**ACADEMIA DE STUDII ECONOMICE BUCUREȘTI**

**FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**

**Specializarea: Cibernetică economică**

****

**Modelarea stochastică bazată pe serii de timp, cu aplicații în hidrologie**

**~Lucrare de licență~**

**Coordonator științific: PROF. UNIV. DR. STANCU STELIAN**

**Absolvent: IFTIMIE ALEXANDRA-MARIA**

**BUCUREȘTI**

**2021-2022**

Cuprins:

[**1. Introducere** 3](#_Toc107415984)

[**2. Recenzia literaturii** 5](#_Toc107415985)

[**3. Metodologia cercetării** 9](#_Toc107415986)

[**4.1. Analiza evoluției debitului Dunării** 11](#_Toc107415987)

[**4.2. Studierea sezonalității debitului Dunării** 18](#_Toc107415988)

[**5.1. Analiza nivelului de clorofilă din Marea Neagră** 28](#_Toc107415989)

[**5.2. Regresie** 35](#_Toc107415990)

[**5.3. Anova** 36](#_Toc107415991)

[**6. Concluzii** 37](#_Toc107415992)

# **1. Introducere**

­­ Apa reprezintă o resursă extrem de importantă, iar calitatea vieții depinde strict de aceasta. Acest element joacă un rol extrem de important atât în natură, precum și în economie, fiind prezentă în industrii precum navigație și transport, agricultură, turism sau producția energiei electrice, cu ajutorul hidrocentralelor ce folosesc puterea apei ca materie primă.

Confucius era de părere că „trebuie să studiezi trecutul, dacă vrei să cunoști viitorul” (Confucius, fără an). Acest citat reprezintă unul dintre obiectivele majore al analizei seriilor de timp, găsirea și utilizarea unui model stochastic de un anumit tip, capabil să reproducă comportamentul viitor al unei serii pe baza observațiilor din trecut.

Tema abordată în această lucrare, intitulată „Modelarea stochastică bazată pe serii de timp, cu aplicații în hidrologie” reprezintă una de actualitate în contextul crizei apei. Subiectul este unul de interes atât pentru hidrologi, cât și cei a căror activitate economică depinde direct de sursele de apă naturale, precum râurile sau lacurile. Motivația alegerii acestei teme are la bază dorința personală de a studia evoluția în timp a unei serii hidrologice, întrucât acest subiect este unul puțin abordat în spațiul românesc.

Obiectivul primei părți a studiului de caz din cadrul acestei lucrări este de a studia evoluția debitelor zilnice ale fluviului Dunărea și găsirea celui mai bun model care are abilitatea de a înțelege stările hidrologice ale fluviului și să previzioneze valorilor viitoare cu o acuratețe cât mai mare.

Scopul celei de-a doua părți al acestui studiu este de a analiza modificările nivelului de clorofilă din apele Mării Negre pe teritoriul țării, obținerea unui model stochastic cu cele mai mici erori de previziune și investigarea fenomenului de variație a clorofilei, utilizând instrumente precum testul anova și regresii unifactoriale și multifactoriale. Analiza de regresie utilizată pornește de la ipoteza că nivelul clorofilei în apele marine este influențat de aciditatea apei, nivelul hidrocarburilor, nivelul metelelor grele prezente în apă, a fertilizatorilor utilizați în agricultură care s-au scurs în mare și a deșeurilor marine, ce vor fi utilizate ca variabile independente.

Studiul este organizat în cinci părți principale.

Prima parte, este cea care pune bazele cadrului conceptual și a principalelor noțiuni teoretice necesare pentru înțelegerea lucrării. Tot aici se regăsesc descrise principalele studii ce abordează o temă asemănătoare.

A doua parte ilustrează metodologia utilă pentru realizarea analizei, în scopul aplicării unor modele stochastice, a testului ANOVA și a modelelor de regresie unifactorială și multifactorială.

Partea a treia prezintă analiza seriei de timp a fluviului Dunărea, set de date cuprins din 1236 de valori zilnice în intervalul 1 ianuarie 2019-20 mai 2022. Se va analiza staționaritatea seriei, sezonalitatea acesteia și se va căuta modelul care produce cele mai mici erori pentru a explica evoluția în timp a fluviului. De asemenea, se va studia dacă modificări aduse seriei influențează sezonalitatea acesteia.

Penultima parte expune cercetarea efectuată asupra nivelului clorofilei din mare, set de date cuprins din 619 de valori bilunare în intervalul 1 ianuarie 1995-1 octombrie 2020. Se va analiza staționaritatea și sezonalitatea seriei și modelul optim pentru previziune. De asemenea, tot în cadrul acestei părți se vor prezenta rezultatele aplicării modelelor de regresie unifactoriale și multifactoriale și a testului ANOVA pe setul de date al nivelului clorofilei, în scopul de a înțelege evoluția acestuia.

Ultima parte a lucrării vine ca o încheiere a studiului, prezentând concluziile și limitele analizei.

# 

# **2. Recenzia literaturii**

În cadrul prezentului capitol voi descrie principalele noțiuni teoretice relevante în ansamblul lucrării și voi indica în mod succint o selecție de studii științifice reprezentative pentru tema abordată.

Un proces stochastic, așa cum îl descrie Haidu Ion în lucrarea sa intitulată „Analiza seriilor de timp Aplicații în hidrologie” reprezintă „o mulțime de realizări ale unui element, realizări ce implică unele incertitudini și țin cont de timp” (Haidu, 1997). Comparat cu procesul determinist, care presupune o singură variabilă sau realizare posibilă, procesul stochastic reprezintă opusul celui dintâi, implicând în evoluția acestuia incertitudini; din acest motiv, procesul stochastic poate fi regăsit și sub denumirea de proces aleator.

Modelarea unui proces determinist, numită și modelare deterministă, se caracterizează prin obținerea aceluiași rezultat la aplicarea modelului, indiferent de numărul de aplicări. Un exemplu concludent în acest sens îl reprezintă regulile matematice de calcul. Spre deosebire de modelarea deterministă, modelarea proceselor stochastice, sau modelarea stochastică, implică factorul aleatoriu, ceea ce poate produce schimbări în rezultate la reaplicarea modelului.

Modelarea stochastică își arată importanța în domenii precum biologie, fizică cuantică, hidrologie, sau în alte industrii, spre exemplu industriile contabilo-financiare și ale asigurărilor, care își propun să prognozeze schimbările posibile ale bilanțului contabil al firmei.

Modelarea stochastică în domeniul hidrologiei își conturează utilitatea prin numeroasele influențe benefice asupra diverselor sectoare economice. Luând în analiză de exemplu fluviul Dunărea, pe teritoriul țării noastre, a utiliza modele stochastice pentru prognozarea debitelor poate avea efecte asupra mai multor sectoare, precum:

* producerea energiei electrice: datorate barajelor de la Porțile de Fier I și II care utilizează apa Dunării ca pe o sursă de hidroenergie;
* agricultură: Dunărea reprezintă o sursă de apă pentru industrii, jucând un rol major în producția agricolă;
* transporturi și navigație: fluviul reprezintă un important drum navigabil, iar așa cum specifică cei de la Fundația „World Wildlife” în studiul lor de caz, „condiţiile de navigaţie pe porţiunea de la gura Djerdap, pe Dunăre, cândva foarte periculoase, au fost îmbunătăţite radical datorită construcției barajelor de la Porțile de Fier I și II, iar navigaţia se desfăşoară pe tot parcursul anului” (Hulea & Ionescu, fără an);
* turistic;
* pescuit;
* construcția de vapoare.

A cunoaște valorile viitoare ale debitului Dunării poate duce la alegeri cu impact benefic din punct de vedere economic. Ca o ilustrare a acestui fapt, voi propune următoare situație: după o perioadă de secetă, se prognozează o perioadă în care debitul Dunării va crește, ceea ce va duce la creșterea numărului de turiști în zona Cazanelor Dunării. Astfel, managerii din domeniul turistic pot lua decizii și schimbări financiare ținând cont de numărul crescut de turiști.

Conform lui Haidu, „conceptul stochastic de modelare a unui râu presupune faptul că un anume tip de model reprezintă cele mai relevante caracteristici statistice ale seriei istorice, iar cele mai utilizate modele sunt cele autoregresive și cele mixte autoregresive și de medie mobilă” (Haidu, 1997).

Un proces autoregresiv de ordin p, notat și AR(p), reprezintă un proces unde valoarea actuală a lui *y* este influențată de propriile valori din trecut, și o perturbație εt, numită și inovație.

Ecuația unui proces autoregresiv este:

(1)

Ordinul p al procesului autoregresiv se determină pe baza funcției de autocorelație parțială. Coeficientul de corelație parțială de ordin k, sau funcția de autocorelație parțială, măsoară efectul direct al lui *yt-k* asupra lui *yt*.

Un proces de medie mobilă de ordin q, notat și MA(q), reprezintă un proces unde valoarea curentă a lui  *y* este influențată atât de valoarea contemporană cât și de valorile din trecut ale termenului de inovație εt.

Ecuația unui proces de medie mobilă este:

(2)

Ordinul q al procesului de medie mobilă se determină pe baza funcției de autocorelație. Coeficientul de autocorelație de ordin k (la lag-ul k), sau funcția de autocorelație, măsoară corelația dintre *yt* și valorile sale anterioare *yt-k*.

Un model autoregresiv de medie mobilă, notat și ARMA(p,q), are o componentă autoregresivă și o componentă medie mobilă.

Ecuația modelului autoregresiv de medie mobilă este:

(3)

Ordinul p al procesului autoregresiv se determină pe baza funcției de autocorelație parțială , în timp ce ordinul q al procesului de medie mobilă se determină pe baza funcției de autocorelație.

Davidescu Adriana a afirmat în cadrul cursului său că „ seria de timp staționară este acea serie ale cărei valori oscilează, mai mult sau mai puțin aleator, în jurul unui nivel de referință, media fiind deci într-o stare de echilibru” (Davidescu, 2022). Astfel, dacă seria nu este staționară, se impune staționarizarea funcției prin procese precum diferențierea. În consecință, a apărut modelul autoregresiv de medie mobilă integrat, notat ARIMA(p,i,q), unde I definește diferențierea seriei, iar parametrul i reprezintă numărul de diferențieri aplicate seriei pentru e deveni staționară.

Ecuația modelului ARIMA(p,i,q) este:

(4)

Sezonalitatea unei serii de date poate fi definită prin abateri de la medie care revin sistematic, și apare datorită faptului că datele au un caracter ciclic. Exemple concludente în acest sens sunt producția lunară de produse agricole, care va prezenta valori ridicate pe perioada verii, sau valorile lunare ale numărului persoanelor care închiriază un apartament în orașul Cluj-Napoca, unde se vor observa niveluri ridicate pe perioada verii și începutul toamnei, datorită festivalurilor de muzică și a începutului anului universitar. Prin urmare, s-a dorit realizarea unui model care ia în considerare sezonalitatea seriei studiate, astfel apărând modelul SARIMA, unde S provine de la sezonalitate. Modelul SARIMA are patru componente: de tip autoregresiv (AR), de tip medie mobilă (MA), de tip autoregresiv sezonieră (SAR) și de tip medie mobilă sezonieră (SMA). Primii trei parametri ai modelului SARIMA, notat SARIMA(p,i,q)(P,D,Q)[m], se determină ca la un model ARIMA, în timp ce P al procesului autoregresiv sezonier se determină pe baza funcției de autocorelație parțială, iar Q al procesului de medie mobilă se determină pe baza funcției de autocorelație, ambele utilizând lag-urile sezoniere. Termenul m reprezintă numărul de observații care definește perioada de sezonalitate, în timp ce D reprezintă numărul diferențierilor sezoniere aplicate seriei pentru a deveni staționară.

