nivel național în anul 2021, și de la 15749 de biblioteci naționale active în primul an analizat, la 8458 pentru ultimul an prezent în model.

Deși aceste schimbări în infrastructura națională la nivel academic pot fi considerate semnificative și devastatoare la suprafață, există totuși varianta ca populația angrenată în studii preuniversitare și universitare să nu aibă de suferit din cauza acestor limitări ale oportunităților de învățare. Pentru testarea acestei ipoteze, vom consulta articolul oficial intitulat „Raport privind starea învățământului preuniversitar din România", semnat de Ministerul Educației (București, 2022).

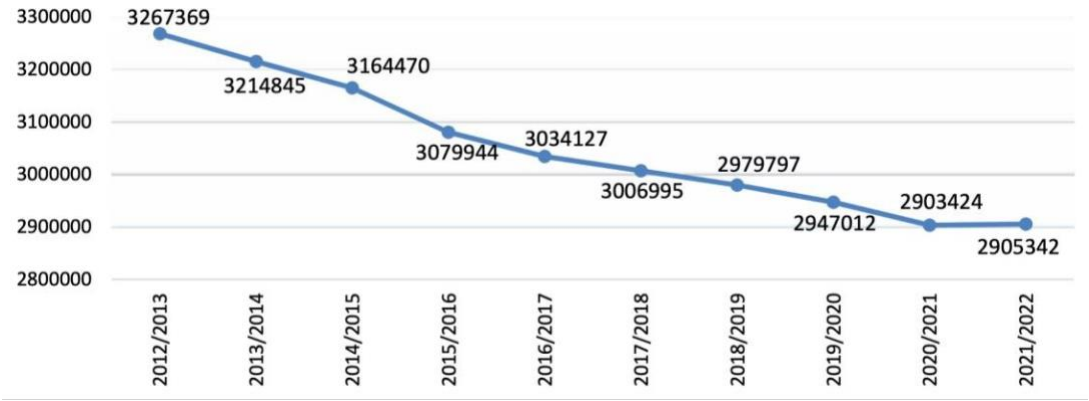


Figura 42 - Evoluția efectivelor de elevi din învățământul preuniversitar

Printre altele, se vorbește despre un trend prioritar descendent al numărului de elevi participanți ai unui ciclu preșcolar, primar, gimnazial, liceal, profesional sau postliceal, fapt datorat unei multitudini de factori precum schimbările semnificative ale sistemului educațional (prin introducerea clasei 0), dezinteresul pentru lumea academică și infrastructura.
Indiferent de valoarea numerică a participanților la studii preuniversitare, care poate fi manipulată de factori precum dorința de a studia în țări străine, se poate observa și o scădere minoră în procentul de tineri care sunt angrenați în studii, numită „Rata brută de cuprindere în învățământul preuniversitar (de la învățământul preșcolar la cel postliceal) ca raport din populația în vârstă de 3-21 de ani" în articolul analizat. Rezultatele analizei evidențiază o scădere de la 78.1% în anul școlar 2014/2015 la 75.4% în 2021/2022, cu rezultate similare pentru grupe de vârstă mai detaliate.

Conform studiilor parcurse, se poate constata un trend similar urmat atât de scăderea calității infrastructurii sistemului educațional român, cât și de nivelul de participanți sau absolvenți de studii preuniversitare și/sau universitare la nivel național. Însă, există posibilitatea ca aceste variabile considerate să fie similare prin simplă coincidență, caz în care problema acestui trend descendent este localizată altundeva.

Pentru a determina dacă dezinteresul statului român în a investi în spații prielnice pentru învățare, precum unități de învățământ și biblioteci, are sau nu un efect devastator asupra numărului de absolvenți de studii noi intrați pe piața muncii, se va alcătui o analiză de cointegrare pe variabilele discutate.

Analiză pe Serie de Timp

Înainte de a iniția analiza propriu-zisă este nevoie de un stadiu de preprocesare a datelor statistice pentru a putea genera seria de timp necesară.

Se observă în captura de ecran inserată mai jos cum datele au fost mai întâi structurate în trei data frame-uri diferite, în care a fost selectată strict perioada de timp dorită, întrucât fiecare set de date conține intervale diferite pentru datele statistice necesare.

```
# Citirea valorilor din fisierul absolventi.csv si crearea unui dataframe
absolventi <- read.csv("absolventi.csv", header = TRUE, sep = ",")
absolventi <- absolventi[2:32,]
absolventi <- absolventi[order(absolventi$Ani),]

# Citirea valorilor din fisierul unitati_invatamant.csv si crearea unui dataframe
unitati_invatamant <- read.csv("unitati_invatamant.csv", header = TRUE, sep = ",")
unitati_invatamant <- unitati_invatamant[2:32,]
unitati_invatamant <- unitati_invatamant[order(unitati_invatamant$Ani),]

# Citirea numarului de biblioteci din fisierul biblioteci.csv si crearea unui dataframe
biblioteci <- read.csv("biblioteci.csv", header = TRUE, sep = ",")
biblioteci <- biblioteci[2:32,]
biblioteci <- biblioteci[order(biblioteci$Ani),]
```

Figura 43 - Citirea setului de date

Este de menționat faptul că se va structura în paralel și un data frame pentru creșterea anuală a PIB-ului național, care va fi folosit pentru o comparație grafică între indicatorii selectați. Acest set de date nu va fi folosit însă în analizele viitoare.

```
# Citirea cresterii pib-ului din fisierul crestere_pib.csv si crearea unui dataframe
crestere_pib <- read.csv("crestere_pib.csv", header = TRUE, sep = ",")
crestere_pib <- crestere_pib[2:32,]
crestere_pib <- crestere_pib[order(crestere_pib$Ani),]
```

Figura 44 - Citirea setului de date cu Produsul Intern Brut

Odată ce toate seturile de date sunt în conformitate cu cerințele analizei, se va structura un nou data frame la care se vor concatena toate coloanele cu valorile obținute.

```
# Citirea valorilor din absolventi si unitati_invatamant si crearea unui dataframe cu row.names = Ani
absolventi_unitati <- data.frame(absolventi[,2], unitati_invatamant[,2], row.names = absolventi$Ani)

# Adaugarea cresterii pib-ului in dataframe-ul absolventi_unitati
absolventi_unitati$Crestere_PIB <- crestere_pib[,2]

# Adaugarea numarului de biblioteci in dataframe-ul absolventi_unitati
absolventi_unitati$Biblioteci <- biblioteci[,2]
colnames(absolventi_unitati)=c("Absolventi","Unitati","Crestere PIB","Biblioteci")
```