Ecuația modelului SARIMA(p,i,q)(P,D,Q)[m]este:

(5)

Lucrările statistice care abordează tema seriilor de timp cu aplicație în hidrologie se împart în două direcții, în funcție de tipul de apă studiată, curgătoare sau stătătoare. Astfel, autorii fie analizează debitul unui râu sau fluviu, fie starea hidrologică a unui lac. În lucrarea lor despre modelarea stochastică a seriei de timp ce caracterizează stările hidrologice ale râului Elan, Lawrance și Kottegoda evidențiază importanța luării în considerare a sezonalității în modelarea unei serii hidrologice (Lawrance & Kottegoda, 1997).

Mirzavand și Ghazavi au abordat modelarea stochastică a seriilor de timp hidrologice în mediile aride, întrucât „în ultimele decenii, nivelul apelor subterane a scăzut din cauza cererii tot mai mari de apă, a managementului slab al irigațiilor și deteriorarea solului. Pentru gestionarea eficientă a apelor subterane, este important să se modeleze și să se prezică fluctuațiile nivelurilor apelor subterane” (Mirzavand & Ghazavi, 2015). Aceștia au aplicat mai multe modele stochastice, autoregresiv (AR), de medie mobilă (MA), autoregresiv de medie mobilă (ARMA), autoregresiv de medie mobilă integrat (ARIMA) și autoregresiv de medie mobilă integrat cu evoluții sezoniere (SARIMA), pe un set de date compus din observații ale nivelului apei subterane din provincia Isfahan, Iran, și au concluziat că cel mai bun model pentru previziune este un model autoregresiv de ordin 2, AR(2).

În încercarea de a prognoza afluxul lacului de acumulare a barajului Dez, Valipour, Banihabib și Behbahani au comparat modelele ARMA și ARIMA. Aceștia au antrenat modelel pe un set de date constituit din date lunare în perioada 1960-2002, crescând parametri modelelor pentru a găsi varianta optimă ce caracterizează evoluția seriei în măsura cea mai mare, iar setul de testare a fost compus din perioada 2003-2007. Confruntând eroarea medie pătratică a fiecărui model, autorii au concluzionat că modelul ARIMA are o eroare mai mică decât modelul ARMA, ceea ce înseamnă că prognozele modelului autoregresiv de medie mobilă integrat sunt mult mai apropiate de valorile reale decât valorile obținute prin modelul ARMA (Valipour, et al., 2013).

Aceeași temă, a modelării unei serii hidrologice utilizând ARIMA, au abordat-o atât B.N. Ghimire (Ghimire, 2017) în lucrarea sa despre dezvoltarea modelului ARIMA pentru două stații hidrologice ale unui râu din Statele Unite ale Americii, cât și Z. Wang și Y. Lou (Wang & Lou, 2019) care și-au propus să prognozeze evoluțiile „nivelului mediu zilnic al apei unei stații hidrologice din bazinul râului Chuhe”. Ambele studii au atins aceeași concluzie: modelul ARIMA a putut fi utilizat cu succes pentru previziunile unei serii hidrologice.

Prognoza perioadelor de secetă este o tema abordată de mulți autori, în special în zonele aride, unde sursele de apă sunt puține și calitatea vieții depinde strict de nivelul acesteia. Atât El-Shafie et al (El-Shafie, et al., 2020), cât și Bazrafshan et al (Bazrafshan, et al., 2015) și-au propus să prezică perioadele de secetă în lucrările lor, cei din urmă ajungând la concluzia că modelul cu evoluții sezoniere lunare SARIMA se pretează cel mai bine cauzei lor.

În cadrul celei de-a 5-a Conferințe Internaționale privind resursele de apă și mediu, ce a avut loc în Macao, China în 2019, C. A. Mocanu-Vargancsik și A. Barbulescu au studiat evoluția debitului de apă a râului Buzău și a schimbărilor sezoniere ale acestuia, înainte și după construirea Barajului Siriu (Mocanu-Vargancsik & Barbulescu, 2019). Cei doi au evidențiat că utilizarea modelului AR(5) se potrivește cel mai bine evoluției râului Buzău.

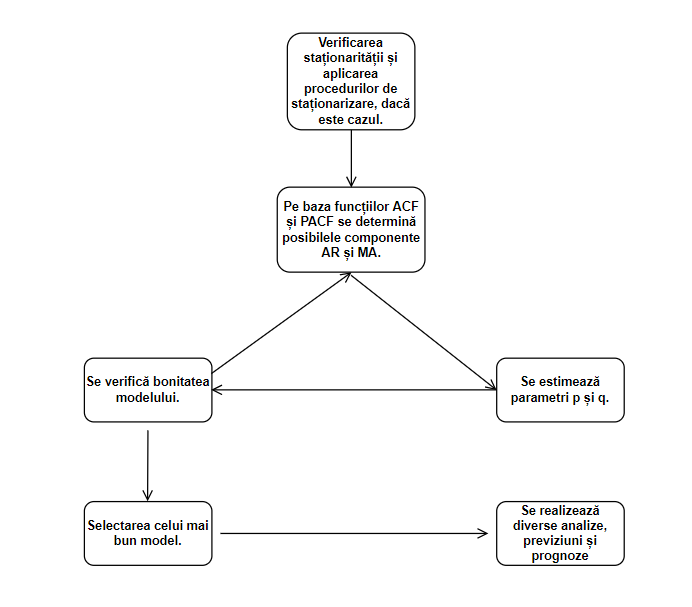
Haidu, în lucrarea sa despre analiza seriilor de timp în hidrologie, „pornește de la un bazin hidrografic, considerat ca sistem, și studiază cum anume input-ul, precipitațiile anuale bazinale, sunt transformate de către bazinul hidrografic în output, debitele medii anuale” (Haidu, 1997), utilizând spre exemplificare modelele autoregresiv (AR), de medie mobilă (MA), autoregresiv de medie mobilă (ARMA) și autoregresiv de medie mobilă integrat (ARIMA).

În plus față de predicția stărilor hidrologice ale unui râu sau lac, unii autori și-au propus să studieze evoluția nivelului de clorofilă în apele oceanice (Rousseaux, et al., 2021) și s-a constatat că valorile reale s-au încadrat în intervalul de încredere al previziunii, cu trei excepții zona nordică a oceanelor Atlantic și al Indiei și zona sudică a Oceanului Pacific. Concentrația de clorofilă din apă poate fi considerat un indicator în studiul sănătății mărilor și oceanelor și prognoza acestei concentrații poate ajuta în managementul pescuitului.

# **3. Metodologia cercetării**

Metodologia de cercetare a implicat culegerea datelor necesare studiului de caz. În ceea ce privește prima parte a acestui studiu, datele au fost culese de pe site-urile Institutului Național de Hidrologie și Gospodărire a Apelor (link-ul către site se găsește [aici](http://www.inhga.ro/)) și al Administrației Naționale „Apele Române” (adresa site-ului se găsește [aici](https://rowater.ro/)). Colectarea datelor a avut drept rezultat obținerea unui set de date cuprins din 1236 de valori zilnice în intervalul 1 ianuarie 2019-20 mai 2022, ce caracterizează debitelul Dunării la intrarea în țară, pe secțiunea Baziaș, măsurat zilnic la ora 700 și exprimat în m3/s.

Acestui set de date i s-a aplicat metoda Box-Jenkins, creată de cei doi statisticieni George Box și Gwilyn Jenkins, pentru estimarea unui model de tip autoregresiv de medie mobilă (ARMA) sau autoregresiv de medie mobilă integrat (ARIMA). Grafic, etapele metodei sunt redate în figura următoare:



Figură 1-Schema logică a metodei Box-Jenkins

*Sursa: Prelucrarea autorului în Visual Paradigm*

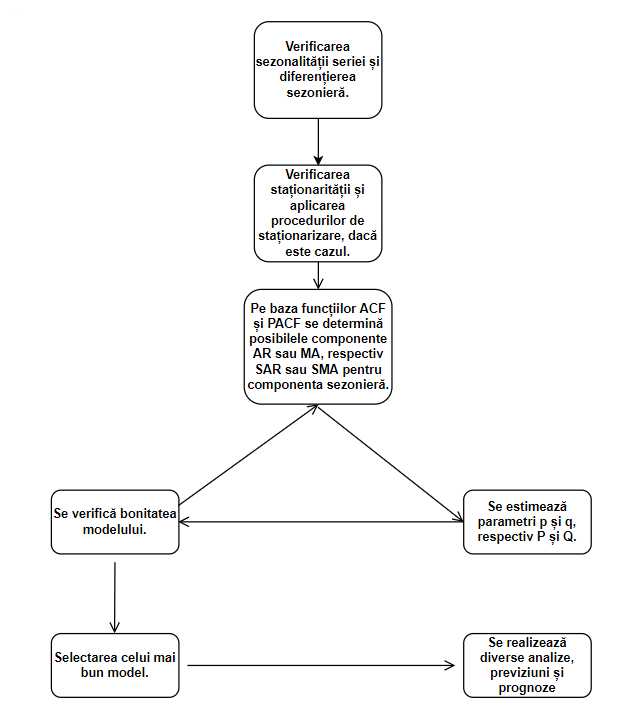
Metodologia Box-Jenkins se aplică seriilor staționare (în cazul în care seria nu prezintă staționaritate, se aplică diverse proceduri pentru a deveni staționară) și este compusă din mai multe etape:

1. Verificarea staționarității seriei și inducerea acesteia dacă este cazul.
2. Identificarea modelulului: pe baza funcțiilor de autocorelație (ACF) și autocorelație parțială (PACF) se vor alege posibililele componente AR sau MA ce vor constitui modelul.
3. Estimarea parametrilor p și q care construiesc cel mai bun model.
4. Verificarea bonității modelului: se cercetează diverse criterii de validitate și performanță, precum și dacă reziduurile modelului sunt zgomot alb, acestea trebuie să nu fie autocorelate și să aibe media și varianța nule. În situația în care modelul ales nu prezintă caracteristicile dorite, se va relua algoritmul de la pasul 1 pentru estimarea unui model nou.
5. Se selectează modelul care îndeplinește condițiile necesare.
6. Pe baza modelului ales se fac diverse preziviuni și prognoze.

De asemenea, s-a cercetat posibilitatea estimării evoluției setului de date utilizând un model de tip autoregresiv medie mobilă pentru evoluții sezoniere (SARIMA). Pașii pentru aplicarea unui astfel de model sunt următorii:

1. Verificarea sezonalității seriei și diferențierea sezonieră.
2. Verificarea staționarității seriei și inducerea acesteia dacă este cazul.
3. Identificarea modelulului: pe baza funcțiilor de autocorelație (ACF) și autocorelație parțială (PACF) se vor alege posibililele componente AR sau MA, respectiv SAR sau SMA pentru componenta sezonieră, ce vor constitui modelul.
4. Estimarea parametrilor p și q, respectiv P și Q, care construiesc cel mai bun model.
5. Verificarea bonității modelului: se cercetează diverse criterii de validitate și performanță, precum și dacă reziduurile modelului sunt zgomot alb, acestea trebuie să nu fie autocorelate și să aibe media și varianța nule. În situația în care modelul ales nu prezintă caracteristicile dorite, se va relua algoritmul de la pasul 1 pentru estimarea unui model nou.
6. Se selectează modelul care îndeplinește condițiile necesare.
7. Pe baza modelului ales se fac diverse preziviuni și prognoze.

Grafic, pașii aplicării modelului SARIMA sunt redați în figura următoare:



Figură 2-Schema logică a modelului SARIMA

*Sursa: Prelucrarea autorului în Visual Paradigm*

În ceea ce privește a doua parte a studiului, datele au fost culese de pe site-ul „[European Marine Observation and Data Network](https://emodnet.ec.europa.eu/en)” (link-ul către site se găsește [aici](https://emodnet-chemistry.maris.nl/search)). Colectarea datelor a avut drept rezultat obținerea a două seturi de date: primul cuprinde 619 de valori bilunare în intervalul 1 ianuarie 1995-1 octombrie 2020, în timp ce al doilea este alcătuit din 79 de observații zilnice, ambele caracterizând nivelul clorofilei din apa marină. Primul set de date va fi utilizat pentru a studia evoluția clorofilei aplicând modele de serii de timp, în timp ce pe al doilea set se va studia relația dintre nivelul de clorofilă și aciditatea apei, nivelul hidrocarburilor, nivelul metelelor grele, a fertilizatorilor utilizați în agricultură și a deșeurilor marine. În acest scop se vor exprima modelele de regresie și se vor verifica parametri din punct de vedere al reprezentativității. În situația în care aceștia sunt valizi, se va verifica bonitatea modelului și reziduurile acestuia. De asemenea, utilizând valorile zilnice se va aplica un test ANOVA pentru studierea existenței unor diferențe semnificative în medie, din perspectiva celor patru zone geografice maritime, nord, sud, est, vest, de pe teritoriul țării.

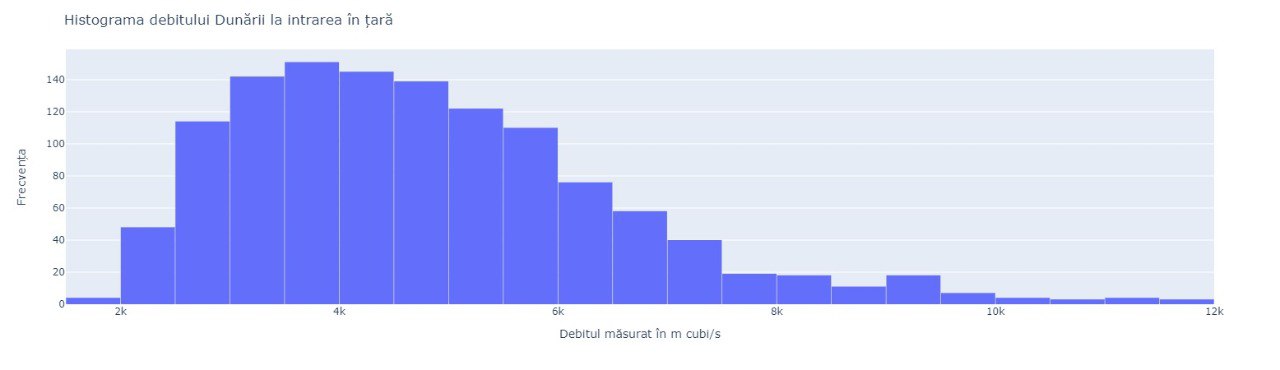
# **4.1. Analiza evoluției debitului Dunării**

**Tabel 1. Statisticile descriptive ale setului de date**

|  |  |
| --- | --- |
| **Statisticile descriptive ale debitelor medii, măsurate în m3/s** | |
| Mean | 4797,66 |
| Standard Error | 166,63 |
| Median | 4460 |
| Mode | 4310 |
| Standard Deviation | 1739,72 |
| Sample Variance | 3026618,32 |
| Kurtosis | 0,91 |
| Skewness | 0,95 |
| Range | 8245 |
| Minimum | 2015 |
| Maximum | 10260 |

*Sursa: Prelucrarea autorului în Microsoft Excel, pe baza datelor culese*

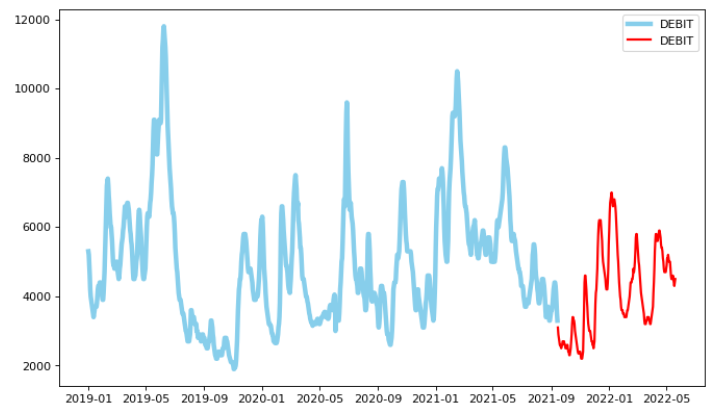
În tabelul 1 este ilustrată descrierea statistică a seriei de date, cu ajutorul căreia putem caracteriza setul de date. Se observă în primul rând indicatorii tendinței centrale, media, mediana și modul. Aceștia rezumă într-o măsură ridicată datele și sunt strâns legați de centrul setului de date. Media are o valoare aproximativă de 4800 m3/s, mai mare decât valorile medianei și a modului, de 4460 m3/s, respectiv 4310 m3/s. Acest lucru ilustrează o asimetrie la stânga, pozitivă, fapt confirmat și de coeficientul skewness cu valoarea de 0,95. În al doilea rând, putem caracteriza distribuția datelor, utilizând coeficienții skewness și kurtosis. Așa cum am specificat anterior, asimetria este una pozitivă, la stânga, iar boltirea este una leptokurtică. În ceea ce privește indicatorii variației datelor, se observă că amplitudinea are o valoare aproximativ dublă față de medie, în consecință putem trage concluzia că datele au o împrăștiere mare față de medie, acestea abătându-se cu aproximativ 1740 m3/s.



Figură 3-Histograma debitului Dunării pe secțiunea Baziaș, măsurat în m3/s

*Sursa: Prelucrarea autorului în Visual Studio Code, pe baza datelor culese*

Setul de date compus din valori zilnice a fost împărțit în set de antrenare și set de testare, în proporție de 80-20 %. Primul set, ilustrat în culoare albastră,va fi utilizat pentru definirea modelului și al doilea, ilustrat în culoare roșie, va fi folosit pentru a testa calitatea acestuia. Cele două seturi și împărțirea acestora se vede în graficul următor:



Figură 4-Împărțirea setului de date în set de antrenare și set de testare

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

Pentru a putea aplica un proces autoregresiv sau de medie mobilă unui set de date, trebuie în primul rând verificată staționaritatea seriei. În acest scop, voi utiliza testele Augmented Dickey-Fuller, Phillips-Perron și Kwiatkowski–Phillips–Schmidt–Shin (KPSS). Ipotezele acestor teste sunt următoarele:

* Augmented Dickey-Fuller:

H0: seria are rădăcină unitară și este nestaționară.

H1: seria este staționară.

* Phillips-Perron:

H0: seria are rădăcină unitară și este nestaționară.

H1: seria este staționară.

* KPSS:

H0: seria este staționară.

H1: seria are rădăcină unitară și este nestaționară.

În situația în care valoarea absolută a testului statistic calculat este mai mică decât valoare tabelată în modul sau valoarea probabilității p-value este mai mare decât nivelul de încredere, se va accepta ipoteza nulă H0 și se va respinge ipoteza alternativă H1.

În urma aplicării celor 3 teste statistice asupra seriei de date, au rezultat următoarele output-uri, prezentate în tabelul alăturat:

**Tabel 4.2. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 2,2e-16 | -1,83 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,05 | 0,05396 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 0,002 | -27,54 | -2,56 | -2,86 | -3,43 |

*Sursa:* Prelucrarea autorului utilizând RStudio

Analizând tabelul output-urilor testelor statistice pentru verificarea staționarității se poate concluziona că seria este staționară și nu necesită diferențiere. Astfel, putem aplica modelele AR, MA, și ARMA seriei de date. În continuare trebuie verificată sezonalitatea seriei, pentru a putea decide dacă seria trebuie ajustată sezonier și dacă putem aplica un model de tip SARIMA datelor.

În ceea ce privește textul Hegy, acesta are următoarele ipoteze:

H0: seria are rădăcină sezonieră unitară și prezintă sezonalitate.

H1: seria este nesezonieră.

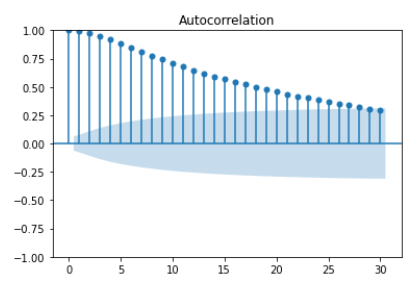
În situația în care valoarea absolută a testului statistic calculat este mai mică decât valoare tabelată în modul sau valoarea probabilității p-value este mai mare decât nivelul de încredere, se va accepta ipoteza nulă H0 și se va respinge ipoteza alternativă H1.

**Tabel 2. Output-urile testului Hegy pentru verificarea sezonalității seriei inițiale**

|  |  |
| --- | --- |
| **Testul utilizat** | **Valoare probabilității**  **p-value** |
| **Hegy** | 0,009 |

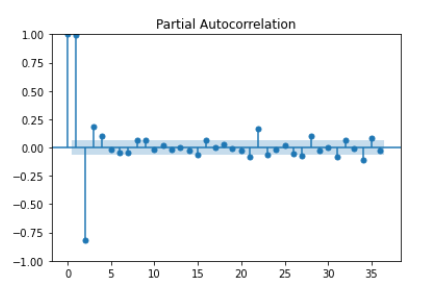
*Sursa:* Prelucrarea autorului utilizând RStudio

Prin utilizarea testului, se confirmă faptul că seria nu prezintă sezonalitate, ceea ce înseamnă că nu se poate aplica un model de tip SARIMA setului de date. Pentru a putea determina ordinul proceselor AR, MA și ARMA, trebuie studiată funcția de autocorelație și funcția de autocorelație parțială.