Figura 45 - Crearea dataFrame-urilor necesare

În continuare se va utiliza variabila *absolventi_unitati* pentru a structura cele patru serii de timp necesare testării. Mai jos se poate vedea utilizarea funcției *ts( )* cu proprietatea *frequency* egală cu 1, pentru a semnala utilizarea de date anuale.

```
# Declararea variabilelor de tip „time series”
absolventi_ts <- ts(absolventi_unitati$Absolventi, start = 1991, frequency = 1)
unitati_ts <- ts(absolventi_unitati$Unitati, start = 1991, frequency = 1)
crestere_pib_ts <- ts(absolventi_unitati$`Crestere PIB`, start = 1991, frequency = 1)
biblioteci_ts <- ts(absolventi_unitati$Biblioteci, start = 1991, frequency = 1)

absolventi_ts <- window(absolventi_ts, start=1991, end=2021)
unitati_ts <- window(unitati_ts, start=1991, end=2021)
crestere_pib_ts <- window(crestere_pib_ts, start=1991, end=2021)
biblioteci_ts <- window(biblioteci_ts, start=1991, end=2021)
```

Figura 46 - Crearea seriilor temporare

Întrucât vom lucra cu date numerice pe intervale de valori drastic diferite, se va recurge la normalizarea datelor după formula matematică

$$ValoareNormalizată = (Valoare - MediaAritmetică)/AbatereaStandard$$

pentru a genera graficele statistice.

```
# Normalizarea datelor
absolventi_ts <- (absolventi_ts - mean(absolventi_ts)) / sd(absolventi_ts)
unitati_ts <- (unitati_ts - mean(unitati_ts)) / sd(unitati_ts)
crestere_pib_ts <- (crestere_pib_ts - mean(crestere_pib_ts)) / sd(crestere_pib_ts)
biblioteci_ts <- (biblioteci_ts - mean(biblioteci_ts)) / sd(biblioteci_ts)
```

Figura 47 - Normalizarea datelor

Așadar, în urma pre-procesării, se poate vizualiza evoluția comparativă a seturilor de date. Din nefericire, această suprapunere regăsită în figura de mai jos arată o discrepanță semnificativă între trendul prioritar ascendent al creșterii anuale a Produsului Intern Brut și scăderile alarmante ale valorilor indicatorilor asociați ariei academice de la an la an.

Figura 48 - Graficul seriilor

Un factor fundamental în garantarea rezultatelor corespunzătoare în analiza cointegrării variabilelor este staționaritatea seriilor de timp. Mai întâi, vom aborda varianta grafică de verificare a persistenței modelului, pentru a identifica posibile semnale de alarmă.

```
# Determinarea persistentei modelului
ggtsdisplay(absolventi_ts)
ggtsdisplay(unitati_ts)
ggtsdisplay(biblioteci_ts)
```

Figura 49 - Determinarea persistenței modelului

1. Evoluția numărului de absolvenți ai unui ciclu academic la nivel național

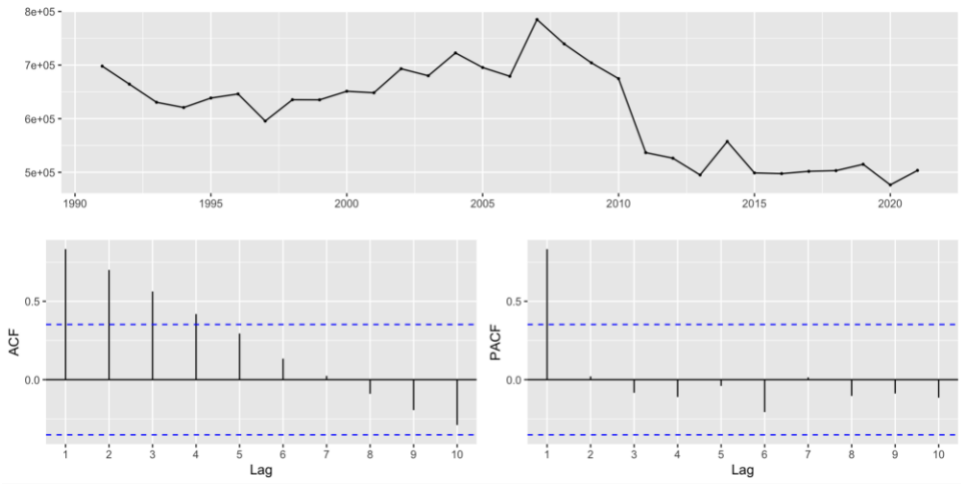
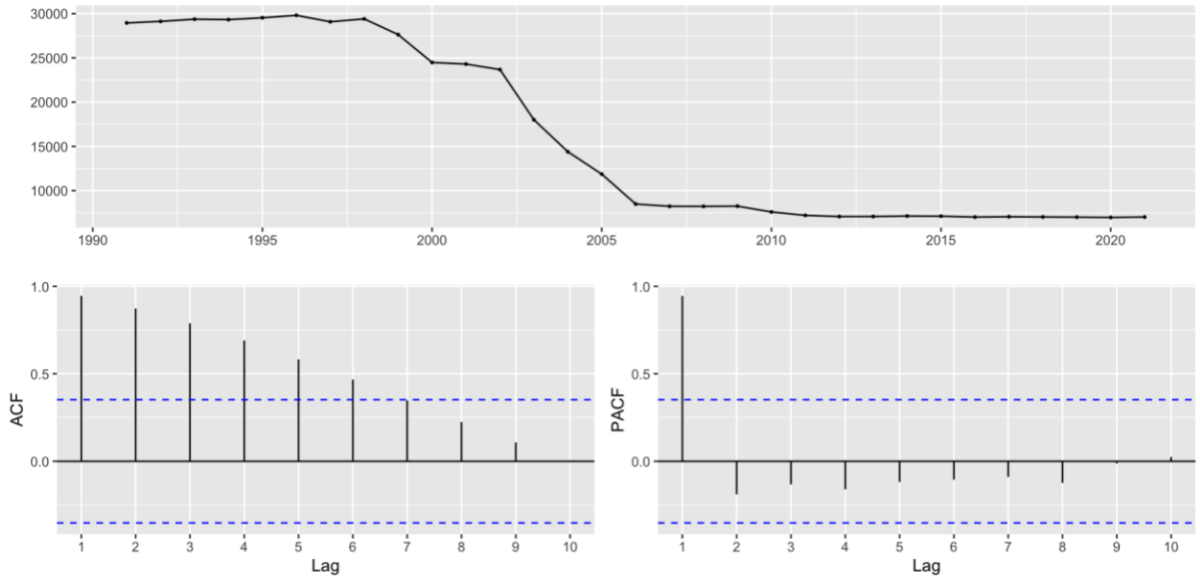


Figura 50 - Evoluția numărului de absolvenți ai unui ciclu academic la nivel național

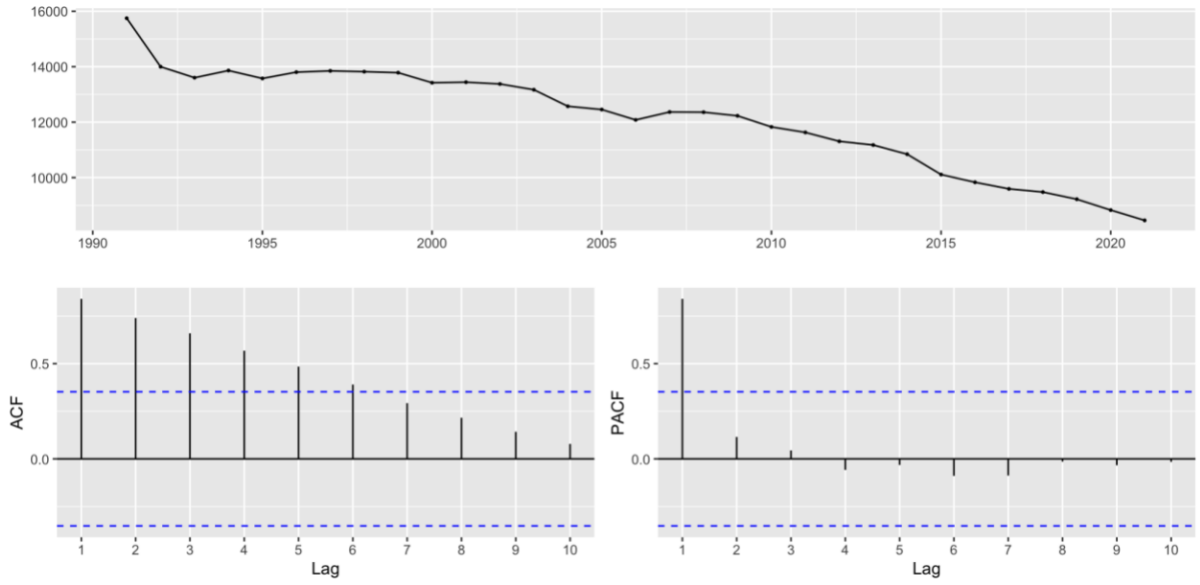
2. Evoluția numărului de unități de învățământ la nivel național

Figura 51- Evoluția numărului de unități de învățământ la nivel național



3. Evoluția numărului de biblioteci la nivel național

Figura 52 - Evoluția numărului de biblioteci la nivel național



Se observă că toate cele trei serii prezintă nestaționaritate conform reprezentării grafice ADF.

Mai departe vom folosi testul statistic ADF (Augmented Dickey–Fuller) pentru a determina situația matematică actuală a indicatorilor:

1. Indicatorul *absolventi_ts*

Figura 53 - Indicatorul absolventi_ts

```
Value of test-statistic is: -1.4872 1.0447 1.2304

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.15 -3.50 -3.18
phi2  7.02  5.13  4.31
phi3  9.31  6.73  5.61
```

Valoarea calculată de 1.2304 este mai mică în modul decât cele trei de pe linia tau3 din tabela prezentă. Așadar, seria este nestaționară conform ADF.

2. Indicatorul *unitati_ts*

Figura 54 – Indicatorul unitati_ts

```
Value of test-statistic is: -1.2267 1.3871 1.1151

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.15 -3.50 -3.18
phi2  7.02  5.13  4.31
phi3  9.31  6.73  5.61
```

Valoarea calculată de 1.1151 este mai mică în modul decât cele trei de pe linia tau3 din tabela prezentă. Așadar, seria este nestaționară conform ADF.

3. Indicatorul *biblioteci_ts*

Figura 55 – Indicatorul unitati_ts

```
Value of test-statistic is: -1.6696 7.1234 4.3717

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -4.15 -3.50 -3.18
phi2  7.02  5.13  4.31
phi3  9.31  6.73  5.61
```

Valoarea calculată de 4.3717 este mai mare în modul decât cele trei de pe linia tau3 din tabela prezentă. Așadar, seria este staționară conform ADF.

Pentru a garanta o variantă finală privind problema staționarității pentru seria *biblioteci_ts* vom folosi comanda *ndiffs( )* pentru a determina numărul de diferențieri necesare pentru a ajunge la staționaritate:

```
> ndiffs(biblioteci_ts)
[1] 1
```

Așadar, conform reprezentării grafice ADF și *ndiffs( )*, seria de timp *biblioteci_ts* este nestaționară.

Se va folosi funcția *diff( )* din R pentru a genera prima diferență a seriilor utilizate în analiză.

Pentru a garanta corectitudinea datelor, vom folosi funcția *ndiffs( )* pentru a verifica numărul de diferențieri rămase pentru fiecare serie până la staționaritate:

```
> ndiffs(absolventi_diff)
[1] 0
> ndiffs(unitati_diff)
[1] 0
> ndiffs(biblioteci_diff)
[1] 0
```

Întrucât toate variabilele implicate în model sunt staționare după prima diferență, putem considera că setul se încadrează în categoria I1 pentru cointegrare.