Figură 5-Funcția de autocorelație

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese



Figură 6- Funcția de autocorelație parțială

*Sursa: Prelucrarea autorului în Visual Studio Code, pe baza datelor culese*

În mediul de programare RStudio au fost încercate următoarele modele: AR(1), AR(2), AR(3), AR(4), AR(5), MA(1), MA(2), MA(3), MA(4), MA(5), ARMA(1,1), ARMA(1,2), ARMA(1,3), ARMA(1,4), ARMA(1,5), ARMA(2,1), ARMA(2,2), ARMA(2,3), ARMA(2,4), ARMA(2,5), ARMA(3,1), ARMA(3,2), ARMA(3,3), ARMA(3,4), ARMA(3,5). Dintre aceste modele, doar câteva au prezentat parametri semnificativi. Descrierile succinte ale output-utilor modelelor se regăsesc în tabelul următor.

**Tabel 4. Principalele criterii pentru alegerea celui mai bun model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modele** | **Numărul parametrilor modelului** | **Parametri semnificativi** | **Criteriul Akaike** | **Criteriul Schwartz** |
| **AR(1)** | 1 | 1 | 13445,98 | 13460,67 |
| **AR(2)** | 2 | 2 | 12593,57 | 12613,16 |
| **MA(1)** | 1 | 1 | 16341,89 | 16356,58 |
| **MA(2)** | 2 | 2 | 15280,62 | 15300,21 |
| **MA(3)** | 3 | 3 | 14538,77 | 14563,25 |
| **MA(4)** | 4 | 4 | 13997,52 | 14026,9 |
| **MA(5)** | 5 | 5 | 13600,11 | 13634,39 |
| **ARMA(1,1)** | 2 | 2 | 12932,13 | 12951,72 |
| **ARMA(1,2)** | 3 | 3 | 12725,34 | 12749,82 |
| **ARMA(1,3)** | 4 | 4 | 12637,87 | 12667,25 |
| **ARMA(1,4)** | 5 | 5 | 12597 | 12631,28 |

*Sursa:* Prelucrarea autorului utilizând RStudio, pe baza datelor culese

Se observă din tabelul anterior că pe baza criteriilor Akaike și Scwartz, care trebuie să fie minime, cele mai bune modele sunt AR(2) și ARMA(1,4).

Din punct de vedere al erorilor, care trebuie să fie necorelate, homoscedastice și normal distribuite, situația pentru fiecare model se află în tabelul de mai jos. Testele utilizate pentru verificarea autocorelării erorilor, a homoscedasticității și a distribuției normale sunt următoarele:

* Testul Jarque-Bera pentru verificarea distribuției normale:

H0: erorile sunt normal distribuite.

H1: erorile nu sunt normal distribuite.

* Testul ARCH-LM pentru verificarea homoscedasticității erorilor:

H0: erorile sunt homoscedastice.

H1: erorile sunt heteroscedastice.

* Testul Ljung-Box pentru verificarea autocorelării erorilor:

H0: erorile nu sunt autocorelate.

H1: erorile prezintă autocorelație.

În situația în care valoarea absolută a testului statistic calculat este mai mică decât valoare tabelată în modul sau valoarea probabilității p-value este mai mare decât nivelul de încredere, se va accepta ipoteza nulă H0 și se va respinge ipoteza alternativă H1.

**Tabel 5. Caracteristicile erorilor pentru fiecare model utilizat**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modele** | **p-value** | | |
| **Jarque-Bera** | **ARCH-LM** | **Ljung-Box** |
| **AR(1)** | 0.00 | 0.00 | 0.00 |
| **AR(2)** | 0.00 | 0.09 | 0.77 |
| **MA(1)** | 0.00 | 0.00 | 0.00 |
| **MA(2)** | 0.00 | 0.00 | 0.00 |
| **MA(3)** | 0.00 | 0.00 | 0.00 |
| **MA(4)** | 0.00 | 0.00 | 0.00 |
| **MA(5)** | 0.00 | 0.00 | 0.84 |
| **ARMA(1,1)** | 0.00 | 0.02 | 0.00 |
| **ARMA(1,2)** | 0.00 | 0.03 | 0.00 |
| **ARMA(1,3)** | 0.00 | 0.04 | 0.02 |
| **ARMA(1,4)** | 0.00 | 0.03 | 0.61 |

*Sursa:* Prelucrarea autorului utilizând Visual Studio Code, pe baza datelor culese

Se constată studiind tabelul caracteristicilor erorilor că niciun model nu prezintă reziduuri cu distribuție normală. Din punct de vedere al autocorelării erorilor, modelele valide sunt AR(2), MA(5), și ARMA(1,4), în timp ce singurul model semnificativ din punct de vedere al homoscedasticității este AR(2). Se va alege modelul AR(2) pentru prognoza datelor, întrucât prezintă unele dintre cele mai mici valori pentru criteriile informaționale Akaike și Schwartz și de asemenea reziduurile modelului nu prezintă autocorelare și heteroscedasticitate.

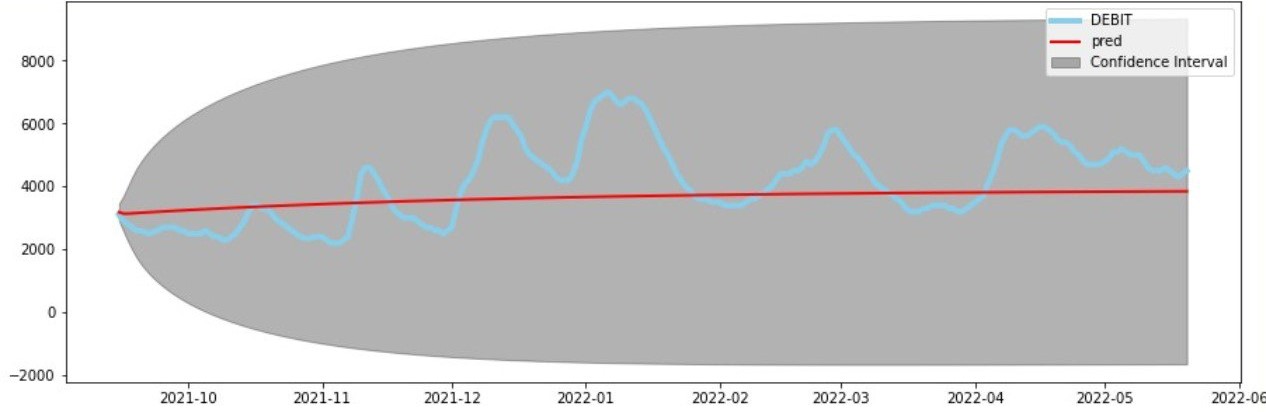
Utilizând setul de testare pentru a testa calitatea acestuia, au rezultat următoarele valori pentru erorile de predicție:

**Tabel 6. Caracteristicile erorilor pentru fiecare model utilizat**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **ME** | **RMSE** | **MAE** | **MPE** | **MAPE** | **MASE** |
| **Set de antrenare** | -0.58 | 139.88 | 92.75 | -0.11 | 1.97 | 0.046 |
| **Set de testare** | -587.51 | 1278.01 | 1075.09 | -23.82 | 31.78 | 0.54 |

*Sursa:* Prelucrarea autorului utilizând RStudio, pe baza datelor culese

Se poate observa cu ușurință erorile foarte mari de predicție, lucru ilustrat și în graficul prognozei.



Figură 7- Graficul setului de testare și al predicției

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

Cu toate că erorile sunt unele destul de mari, se constată că prognoza se încadrează în intervalul de încredere.

# **4.2. Studierea sezonalității debitului Dunării**

În continuare se va studia dacă debitul Dunării prezintă totuși sezonalitate pe termen scurt, fâcându-se modificări asupra seriei. În acest sens, se vor grupa datele câte 7, 10, 15, 20 și 30, se vor calcula mediile și astfel se va verifica sezonalitatea.

**Tabel 7. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

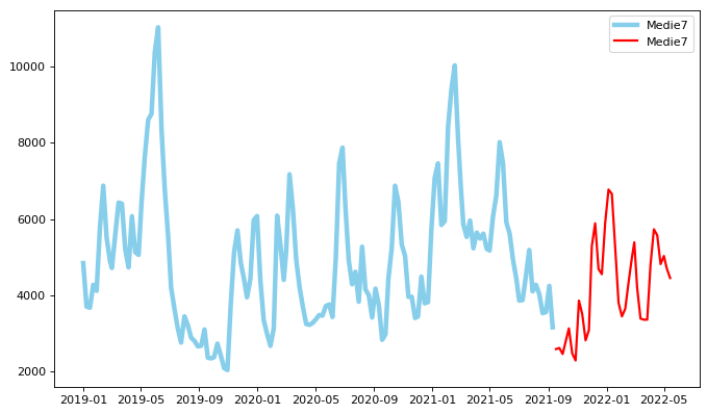
|  |  |
| --- | --- |
|  | **Valoarea p-value a testului Hegy pentru determinarea sezonalității** |
| **Medii săptămânale** | 0,22 |
| **Medii pe 10 zile** | 0,16 |
| **Medii pe 15 zile** | 0,11 |
| **Medii pe 20 de zile** | 0,00 |
| **Medii lunare** | 0,00 |

*Sursa:* Prelucrarea autorului utilizând RStudio

Se ilustrează astfel în tabel că doar seriile de date reprezentând mediile săptămânale, pe 10 și respectiv 15 zile prezintă sezonalitate și în continuare se va studia posibilitatea aplicării modelului SARIMA pe aceste serii. Unul dintre motivele pentru care seriile cu valori medii pentru 20 de zile și lunare nu prezintă sezonalitate ar putea fi numărul mic de observații rămase în urma modificărilor aduse seriei inițiale.

* Medii săptămânale

Setul de date a fost împărțit în set de antrenare și set de testare, în proporțiile 80-20%, ilustrate cu albastru, respectiv cu roșu în graficul din figura următoare.



Figură 8-Împărțirea setului de date în set de antrenare și set de testare

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

Se vor utiliza testele Augmented Dickey-Fuller, Phillips-Perron și Kwiatkowski–Phillips–Schmidt–Shin (KPSS) pentru verificarea staționarității seriei.

În urma aplicării celor 3 teste statistice asupra seriei de date, au rezultat următoarele output-uri, prezentate în tabelul alăturat:

**Tabel 8. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 0,06 | -1,3619 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,07 | 0,17 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 0,01 | -25,69 | -2,56 | -2,86 | -3,43 |

*Sursa:* Prelucrarea autorului utilizând RStudio

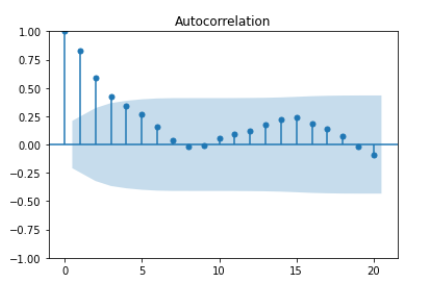
Analizând tabelul output-urilor testelor statistice pentru verificarea staționarității se poate concluziona că seria nu este staționară prin toate testele. Pentru început, se va diferenția sezonier seria și se va verifica iarăși staționaritatea acesteia.