Pentru a analiza relația de cointegrare dintre seriile expuse anterior vom aplica testul Engle-Granger. Prin definiție, cointegrarea are rolul de a evidenția relația pe termen lung între variabile în ciuda fluctuațiilor de trend care rezidă sau nu în urma unor fenomene economice. Pentru a avea un rezultat veridic considerăm output-ul testul de staționaritate Dickey-Fuller efectuat anterior.

Ipoteze testului sunt: *H0 - seriile nu sunt cointegrate* și *H1 - seriile sunt cointegrate*, ipoteze de care vom ține cont în diagnosticarea naturii seriilor. Analiza se face la nivelul reziduurilor dintre cele 2 variabile care justifică staționaritatea sau nestaționaritatea lor.

Dacă ne referim la seriile asociate absolvenților și unităților de învățământ obținem următorul rezultat:

```
Engle-Granger Cointegration Test
alternative: cointegrated

Type 1: no trend
    lag      EG p.value
 3.0000 -2.6591  0.0942
-----
 Type 2: linear trend
    lag      EG p.value
  3.000  -0.716   0.100
-----
 Type 3: quadratic trend
    lag      EG p.value
 3.0000 -0.0971  0.1000
-----------
```

Pentru a interpreta testul evaluăm p-value pentru no trend și linear trend. Valorile sunt mai mari decât 0.05, deci respingem ipoteza nulă și acceptăm H1 => seriile sunt cointegrate. În continuare am testat și reciproca care s-a dovedit a fi adevărată.

Figura 56 - Cointegrare absolvenți, unități de învățământ

```
Engle-Granger Cointegration Test
alternative: cointegrated

Type 1: no trend
    lag      EG p.value
   3.00   -2.56    0.10
-----
 Type 2: linear trend
    lag      EG p.value
  3.000  -0.126   0.100
-----
 Type 3: quadratic trend
    lag      EG p.value
 3.0000 -0.0321  0.1000
-----------
```

Figura 57 - Cointegrare absolvenți, biblioteci

```
Engle-Granger Cointegration Test
alternative: cointegrated

Type 1: no trend
    lag      EG p.value
   3.00   -1.15    0.10
-----
 Type 2: linear trend
    lag      EG p.value
  3.000   0.113   0.100
-----
 Type 3: quadratic trend
    lag      EG p.value
   3.00   -1.39    0.10
```

Figura 58 - Cointegrare unități de învățământ, biblioteci

Rezultatele astfel obținute indică recurența cointegrării și între celelalte serii.

În aceeași măsură ne propunem și aplicarea testului Johansen pentru a testa relațiile de cointegrare simultan. La nivel de funcționalitate al testului acesta se bazează pe un model VECM care este o extensie a modelului VAR.

- *Selectarea lagului*

```
> lagselect$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     6      6      6      7
```

- *Metoda Trace*

```
####################
# Johansen-Procedure #
####################

Test type: maximal eigenvalue statistic (lambda max) , without linear trend and constant
in cointegration

Eigenvalues (lambda):
[1] 8.886458e-01 6.638429e-01 2.214917e-01 1.942890e-16

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 2 |  6.51  7.52  9.24 12.97
r <= 1 | 28.34 13.75 15.67 20.20
r = 0  | 57.07 19.77 22.00 26.81

Eigenvectors, normalised to first column:
(These are the cointegration relations)
```

Figura 59 - Testul Johansen, trace method

Valoarea *r* este rankul matricei care include date asociate absolvenților, unităților de învățământ și bibliotecilor. Pentru *r = 0*, testul are o valoare mai mare decât toate valorile critice, deci avem cel puțin o relație de cointegrare. Pentru *r <= 1*, valoarea testului respectă

același termen de comparație, deci avem 2 relații de cointegrare, în vreme ce pentru *r <= 2*, valoarea testului este mai mică decât valorile critice, deci avem cel mult 2 relații de cointegrare.

- *Metoda valorii proprii maxime*

```
######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , without linear trend and constant
in cointegration

Eigenvalues (lambda):
[1] 8.886458e-01 6.638429e-01 2.214917e-01 1.942890e-16

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 2 |  6.51  7.52  9.24 12.97
r <= 1 | 28.34 13.75 15.67 20.20
r = 0  | 57.07 19.77 22.00 26.81

Eigenvectors, normalised to first column:
(These are the cointegration relations)
```

Figura 60 - Testul Johansen, metoda valorii proprii maxime

Sunt furnizate aceleași valori, deci avem cel mult 2 relații de cointegrare în conformitate și cu acest model.

Datorită cointegrării dintre serii vom aplica metoda VECM pentru a vedea îndeaproape legătura dintre cele 3 serii.

```
> summary(Model1)
#############
###Model VECM
#############
Full sample size: 31    End sample size: 25
Number of variables: 3  Number of estimated slope parameters 48
AIC -241.5258   BIC -180.582    SSR 1.402609
Cointegrating vector (estimated by 2OLS):
   Absolventi   Unitati Biblioteci        const
r1          1 0.7073212  -1.314319 1.156531e-15
```

Figura 61 - Metoda VCM

Relația de cointegrare este dată de rândul r1. Astfel, valoarea 1 asociată categoriei *Absolvenți* indică faptul că este variabila de referință. Restul relațiilor se interpretează cu semn opus astfel:

- absolvenții și unitățile de învățământ au o relație negativă (crește una, scade cealaltă);

- absolvenții și bibliotecile au o relație pozitivă;

- absolvenți și constanta au o relație pozitivă;

Ecuațiile ECT au rolul de a interpreta relațiile de cointegrare între serii pe termen scurt sau lung.

```
                        ECT                 Absolventi -1       Unitati -1          Biblioteci -1       Absolventi -2       Unitati -2          Biblioteci -2       Absolventi -3
Equation Absolventi -1.5777(0.7195).  0.6792(0.4446)  1.0954(0.8941)  0.0184(1.0044)  0.9562(0.3342)* -2.3247(0.8516)* 0.5712(0.8635)  0.6615(0.3804)
Equation Unitati     0.4778(0.2957)  -0.3598(0.1827). 0.4870(0.3674)  0.7608(0.4127). -0.1358(0.1373) -0.5632(0.3499)  -0.3098(0.3549) -0.2341(0.1563)
Equation Biblioteci 0.0039(0.2888)  0.0089(0.1785)  0.6568(0.3589)  0.2396(0.4032)  0.0229(0.1341) -0.8523(0.3418)* 0.2379(0.3466)  -0.0065(0.1527)
                        Unitati -3          Biblioteci -3       Absolventi -4       Unitati -4          Biblioteci -4       Absolventi -5       Unitati -5
Equation Absolventi 0.9646(0.9506)  -0.2919(0.7562)  0.7133(0.3659).  -2.8705(1.0335)*  -0.6149(0.7685)  0.3614(0.2744)   1.8535(0.9054).
Equation Unitati    1.3150(0.3906)** -0.6297(0.3107). -0.2687(0.1504) -0.5233(0.4247)  0.4161(0.3158)  -0.1945(0.1128)  0.3702(0.3721)
Equation Biblioteci 0.5889(0.3816)  -0.1055(0.3035)  0.1051(0.1469)  -0.7348(0.4148)  0.3595(0.3085)  0.0404(0.1101)   0.6540(0.3634)
                        Biblioteci -5
Equation Absolventi -0.1417(0.4689)
Equation Unitati     0.2353(0.1927)
Equation Biblioteci -0.1084(0.1882)
```

Figura 62 - Ecuațiile ECT

Valoarea negativă de -1.58 a ecuației ECT a absolvenților indică prezența unei relații pe termen lung cu celelalte serii. Numărul absolvenților este influențat de numărul absolvenților și cel al unităților de învățământ de acum 2 și 4 ani. ECT asociat bibliotecilor este negativ, fiind influențat pozitiv doar de numărul de unități de învățământ de acum 2 ani. Numărul unităților de învățământ este influențat de numărul de unități de învățământ de acum 3 ani.

În continuare, vom furniza diagnosticul pe reziduuri analizând:

- *autocorelarea*

Figura 63 – Testul Portmanteau

```
        Portmanteau Test (asymptotic)

data:  Residuals of VAR object Model1VAR
Chi-squared = 29.548, df = 3, p-value = 1.718e-06
```

, p-value < 0.05 => avem autocorelare

- *heteroschedascitate*

Figura 64 – Testul ARCH

```
        ARCH (multivariate)

data:  Residuals of VAR object Model1VAR
Chi-squared = 25.167, df = 36, p-value = 0.9119
```

, p-value > 0.1 => reziduuri homosechedastice

- *normalitate*

Figura 65 – Teste de Normalitate

```
        JB-Test (multivariate)

data:  Residuals of VAR object Model1VAR
Chi-squared = 18.111, df = 6, p-value = 0.00596


$Skewness

        Skewness only (multivariate)

data:  Residuals of VAR object Model1VAR
Chi-squared = 8.7576, df = 3, p-value = 0.03269


$Kurtosis

        Kurtosis only (multivariate)

data:  Residuals of VAR object Model1VAR
Chi-squared = 9.3538, df = 3, p-value = 0.02494
```

=> reziduurile nu sunt normal distribuite;

Testul de cauzalitate Granger este o metodă statistică folosită pentru a determina dacă o serie temporală poate oferi informații utile în prezicerea unei alte serii temporale. Acesta se bazează pe următoarea idee: dacă o variabilă X "Granger - cauzează" o variabilă Y, atunci valorile trecute ale lui X ar trebui să conțină informații utile pentru prezicerea valorilor viitoare ale lui Y. Testul de cauzalitate Granger are ipotezele:

H0: Nu avem cauzalitate Granger între date

H1: Avem cauzalitate Granger

```
> GrangerAbsolventi
$Granger

        Granger causality H0: absolventi_norm do not Granger-cause unitati_norm
        biblioteci_norm

data:  VAR object modelVar
F-Test = 1.1891, df1 = 10, df2 = 30, p-value = 0.3365


$Instant

        H0: No instantaneous causality between: absolventi_norm and unitati_norm
        biblioteci_norm

data:  VAR object modelVar
Chi-squared = 0.71261, df = 2, p-value = 0.7003
```

Figura 66 – Cauzalitate Granger pentru Absolvenți

Seria de date "Absolvenți" nu prezintă cauzalitate Granger în raport cu seria numărului de biblioteci și numărul de unități de învățământ, întrucât p - value este mai mare decât pragul de 0.1, atât pentru testul Granger, cât și pentru testul Volt (Instant). Acest rezultat sugerează că seria de date "Absolvenți" nu oferă o contribuție semnificativă în predicția celorlalte serii.

```
> GrangerUnitati
$Granger

        Granger causality H0: unitati_norm do not Granger-cause absolventi_norm
        biblioteci_norm

data:  VAR object modelVar
F-Test = 6.9196, df1 = 10, df2 = 30, p-value = 1.66e-05


$Instant

        H0: No instantaneous causality between: unitati_norm and absolventi_norm
        biblioteci_norm

data:  VAR object modelVar
Chi-squared = 6.7045, df = 2, p-value = 0.035
```

Figura 67 – Cauzalitate Granger pentru Unități

Observăm că în ambele p-value este mai mic decât 0.1, ceea ce indică faptul că respingem ipoteza nulă. Acest lucru sugerează că există o relație de cauzalitate Granger între numărul de unități de învățământ și numărul de absolvenți/ numărul de biblioteci. Cu alte cuvinte, valorile trecute ale numărului de unități de învățământ sunt utile pentru a prezice valorile viitoare ale numărului de absolvenți sau biblioteci.

```
> GrangerBiblioteci
$Granger

        Granger causality H0: biblioteci_norm do not Granger-cause absolventi_norm
        unitati_norm

data:  VAR object modelVar
F-Test = 0.60976, df1 = 10, df2 = 30, p-value = 0.7932


$Instant

        H0: No instantaneous causality between: biblioteci_norm and absolventi_norm
        unitati_norm

data:  VAR object modelVar
Chi-squared = 6.4974, df = 2, p-value = 0.03882
```

Figura 68 – Cauzalitate Granger pentru Bilbioteci

Rezultatul testului de cauzalitate Granger arată că nu există dovezi statistice suficiente pentru a concluziona că numărul de biblioteci "Granger - cauzează" numărul de absolvenți sau unități de învățământ. Cu alte cuvinte, valorile trecute ale numărului de biblioteci nu sunt

predictori semnificativi pentru valorile viitoare ale numărului de absolvenți sau unităților de învățământ.