**Tabel 9. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 1,104e-5 | -4,15 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,06 | 0,322 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 0,04 | -19,65 | -2,56 | -2,86 | -3,43 |

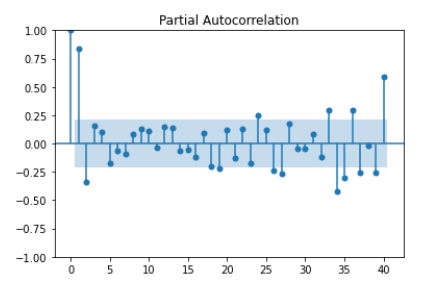
*Sursa:* Prelucrarea autorului utilizând RStudio

Se remarcă faptul că după diferențierea sezonieră, seria nu are nevoie de o diferențiere non-sezonieră.Pentru a putea determina parametri modelului SARIMA, trebuie studiată funcția de autocorelație și funcția de autocorelație parțială.



Figură 9-Funcția de autocorelație

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese



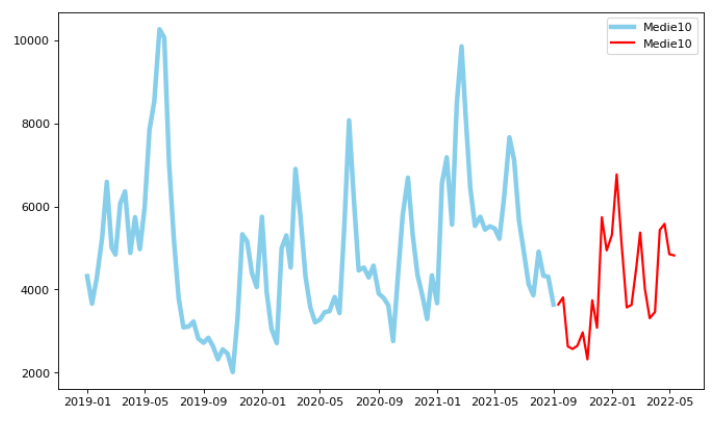
Figură 10- Funcția de autocorelație parțială

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

În mediul de programare RStudio au fost încercate numeroase modele, printre care și următoarele: SARIMA(0,0,0)(1,1,0), SARIMA(0,0,0)(0,1,1), SARIMA(1,0,0)(0,1,1), SARIMA(1,0,0)(1,1,0), SARIMA(0,0,1)(1,1,0), SARIMA(0,0,1)(0,1,1), SARIMA(1,0,1)(1,1,0), SARIMA(1,0,1)(0,1,1). Dintre toate modele verificate, niciunul nu a prezentat parametri sezonieri semnificativi, deși cei non-sezonieri erau seminificativi, în consecință, se poate trage concluzia că seria de date reprezentată de mediile săptămânale ale debitului Dunării nu poate fi prognozată utilizând un model SARIMA, ci un model ARIMA cu seria ajustată sezonier.

* Mediile debitului pe 10 zile

Setul de date a fost împărțit în set de antrenare și set de testare, în proporțiile 80-20%, ilustrate cu albastru, respectiv cu roșu în graficul din figura următoare.



Figură 11-Împărțirea setului de date în set de antrenare și set de testare

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

Se vor utiliza testele Augmented Dickey-Fuller, Phillips-Perron și Kwiatkowski–Phillips–Schmidt–Shin (KPSS) pentru verificarea staționarității seriei.

În urma aplicării celor 3 teste statistice asupra seriei de date, au rezultat următoarele output-uri, prezentate în tabelul alăturat:

**Tabel 10. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 0,2549 | -1,1797 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,1 | 0,1939 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 0,03 | -24,51 | -2,56 | -2,86 | -3,43 |

*Sursa:* Prelucrarea autorului utilizând RStudio

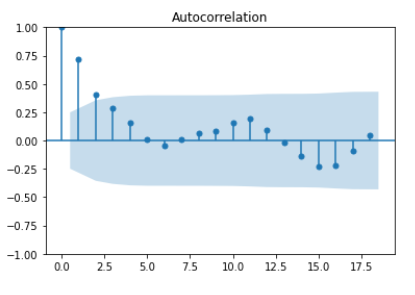
Analizând tabelul output-urilor testelor statistice pentru verificarea staționarității se poate concluziona că seria trebuie staționarizată. Se va diferenția sezonier seria și se va verifica iarăși staționaritatea acesteia.

**Tabel 11. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 0,00053 | -3,7472 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,07 | 0,3098 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 2,397e-12 | -16,508 | -2,56 | -2,86 | -3,43 |

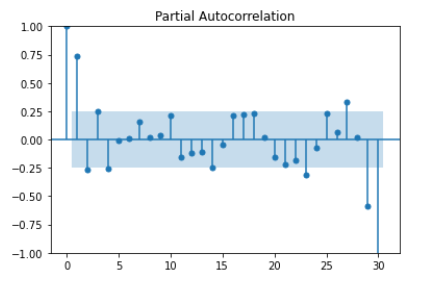
*Sursa:* Prelucrarea autorului utilizând RStudio

Se remarcă faptul că după diferențierea sezonieră, seria nu are nevoie de o diferențiere non-sezonieră.Pentru a putea determina parametri modelului SARIMA, trebuie studiată funcția de autocorelație și funcția de autocorelație parțială.



Figură 12-Funcția de autocorelație

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese



Figură 13- Funcția de autocorelație parțială

*Sursa: Prelucrarea autorului în Visual Studio Code, pe baza datelor culese*

Printre modelele încercate în RStudio se află și următoarele: SARIMA(0,0,0)(0,1,1), SARIMA(0,0,1)(0,1,1), SARIMA(1,0,0)(1,1,1), SARIMA(2,0,0)(0,1,1), SARIMA(2,0,1)(0,1,1), SARIMA(1,0,1)(0,1,1), SARIMA(1,0,3)(0,1,1), SARIMA(2,0,3)(1,1,1). Dintre toate modele verificate, doar două au prezentat parametri semnificativi din punct de vedere statistic. Aceste modele sunt sumarizate în tabelul următor.

**Tabel 12. Principalele criterii pentru alegerea celui mai bun model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modele** | **Numărul parametrilor modelului** | **Parametri semnificativi** | **Criteriul Akaike** | **Criteriul Schwartz** |
| **SARIMA(1,0,1)(1,1,0)** | 3 | 3 | 1047,65 | 1056,03 |
| **SARIMA(0,0,0)(1,1,0)** | 1 | 1 | 1097,17 | 1101,36 |
| **SARIMA(2,0,1)(1,1,0)** | 4 | 4 | 1036,84 | 1047,23 |

*Sursa:* Prelucrarea autorului utilizând RStudio, pe baza datelor culese

Se observă din tabelul anterior că pe baza criteriilor Akaike și Scwartz, care trebuie să fie minime, cel mai bun model dintre cele 3 este modelul SARIMA(2,0,1)(1,1,0).

Din punct de vedere al erorilor, care trebuie să fie necorelate, homoscedastice și normal distribuite, situația pentru fiecare model ce prezintă parametri semnificativi se află în tabelul de mai jos. Testele utilizate pentru verificarea autocorelării erorilor, a homoscedasticității și a distribuției normale sunt testul Jarque-Bera, ARCH-LM și Ljung-Box.

În situația în care valoarea probabilității p-value este mai mare decât nivelul de încredere, se va accepta ipoteza nulă H0 și se va respinge ipoteza alternativă H1.

**Tabel 13. Caracteristicile erorilor pentru fiecare model utilizat**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modele** | **p-value** | | |
| **Jarque-Bera** | **ARCH-LM** | **Ljung-Box** |
| **SARIMA(1,0,1)(1,1,0)** | 0.06 | 0.37 | 0.90 |
| **SARIMA(0,0,0)(1,1,0)** | 0.63 | 0.57 | 0.00 |
| **SARIMA(2,0,1)(1,1,0)** | 0.09 | 0.50 | 0.93 |

*Sursa:* Prelucrarea autorului utilizând Visual Studio Code, pe baza datelor culese

Se constată studiind tabelul caracteristicilor erorilor că modelele prezintă reziduuri cu distribuție normală. Din punct de vedere al autocorelării erorilor, modelele valide sunt SARIMA(1,0,1)(1,1,0) și SARIMA(2,0,1)(1,1,0), în timp ce din punct de vedere al homoscedasticității toate modele sunt valide. Se va alege modelul SARIMA(2,0,1)(1,1,0) pentru prognoza datelor, întrucât prezintă unele dintre cele mai mici valori pentru criteriile informaționale Akaike și Schwartz.

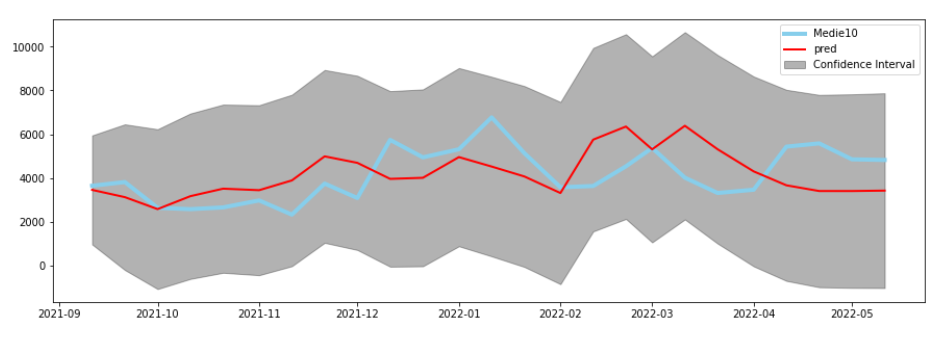
Utilizând setul de testare pentru a verifica calitatea modelului ales, au rezultat următoarele valori pentru erorile de predicție:

**Tabel 14. Caracteristicile erorilor pentru fiecare model utilizat**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **ME** | **RMSE** | **MAE** | **MPE** | **MAPE** | **MASE** |
| **Set de antrenare** | 0,04 | 0,46 | 0,21 | 0,11 | 0,56 | 0,24 |
| **Set de testare** | -942,95 | 1672,47 | 1291,09 | -30,08 | 36,23 | 0,64 |

*Sursa:* Prelucrarea autorului utilizând RStudio, pe baza datelor culese

Se pot observa cu ușurință valorile mari ale erorilor de predicție pentru setul de testare. Modul în care datele previzionate diferă de cele din setul de testare este ilustrat în figura următoare. Se poate constata că evoluția previzionată se încadrează în intervalul de încredere, și de asemenea, la o analiza mai amănunțită, se remarcă faptul că valorile prognozate înfățisează într-o anumită măsură perioadele de creștere și scădere parcurse de setul de testare.