Rezultatul testului Volt indică existența unei relații de cauzalitate instantanee între numărul de biblioteci și numărul de absolvenți / unități de învățământ. Aceasta înseamnă că schimbările simultane în numărul de biblioteci sunt asociate cu schimbările în numărul de absolvenți și unități de învățământ.

În ansamblu, rezultatele sugerează o cauzalitate parțială și la limită între numărul de biblioteci și numărul de absolvenți / unități de învățământ.

O modalitate eficientă de testare a efectului pe care o variabilă din setul de date îl are asupra alteia este realizarea funcției de răspuns la impuls. Pentru această analiză vom considera două scenarii diferite, întrucât atât unitățile de învățământ, cât și bibliotecile influențează în diferite măsuri numărul elevilor sau studenților care încheie unul din ciclurile academice. Astfel, putem vedea cum ar evolua starea actuală a absolvenților în cazul în care s-ar insera un influx de instituții de natură academică în infrastructura țării, atât pe termen scurt, cât și în anii următori.

1. Efectul unui impuls de absolvenți asupra numărului de unități de învățământ

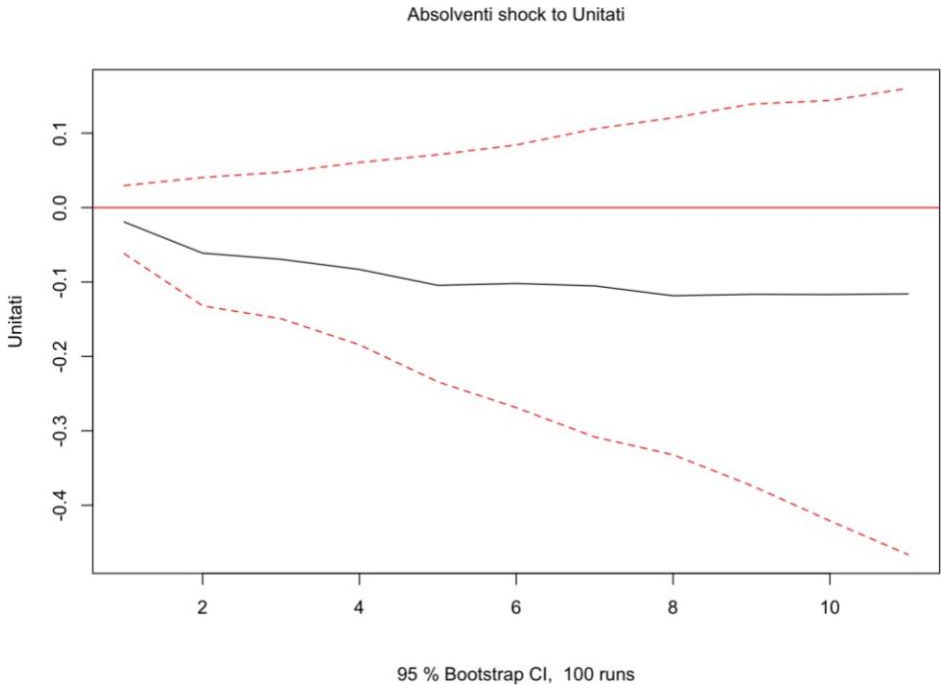


Figura 69 - Impuls absolvenți-unități învățământ

Conform funcției generate, un impuls pozitiv din partea variabilei *absolventi*, materializat prin rezultate bune la nivel național, va genera o scădere semnificativă a numărului de unități de învățământ, posibil datorită ajustării pieței la cerințele elevilor. Deși o astfel de schimbare

pozitivă ar trebui să inspire încredere în potențialul unei noi generații pe piața muncii, se observă o previziune prioritar negativă, cu o medie constant sub limita de 0.0.

2. Efectul unui impuls al numărului de biblioteci la nivel național asupra numărului de unități de învățământ

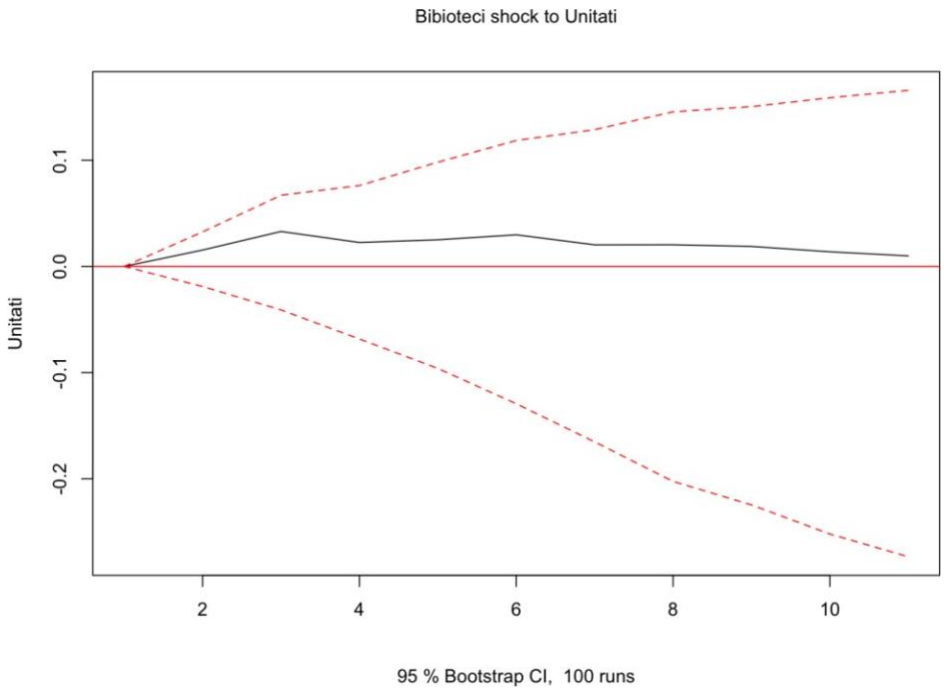


Figura 70 - Impuls biblioteci-unități învățământ

Cazul în care numărul de biblioteci naționale ar suferi o creștere accelerată generează un răspuns prioritar pozitiv asupra numărului de unități de învățământ. Se observă o perioadă incipientă de creștere mai accelerată, urmată de o scădere treptată, însă cu valori în medie peste nivelul de 0.0 pe toată durata prognozării. O astfel de schimbare ar putea fi datorată unui impuls la nivel guvernamental generat de interesul sporit al populației pentru zona academică, însă fără schimbări sau investiții semnificative.

Un alt mod de determinare a interdependenței acestor trei variabile este prin analiza descompunerii varianței erorii de prognoză (FEVD - Forecast Error Variance Decomposition), care determină compoziția varianței fiecărui indicator raportat la toți trei indicatorii pe un orizont de timp realizat din 10 perioade diferite.

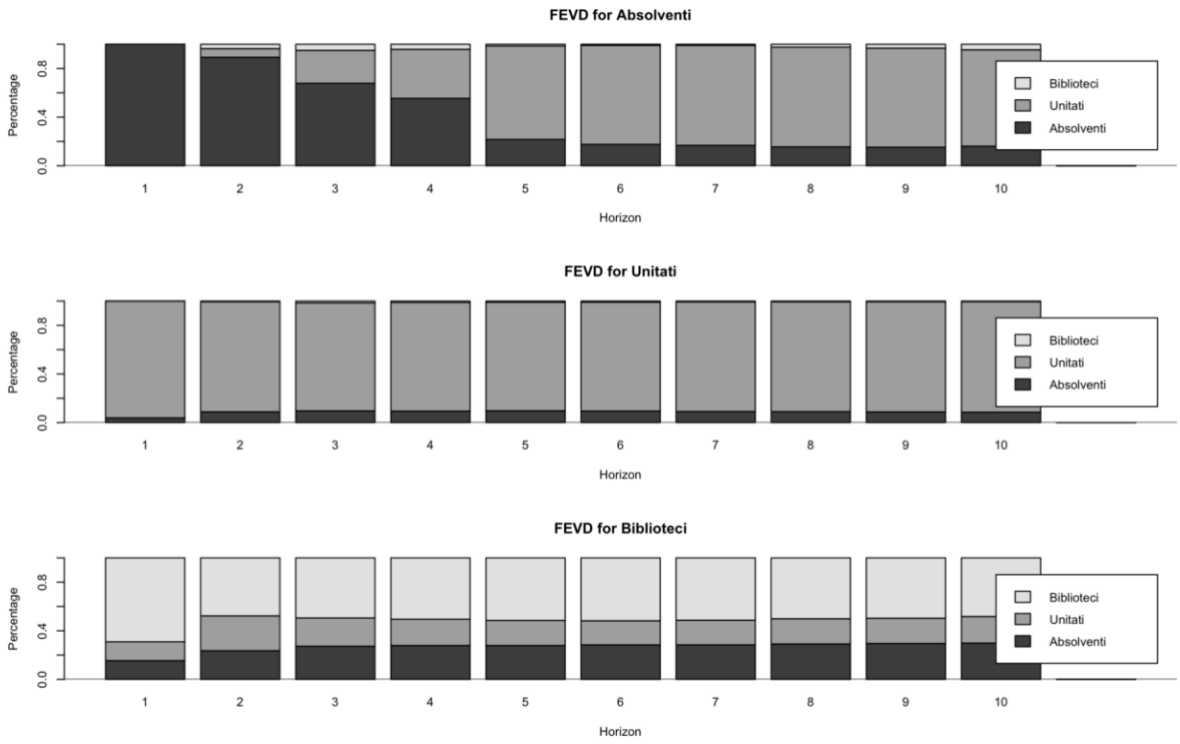


Figura 71 - FEVD

Conform reprezentării grafice de mai sus, putem extrage câteva observații relevante în înțelegerea modului în care una sau mai multe din variabile influențează parcursul celeilalte:

1. Varianța variabilei Absolventi este explicată inițial în totalitatea de propriile șocuri, însă se observă cum devine din ce în ce mai dependentă de șocurile în materie de unități de învățământ, ajungând la un procentaj chiar mai mare decât cel propriu.
2. Variabila Unitati este ușor influențată de șocurile în materie de absolvenți, însă aproape deloc afectată de numărul de biblioteci.
3. Valoarea variației pentru variabila Biblioteci este reprezentată în procentaje similare de toți cei trei indicatori discutați, păstrând valori similare pe toată perioada generată.

Pornind de la informațiile din această analiză, putem constata faptul că numărul de biblioteci inaugurate la nivel național este dependent de mai mulți factori din zona academică, pe când stabilirea numărului de unități de învățământ este doar ușor ghidată de rezultatele elevilor și studenților.

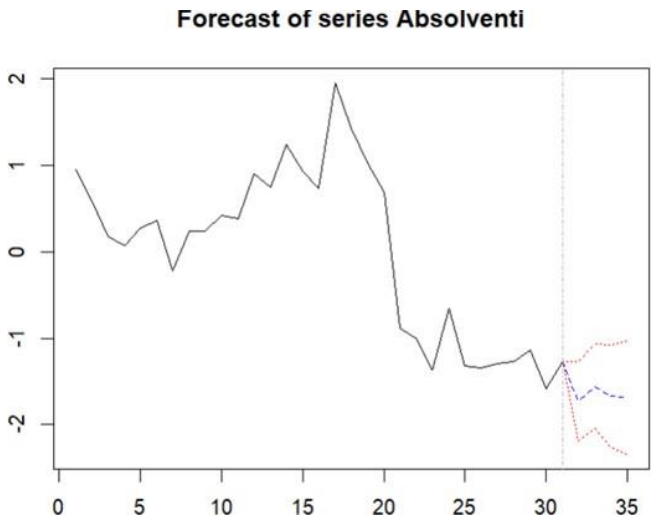


Figura 72 - Prognoze absolvenți

Graficul de mai sus prezintă prognoza seriei "Absolventi" pentru următorii 4 ani. Linia neagră arată datele istorice, iar prognoza începe după linia verticală punctată. Valorile estimate sunt indicate pe linia albastră punctată, în timp ce liniile roșii punctate marchează intervalele de încredere de 99% ale prognozei.