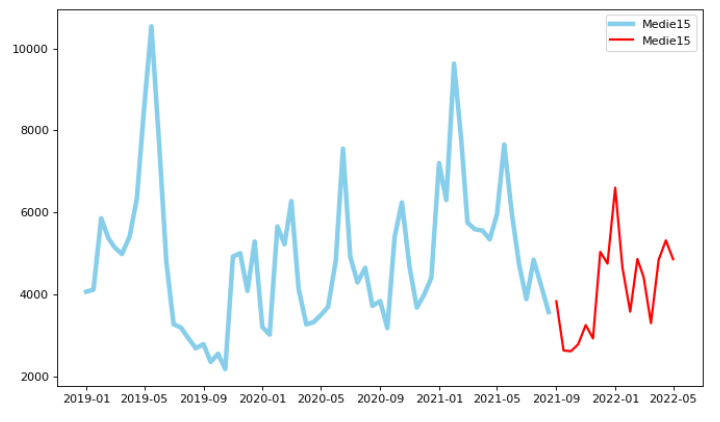


Figură 14- Graficul setului de testare și al predicției

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

* Medii bilunare

Setul de date a fost împărțit în set de antrenare și set de testare, în proporțiile 80-20%, ilustrate cu albastru, respectiv cu roșu în graficul din figura următoare.



Figură 15-Împărțirea setului de date în set de antrenare și set de testare

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

Se vor utiliza testele Augmented Dickey-Fuller, Phillips-Perron și Kwiatkowski–Phillips–Schmidt–Shin (KPSS) pentru verificarea staționarității seriei.

În urma aplicării celor 3 teste statistice asupra seriei de date, au rezultat următoarele output-uri, prezentate în tabelul alăturat:

**Tabel 15. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 0,5588 | -1,0729 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,05 | 0,134 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 0,039 | -23,3015 | -2,56 | -2,86 | -3,43 |

*Sursa:* Prelucrarea autorului utilizând RStudio

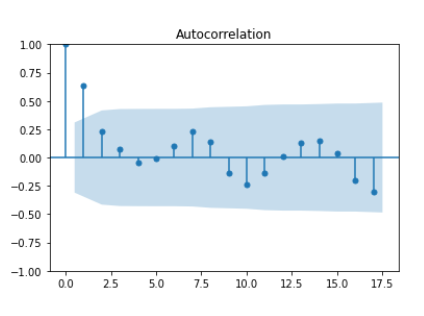
Analizând tabelul output-urilor testelor statistice pentru verificarea staționarității se poate concluziona că seria nu prezintă staționaritate și trebuie diferențiată. Se va aplica mai întâi o diferențiere sezonieră, apoi se va verifica iarăși staționaritatea setului de date.

**Tabel 16. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 0,006 | -3,3893 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,06 | 0,305 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 1,063e-5 | -15,17 | -2,56 | -2,86 | -3,43 |

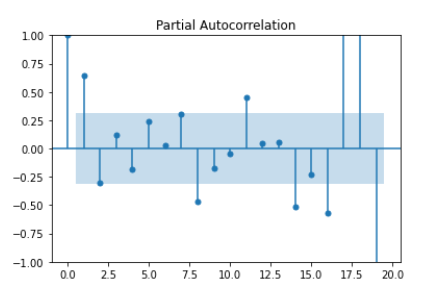
*Sursa:* Prelucrarea autorului utilizând RStudio

Se remarcă faptul că după diferențierea sezonieră, seria nu are nevoie de o diferențiere non-sezonieră, întrucât este staționară.Pentru a putea determina parametri modelului SARIMA, trebuie studiată funcția de autocorelație și funcția de autocorelație parțială.



Figură 16-Funcția de autocorelație

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese



Figură 17- Funcția de autocorelație parțială

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

În mediul de programare RStudio au fost încercate multe modele, printre care și următoarele: SARIMA(0,0,0)(0,1,1), SARIMA(1,0,0)(0,1,1), SARIMA(0,0,1)(0,1,1), SARIMA(1,0,0)(0,1,1), SARIMA(1,0,1)(0,1,1), SARIMA(0,0,2)(0,1,1), SARIMA(1,0,2)(0,1,1), SARIMA(1,0,1)(1,1,1). Dintre toate modele verificate, niciunul nu a prezentat parametri sezonieri semnificativi, deși cei non-sezonieri erau seminificativi, în consecință, se poate trage concluzia că seria de date reprezentată de mediile bilunare ale debitului Dunării nu poate fi prognozată utilizând un model SARIMA, ci un model ARIMA cu seria ajustată sezonier.

# **5.1. Analiza nivelului de clorofilă din Marea Neagră**

În partea a doua a acestei lucrări va fi analizat nivelul de clorofilă din apele maritime de pe teritoriul țării. Clorofila, (cuvânt provenit din limba greacă, unde *cloros* înseamnă verde) reprezintă acea moleculă ce dă pigmentul verde plantelor și este esențială procesului de fotosinteză. Aceasta se regăsește în toate tipurile de alge, verzi, brune, roșii și albastre-verzi. Algele au un rol foarte important asupra întregii vieți de pe planetă, întrucât un procent foarte mare din oxigenul care există este produs de algele marine, aproape 60%. În plus, acestea funcționează ca un sistem de oxigenare pentru viețuitoarele acvatice, deoarece algele produc majoritatea oxigenului din apă, în urma procesului de fotosinteză. La baza alimentației unor viețuitoare acvatice, precum racii, moluștele sau chiar mamiferele acvatice, stă fitoplanctonul. Acesta este compus din alge ce se regăsesc în biomasa maritimă și oceanică într-un procent foarte mare. Se poate concluziona astfel că nivelul de clorofilă reprezintă un indicator pentru starea de sănătate a mărilor și oceanelor. Nivelul clorofilei în apele Mării Negre sunt ilustrate în figura următoare:



Figură 18-Nivelul clorofilei în apele Mării Negre pe teritoriul țării

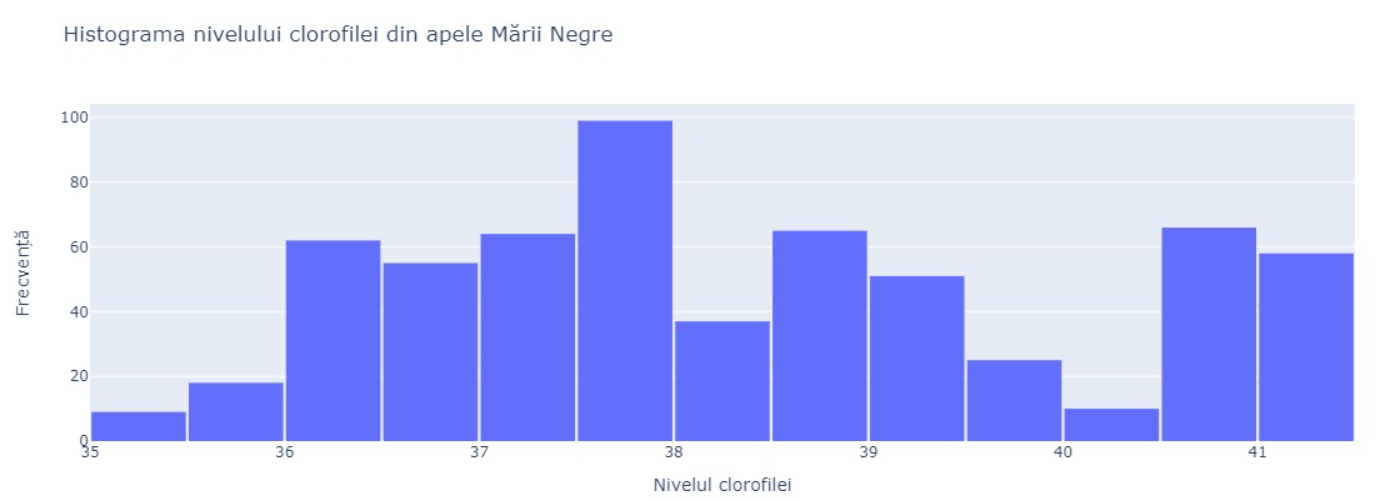
*Sursa:* [*Emodnet*](https://emodnet-chemistry.maris.nl/search)

**Tabel 17. Statisticile descriptive ale setului de date**

|  |  |
| --- | --- |
| **Statisticile descriptive ale nivelului de clorofilă din Marea Neagră** | |
| Mean | 38.29 |
| Standard Error | 0.08 |
| Median | 37.80 |
| Mode | 36.40 |
| Standard Deviation | 1.71 |
| Sample Variance | 2.92 |
| Kurtosis | -1.07 |
| Skewness | 0.33 |
| Range | 5.94 |
| Minimum | 35.29 |
| Maximum | 41.23 |

*Sursa:* Prelucrarea autorului în Microsoft Excel, pe baza datelorculese

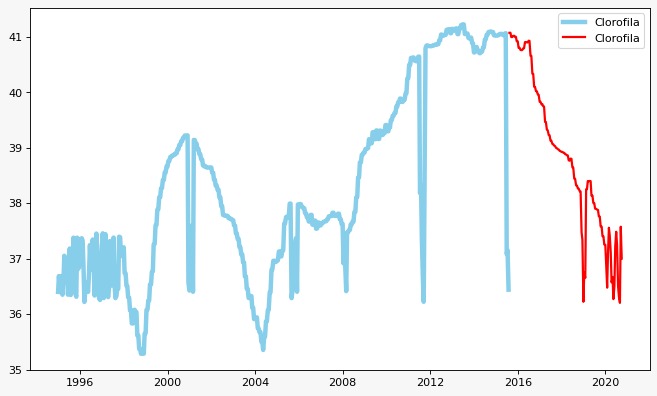
În tabelul 1 este ilustrată descrierea statistică a seriei de date, cu ajutorul căreia putem caracteriza setul de date. Se observă în primul rând indicatorii tendinței centrale, media, mediana și modul. Aceștia rezumă într-o măsură ridicată datele și sunt strâns legați de centrul setului de date. Media are o valoare aproximativă de 38 unități, mai mare decât valorile medianei și a modului, de 37, respectiv 36. Acest lucru ilustrează o asimetrie la stânga, pozitivă, fapt confirmat și de coeficientul skewness cu valoarea de 0,33. În al doilea rând, putem caracteriza distribuția datelor, utilizând coeficienții skewness și kurtosis. Așa cum am specificat anterior, asimetria este una pozitivă, la stânga, iar boltirea este una platikurtică. În ceea ce privește indicatorii variației datelor, se observă că amplitudinea are o valoare foarte mică, în consecință putem trage concluzia că datele nu au o împrăștiere mare față de medie, acestea abătându-se de la medie cu aproximativ 1,7 unități.



Figură 19-Histograma nivelului clorofilei din apele Mării Negre

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

Setul de date compus din valori bilunare a fost împărțit în set de antrenare și set de testare, în proporție de 80-20 %. Primul set va fi utilizat pentru definirea modelului și al doilea va fi folosit pentru a testa calitatea acestuia. Cele două seturi și împărțirea acestora se vede în graficul următor:



Figură 20-Împărțirea setului de date în set de antrenare și set de testare

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

Pentru a putea modela seria prin intermediul seriilor de timp, trebuie în primul rând verificată staționaritatea seriei. În acest scop, voi utiliza testele Augmented Dickey-Fuller, Phillips-Perron și Kwiatkowski–Phillips–Schmidt–Shin (KPSS).

Așa cum a fost specificat anterior, în situația în care valoarea absolută a testului statistic calculat este mai mică decât valoare tabelată în modul, sau valoarea probabilității p-value este mai mare decât nivelul de încredere, se va accepta ipoteza nulă H0 și se va respinge ipoteza alternativă H1.

În urma aplicării celor 3 teste statistice asupra seriei de date, au rezultat următoarele output-uri, prezentate în tabelul alăturat:

**Tabel 18. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 0,3732 | 0,3371 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,01 | 5,4056 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 0,01 | -12,32 | -2,56 | -2,86 | -3,43 |

*Sursa:* Prelucrarea autorului utilizând RStudio

Analizând tabelul output-urilor testelor statistice pentru verificarea staționarității se poate concluziona că seria este staționară utilizând testul Phillips-Perron, dar celelalte două teste afirmă faptul că seria nu prezintă staționaritate, deci ar necesita diferențiere. Astfel, se va alege diferențierea seriei pentru a avea un set de date cât mai comprehensibil pentru modelare. Înainte de acest lucru trebuie verificată sezonalitatea seriei, pentru a putea decide dacă seria are nevoie poate de o diferențiere sezonieră în locul uneia sau pe langă una non-sezonieră. Totodată, outputul testului Hegy va determina și posibilitatea aplicării unui model de tip SARIMA datelor.

**Tabel 19. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |
| --- | --- |
| **Testul utilizat** | **Valoare probabilității**  **p-value** |
| **Hegy** | 0,6 |

*Sursa:* Prelucrarea autorului utilizând RStudio

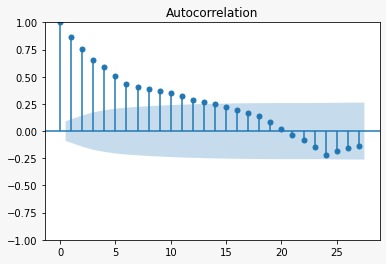
Tabelul reprezentând output-ul testului Hegy ilustrează valoarea probabilității p-value ca fiind mai mare decât nivelul de încredere, de 0,05, în consecință se va accepta ipoteza nulă H0 conform căreia seria prezintă sezonalitate. Se va aplica o diferențiere sezonieră setului de date, iar apoi se va verifica iarăși staționaritatea seriei, utilizând cele trei teste ADF, KPSS și Phillips-Perron.

**Tabel 20. Output-urile testelor statistice pentru verificarea staționarității seriei inițiale**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testul utilizat** | **Probabilitatea** | **Valoarea calculată a testului statistic** | **Valoarea tabelară a testului pentru un prag de semnificație egal cu:** | | |
|  | **p-value** | **tcalculat** | **10 %** | **5%** | **1%** |
| **ADF** | 2,023e-07 | -5,53 | -1,62 | -1,95 | -2,58 |
| **KPSS** | 0,05 | 0,18 | 0,347 | 0,463 | 0,739 |
| **Phillips-Perron** | 2,2e-16 | -60,82 | -2,56 | -2,86 | -3,43 |

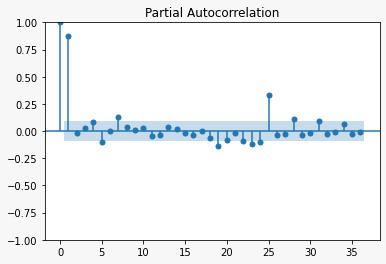
*Sursa:* Prelucrarea autorului utilizând RStudio

Prin utilizarea testelor, se confirmă faptul că seria este acum staționară și nu mai are nevoie și de o diferențiere non-sezonieră față de una sezonieră. Pentru a putea determina parametri modelului SARIMA, trebuie studiată funcția de autocorelație și funcția de autocorelație parțială.



Figură 21-Funcția de autocorelație

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese



Figură 22- Funcția de autocorelație parțială

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

În mediul de programare RStudio au fost încercate numeroase modele, printre care și următoarele: SARIMA(0,0,0)(1,1,0), SARIMA(0,0,0)(1,1,1), SARIMA(0,0,2)(1,1,1), SARIMA(0,0,2)(0,1,1), SARIMA(0,0,0)(1,1,2), SARIMA(0,0,0)(1,1,3), SARIMA(0,0,1)(1,1,0), SARIMA(0,0,2)(1,1,0), SARIMA(0,0,1)(1,1,1), SARIMA(0,0,1)(1,1,2), SARIMA(0,0,1)(1,1,3), SARIMA(0,0,2)(1,1,3). Dintre toate modele verificate, doar câteva au prezentat parametri semnificativi. Descrierile succinte ale output-utilor acestor modele se regăsesc în tabelul următor.

**Tabel 21. Principalele criterii pentru alegerea celui mai bun model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modele** | **Numărul parametrilor modelului** | **Parametri semnificativi** | **Criteriul Akaike** | **Criteriul Schwartz** |
| **SARIMA(0,0,0)(1,1,0)** | 1 | 1 | 1500,81 | 1509,13 |
| **SARIMA(0,0,0)(1,1,1)** | 2 | 2 | 1495,45 | 1507,93 |
| **SARIMA(0,0,0)(0,1,1)** | 1 | 1 | 1497,9 | 1506,22 |
| **SARIMA(1,0,0)(0,1,1)** | 2 | 2 | 675,5 | 687,99 |
| **SARIMA(1,0,0)(2,1,0)** | 3 | 3 | 708,67 | 725,31 |
| **SARIMA(0,0,1)(1,1,0)** | 2 | 2 | 1135,59 | 1148,08 |
| **SARIMA(0,0,1)(0,1,1)** | 2 | 2 | 1130,99 | 1143,47 |
| **SARIMA(0,0,2)(0,1,1)** | 3 | 3 | 915,02 | 931,67 |
| **SARIMA(0,0,2)(2,1,0)** | 4 | 4 | 921,14 | 941,94 |

*Sursa:* Prelucrarea autorului utilizând RStudio, pe baza datelor culese

Se observă din tabelul anterior că pe baza criteriilor Akaike și Scwartz, care trebuie să fie minime, cele mai bune modele sunt SARIMA(1,0,0)(0,1,1) și SARIMA(1,0,0)(2,1,0), urmate îndeaproape de SARIMA(0,0,2)(0,1,1) și SARIMA(0,0,2)(2,1,0).

Din punct de vedere al erorilor, care trebuie să fie necorelate, homoscedastice și normal distribuite, situația pentru fiecare model ce prezintă parametri semnificativi se află în tabelul de mai jos. Testele utilizate pentru verificarea autocorelării erorilor, a homoscedasticității și a distribuției normale sunt testul Jarque-Bera, ARCH-LM și Ljung-Box.

În situația în care valoarea probabilității p-value este mai mare decât nivelul de încredere, se va accepta ipoteza nulă H0 și se va respinge ipoteza alternativă H1.

**Tabel 22. Caracteristicile erorilor pentru fiecare model utilizat**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modele** | **p-value** | | |
| **Jarque-Bera** | **ARCH-LM** | **Ljung-Box** |
| **SARIMA(0,0,0)(1,1,0)** | 0,00 | 0,52 | 0,00 |
| **SARIMA(0,0,0)(1,1,1)** | 0,00 | 0,79 | 0,88 |
| **SARIMA(0,0,0)(0,1,1)** | 0,00 | 0,92 | 0,00 |
| **SARIMA(1,0,0)(0,1,1)** | 0,00 | 0,66 | 0,70 |
| **SARIMA(1,0,0)(2,1,0)** | 0,00 | 0,56 | 0,80 |
| **SARIMA(0,0,1)(1,1,0)** | 0,00 | 0,81 | 0,00 |
| **SARIMA(0,0,1)(0,1,1)** | 0,00 | 0,56 | 0,00 |
| **SARIMA(0,0,2)(0,1,1)** | 0,00 | 0,68 | 0,00 |
| **SARIMA(0,0,2)(2,1,0)** | 0,00 | 0,27 | 0,61 |

*Sursa:* Prelucrarea autorului utilizând Visual Studio Code, pe baza datelor culese

Se constată studiind tabelul caracteristicilor erorilor că niciun model nu prezintă reziduuri cu distribuție normală. Din punct de vedere al autocorelării erorilor, modelele valide sunt SARIMA(0,0,0)(1,1,1), SARIMA(1,0,0)(0,1,1), SARIMA(1,0,0)(2,1,0), SARIMA(2,0,0)(2,1,0), în timp ce din punct de vedere al homoscedasticității toate modele sunt valide. Se va alege modelul SARIMA(1,0,0)(0,1,1) pentru prognoza datelor, întrucât prezintă unele dintre cele mai mici valori pentru criteriile Akaike și Schwartz și de asemenea reziduurile modelului nu prezintă autocorelare și heteroscedasticitate.

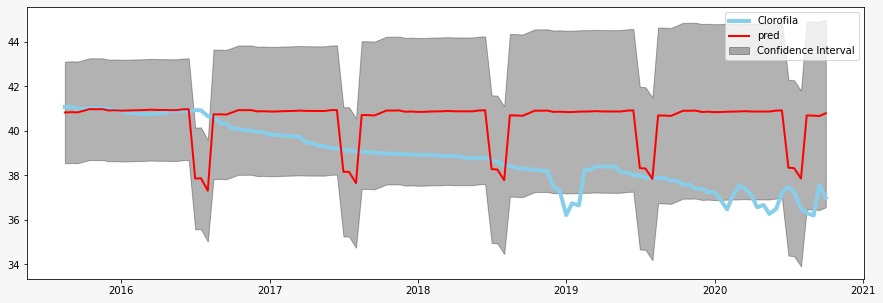
Utilizând setul de testare pentru a testa calitatea acestuia, au rezultat următoarele valori pentru erorile de predicție:

**Tabel 23. Caracteristicile erorilor pentru fiecare model utilizat**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **ME** | **RMSE** | **MAE** | **MPE** | **MAPE** | **MASE** |
| **Set de antrenare** | 0,04 | 0,46 | 0,21 | 0,11 | 0,56 | 0,24 |
| **Set de testare** | -1,05 | 1,51 | 1,17 | -2,80 | 3,11 | 1,34 |

*Sursa:* Prelucrarea autorului utilizând RStudio, pe baza datelor culese

Se pot observa cu ușurință valorile foarte mici ale erorilor de predicție pentru setul de testare, în consecință este de așteptat, din punct de vedere grafic, ca modelul să urmeze aproape fidel evoluția setului de testare. Cu toate acestea, se observă din figura următoare că modelul ales nu surprinde într-o măsură adecvată evoluția seriei analizate. Acest lucru poate fi cauzat de diferențele mici între valorile de testare, și de asemenea, trebuie luat în considerare faptul că diferența maximă dintre cea mai mare valoare și valoarea minimă între seturile de testare și cel de previziune este de 5 unități, ceea ce explică ipoteza conform căreia erorile de predicție sunt mici, deși modelul nu urmează fidel evoluția variabilei analizate.