Prognoza inițială sugerează o tendință ușor ascendentă, urmată de o tendință ușor descendentă. Extinderea intervalelor de încredere pe perioada prognozată reflectă o creștere a incertitudinii privind valorile viitoare pe măsură ce timpul trece.

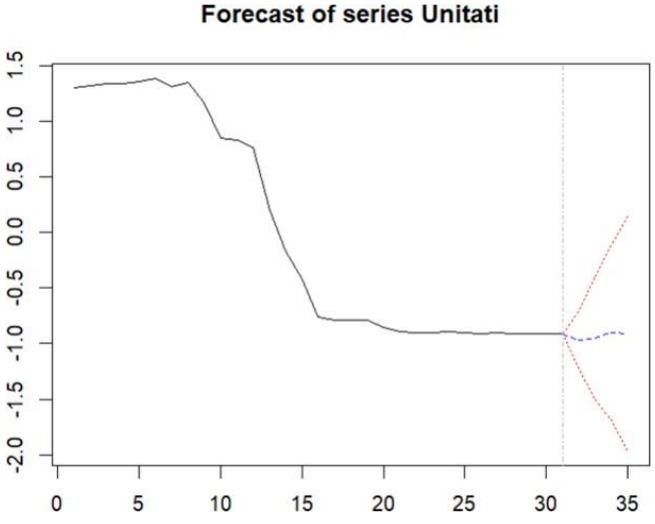


Figura 73 - Prognoză unități de învățământ

Prognoza seriei "Unitati" sugerează stabilitate în prima perioadă, fără tendințe clare de creștere sau scădere imediată. Asemănător prognozei seriei absolvenților, intervalele de încredere se lărgesc pe măsură ce timpul avansează, dar în seria unităților de învățământ, această lărgire este mai abruptă, indicând o incertitudine mai mare în prognozele pe termen lung.

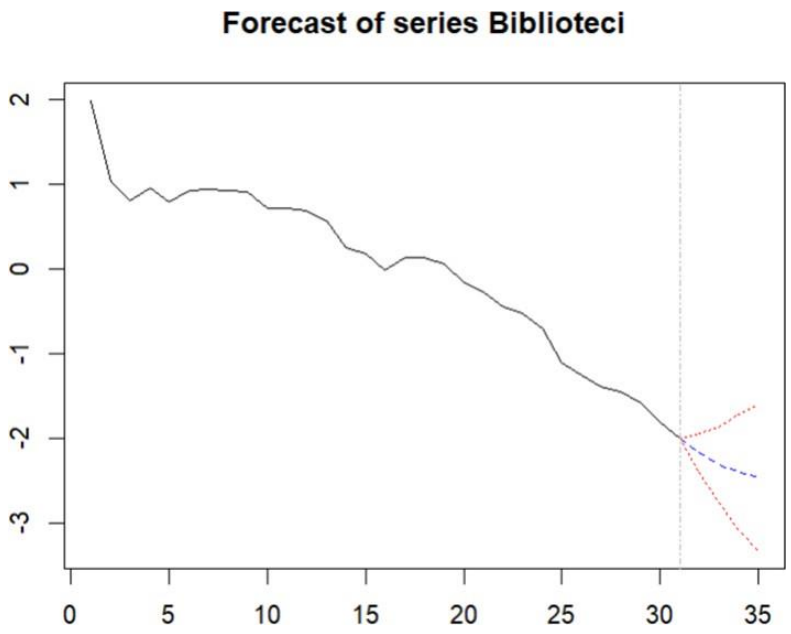


Figura 74 - Prognoză biblioteci

Mai sus avem reprezentată evoluția anuală a numărului de biblioteci de pe teritoriul României pentru perioada 1991 – 2022. Observăm că seria istorică urmează un trend descendent, fiind continuat și de valorile prognozate. Acest comportament se datorează mai multor factori cum ar fi: lipsa de investiții, scăderea interesului pentru lectură, apariția și dezvoltarea cărților electronice etc.

După cum am menționat și pentru prognozele seriilor “Absolvenți” și ”Unități”, liniile roșii punctate reprezintă intervalul de încredere de 99%.

Concluzie

În urma analizelor efectuate observăm că variabila "Absolvenți" devine din ce în ce mai dependentă de șocurile unităților de învățământ, sugerând o legătură puternică între acestea pe termen lung. Pe de altă parte, "Unități" este doar ușor influențată de numărul de absolvenți și aproape deloc de biblioteci, subliniind o influență minoră a performanțelor academice asupra deciziei de deschidere a noilor unități de învățământ. În ceea ce privește variabila "Biblioteci", valoarea variației este distribuită în mod egal între absolvenți, unități și biblioteci, indicând o dependență de mai mulți factori educaționali.

Observațiile privind evoluția anuală a numărului de biblioteci din România între 1991 și 2022 relevă un trend descendent, continuat și în valorile prognozate, atribuit lipsei de investiții, scăderii interesului pentru lectură și dezvoltării cărților electronice. În schimb, prognoza pentru "Unități" sugerează o stabilitate inițială, fără tendințe clare de creștere sau scădere, dar cu o incertitudine crescută pe termen lung. Aceleași tendințe de lărgire a intervalelor de încredere sunt observate și pentru "Absolvenți", indicând dificultăți în prognozarea exactă pe termen lung.

Bibliografie

Constantin Anghelache, C. D. (2013). Evoluția Produsului Intern Brut al României. Romanian Statistical Review, 86-93.

Constantin Anghelache, M. G. (2018). Evoluția centenară a învățământului din România.

(2021-2022). Raport privind starea învățământului preuniversitar din România. București.

Rejuwan Shanim & Dr, T. A. (n.d.). Machine Learning-Based NVIDIA Stock Price Prediction using Autots. Connecting Multidisciplinary Research for Universal Growth.

Tahir Sher, A. R. (n.d.). Exploiting Data Science for Measuring the Performance of Technology Stocks.

Zhenhao Yang, Z. W. (n.d.). The Research of NVIDIA Stock Price Prediction Based on LSTM And ARIMA Model.