Figură 23- Graficul setului de testare și al predicției

*Sursa:* Prelucrarea autorului în Visual Studio Code, pe baza datelor culese

# **5.2. Regresie**

Se va încerca în continuare să se explice fenomenul modificării nivelului de clorofilă din Marea Neagră utilizând ca variabile independente aciditatea apei, nivelul hidrocarburilor, nivelul metelelor grele prezente în apă, a fertilizatorilor utilizați în agricultură care s-au scurs în mare și a deșeurilor marine.

Pentru început, se va încerca determinarea modificării nivelului de clorofilă utilizând toate variabilele independente. Output-ul regresiei se află ilustrat în tabelul următor:

**Tabel 24. Output-ul regresiei utilizând cinci variabile independente**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Coeficient** | **Eroarea standard** | **Tcritic** | **p-value** | **Limita inferioară a intervalului de încredere(95%)** | **Limita superioară a intervalului de încredere(95%)** |
| **Intercept** | 27.20 | 17.99 | 1.51 | 0.13 | -8.64 | 63.05 |
| **Aciditate** | -0.04 | 0.24 | -0.19 | 0.84 | -0.54 | 0.44 |
| **Deșeuri marine** | -0.17 | 0.34 | -0.51 | 0.60 | -0.85 | 0.50 |
| **Hidrocarburi** | 0.09 | 0.26 | 0.36 | 0.71 | -0.43 | 0.62 |
| **Metale grele** | 0.18 | 0.27 | 0.68 | 0.49 | -0.35 | 0.73 |
| **Fertilizatori** | 0.23 | 0.25 | 0.91 | 0.36 | -0.27 | 0.73 |

*Sursa:* Prelucrarea autorului utilizând Microsoft Excel, pe baza datelor culese

Se poate observa din tabel că niciun parametru nu este semnificativ statistic, întrucât valoarea asociată p-value depășește pragul de semnificație de 0,05.

În continuare s-a verificat dacă nivelului de clorofilă din Marea Neagră poate fi explicat de combinații ale celor cinci variabile independente precizate anterior, în combinații de câte patru, trei, două și doar o variabilă independentă. Unul dintre modele este prezentat în tabelul următor, nivelul clorofilei explicat de aciditatea apei.

**Tabel 25. Output-ul regresiei utilizând aciditatea apei ca variabilă independentă**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Coeficient** | **Eroarea standard** | **Tcritic** | **p-value** | **Limita inferioară a intervalului de încredere(95%)** | **Limita superioară a intervalului de încredere(95%)** |
| **Intercept** | 40.30 | 8.75 | 4.61 | 0.00 | 22.88 | 57.71 |
| **Aciditate** | -0.06 | 0.24 | -0.26 | 0.79 | -0.54 | 0.41 |

*Sursa:* Prelucrarea autorului utilizând Microsoft Excel, pe baza datelor culese

Au fost verificate toate posibilele modele care au ca variabile independente cele cinci atribute, dar niciun model nu a prezentat parametri semnificativi, cel mult termenul liber. Se poate concluziona în urma acestei analize asupra datelor culese că nivelul clorofilei nu este influențat de aciditatea apei, de nivelul hidrocarburilor, de nivelul metelelor grele prezente în apă, de fertilizatori utilizați în agricultură care s-au scurs în mare sau de deșeurile marine. Rezultatele acestui studiu pot fi cauzate și de numărul mic de observații luat în considerare.

# **5.3. Anova**

O altă ipoteză de cercetare ce poate fi aplicată pentru înțelegerea evoluției nivelului de clorofilă în apele Mării Negre este reprezentată de verificarea existenței unor diferențe semnificative în medie, din perspectiva celor patru zone geografice maritime, nord, sud, est, vest, de pe teritoriul țării.

Ipotezele testului statistic ANOVA sunt în acest caz:

H0: Mediile de grupă ale nivelului de clorofilă în apele Mării Negre sunt egale (nu diferă semnificativ).

H1: Mediile de grupă ale nivelului de clorofilă în apele Mării Negre nu sunt egale (diferă semnificativ).

Decizia de a respinge sau de a accepta ipoteza nulă H0 se va lua comparând nivelul de semnificație calculat pe baza datelor de selecție, p-value, cu nivelul de semnificație ales, 𝛼, în valoare de 0.05. Valoarea calculată a testului statistic F este de 3294,43, iar probabilitatea p-value asociată este 0,00, mai mică decât nivelul de semnificație de 0,05, prin urmare se poate respinge ipoteza nulă H0 la un nivel de semnificație de 5% și se va accepta ipoteza alternativă H1 conform căreia mediile de grupă ale nivelului de clorofilă în apele Mării Negre nu sunt egale (diferă semnificativ).

**Tabel 26. Statisticile descriptive ale datelor, grupate pe zonele geografice**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Numărul datelor** | **Suma** | **Media** | **Varianța** |
| **Nord** | 619 | 26858.41 | 43.39 | 0.65 |
| **Est** | 619 | 20024.65 | 32.35 | 12.73 |
| **Vest** | 619 | 20470.33 | 33.07 | 10.84 |
| **Sud** | 619 | 27551.69 | 44.51 | 0.88 |

*Sursa:* Prelucrarea autorului utilizând Microsoft Excel, pe baza datelor culese

**Tabel 27. Statisticile descriptive ale datelor, grupate pe zonele geografice**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variația** | **SS** | **df** | **MS** | **F** | **p-value** | **F crit** |
| **Între grupe** | 61961.61 | 3 | 20653.87062 | 3294.43 | 0.00 | 2.608503 |
| **În grupe** | 15497.79 | 2472 | 6.269331286 |  |  |  |

*Sursa:* Prelucrarea autorului utilizând Microsoft Excel, pe baza datelor culese

Întrucât mediile grupelor diferă semnificativ, este de așteptat ca fiecare zonă geografică să producă o cantitate diferită de clorofilă, așa cum este ilustrat și în descrierea statistică anterioară. Astfel, pentru zona nordică este scontat un nivel mediu de clorofilă de 43 de unități, în timp ce la polul opus, în sud se previzionează o valoare de 44 unități. Pe direcție orizontală, se observă la est o valoare de 32 de unități, în timp ce în vest valoarea este de 33 unități.

# **6. Concluzii**

Modelarea stochastică a unui serii de timp hidrologice reprezintă o temă abordată de numeroși autori, influențați de dorința de a cunoaște stările viitoare ale unui râu sau ale unui lac.

Studierea seriei de timp timp a fluviului Dunărea, set de date cuprins din 1236 de valori zilnice în intervalul 1 ianuarie 2019-20 mai 2022 a subliniat faptul că valorile prezintă staționaritate, deși acestea nu sunt caracterizate de sezonalitate. Aceste fapte au exclus din analiză posibilitatea aplicării unui model de tip ARIMA cu un ordin de integrare cel puțin egal cu 1 sau un model de tip SARIMA.

În urma analizei funcțiilor de autocorelație și autocorelație parțială au rezultat numeroase modele posibile, dintre care doar 11 au fost valide din punct de vedere al semnificației parametrilor, deși au prezentat valori foarte ridicate pentru criteriile informaționale Akaike și Schwartz, modele optime din acest punct de vedere fiind AR(2) și ARMA(1,4). Din cele 11 modele, niciunul nu a avut reziduurile normal distribuite, homoscedastice și necorelate, deși modelul AR(2) avea doar două dintre calități, prin urmare decizia a fost de a se testa doar previziunile acestui model.

Erorile de previziune ale modelului AR(2) sunt foarte mari, ceea ce a condus la o predicție departe de valorile reale, observabilă atât în grafic, cât și în tabel.

Din motive ce nu țin de mine, unul dintre neajunsurile acestei lucrări este reprezentată de numărul mic de observații considerat în analiză, ceea ce a dus la valorile ridicate ale erorilor de predicție. Numărul mic de observații este responsabil și de lipsa sezonalității în datele inițiale, ducând astfel la imposibilitatea aplicării unui model de tip SARIMA seriei de date cu valori zilnice.

În urma modificării seriei de date ce cuprinde valori zilnice, pentru a se studia dacă debitul Dunării prezintă totuși sezonalitate pe termen scurt, a rezultat faptul că doar seriile reprezentate de mediile săptămânale, pe o perioadă de 10 zile și bilunare au fost caracterizate de sezonalitate. Cu toate acestea, seturile de date cu valori săptămânale și bilunare nu au putut fi modelate utilizând SARIMA, întrucât parametri sezonieri ai modelelor nu erau valizi din punct de vedere statistic. În urma transformării, seriile cuprinzând mediile pe o perioadă de 20 de zile și mediile lunare, au avut prea puține valori, ceea ce este o posibilă cauză pentru care seturile de date nu au avut sezonalitate.

În partea a doua a lucrării, s-a studiat evoluția nivelului clorofilei în apele Mării Negre, în spațiul românesc. Setul de date a prezentat sezonalitate, așa ca a fost aplicată a diferențiere sezonieră pentru staționarizarea seriei, apoi a fost aplicat un model SARIMA, în încercarea de a previziona evoluția nivelului clorofilei. În urma alegerii modelului, s-au constatat că erorile de previziune erau unele mici, cu toate acestea, modelul ales nu surprinde într-o măsură adecvată evoluția seriei analizate. Acest lucru poate fi cauzat de diferențele mici între valorile de testare, și de asemenea, trebuie luat în considerare faptul că diferența maximă dintre cea mai mare valoare și valoarea minimă între seturile de testare și cel de previziune este de 5 unități, ceea ce explică ipoteza conform căreia erorile de predicție sunt mici, deși modelul nu urmează fidel evoluția variabilei analizate.

În încercarea de a explica fenomenul evoluției nivelului de clorofilă, s-a verificat posibilitatea ca variabile independente aciditatea apei, nivelul hidrocarburilor, nivelul metelelor grele prezente în apă, a fertilizatorilor utilizați în agricultură care s-au scurs în mare și a deșeurilor marine să descrie modificările nivelului clorofilei. În urma aplicării unor modele de regresie, atât unifactoriale cât și cu cinci, patru, trei și două variabile independente, s-a constatat că în urma acestei analize asupra datelor culese că nivelul clorofilei nu este influențat de aciditatea apei, de nivelul hidrocarburilor, de nivelul metelelor grele prezente în apă, de fertilizatori utilizați în agricultură care s-au scurs în mare sau de deșeurile marine. Rezultatele acestui studiu pot fi cauzate și de numărul mic de observații luat în considerare.

Pentru a evidenția distribuția nivelului de clorofilă în apele marine, s-a efectuat un test ANOVA pentru a verifica dacă mediile celor patru zone geografice, nord, sud, est și vest, sunt semnificativ statistic diferite una de cealaltă. În urma aplicării, a rezultat că mediile sunt diferite în funcție de zona geografică. Astfel, pentru zona nordică este scontat un nivel mediu de clorofilă de 43 de unități, în timp ce la polul opus, în sud se previzionează o valoare de 44 unități. Pe direcție orizontală, se observă la est o valoare de 32 de unități, în timp ce în vest valoarea este de 33 unități.