Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

Evidenčné číslo: 103003/B/2022/36124048423279108

Základy analýzy časových radov s využitím jazyka Python

Bakalárska práca

2022 **Lívia Pintérová**

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

Základy analýzy časových radov s využitím jazyka Python

Bakalárska práca

Študijný program: Manažérske rozhodovanie

Študijný odbor: Ekonómia a manažment

Školiace pracovisko: Katedra operačného výskumu a ekonometrie

Vedúci záverečnej práce: prof. Ing. Martin Lukáčik PhD.

Bratislava 2022 **Lívia Pintérová**

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Lívia Pintérová
Študijný program: manažérske rozhodovanie (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: ekonómia a manažment
Typ záverečnej práce: Bakalárska záverečná práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický
Názov: Základy analýzy časových radov s využitím jazyka Python
Anotácia: Jazyk Python je čoraz obľúbenejším jazykom aj pri práci s časovými radmi. Práca sa zameriava na prezentáciu základných metód analýzy časových radov pomocou tohto programovacieho jazyka.

Vedúci: prof. Ing. Martin Lukáčik, PhD.
Katedra: KOVE FHI - Katedra operačného výskumu a ekonometrie FHI
Vedúci katedry: prof. Mgr. Juraj Pekár, PhD.

Dátum zadania: 20.03.2021

Dátum schválenia: 31.03.2021

prof. Mgr. Juraj Pekár, PhD.
vedúci katedry

Čestné vyhlásenie

Čestne vyhlasujem, že záverečnú prácu som vypracovala samostatne, a že som uviedla všetku použitú literatúru.

V Bratislave dňa ..... **.....**

Lívia Pintérová

Poďakovanie

Touto cestou by som sa rada poďakovala svojmu vedúcemu bakalárskej práce prof. Ing. Martinovi Lukáčikovi, PhD. za cenné rady, trpezlivosť, pomoc, pripomienky a ochotu pri písaní záverečnej práce.

ABSTRAKT

PINTÉROVÁ, Lívia: Základy analýzy časových radov s využitím jazyka Python. – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra operačného výskumu a ekonometrie. – Vedúci záverečnej práce: prof. Ing. Martin Lukáčik, PhD. – Bratislava: FHI EU, 2022, 49 s.

Cieľom bakalárskej práce je oboznámenie so základnými metódami analýzy časových radov pomocou jazyka Python, a predložiť stručný popis niektorých populárnych modelov predpovedí časových radov s ich charakteristickými znakmi. Jazyk Python je čoraz obľúbenejším jazykom aj pri práci s časovými radmi. Modelovanie a predpovedanie údajov časových radov má zásadný význam v rôznych praktických oblastiach. Práca je rozdelená do 4 kapitol. Prvá kapitola je venovaná základnému predstaveniu časových radov a jazyka Python. Druhá a tretia kapitola sa venuje vymedzeniu cieľa práce, metodiky práce a metódam skúmania. Štvrtá kapitola sa zaoberá praktickou časťou, kde pomocou jazyka Python na 3 dátových setoch overujeme, či dátový set je vhodný na časovú predikciu, a túto predikciu robíme s využitím rôznych modelov. V poslednej časti sú zhrnuté a interpretované výsledky našich modelov.

Kľúčové slová: časové rady, jazyk Python, stacionarita, Box-Jenkinsová metodológia, lineárne modely, nelineárne modely, analýza časových radov, prognózovanie, ARIMA, SARIMA, HDP EÚ, miera nezamestnanosti EÚ, S&P 500 Index

ABSTRACT

PINTÉROVÁ, Lívia: Basics of time series analysis using Python. – University of Economics in Bratislava. Faculty of Economic Informatics; Department of operational research and econometrics. – Supervisor: prof. Ing. Martin Lukáčik, PhD. – Bratislava: FHI EU, 2022, 49 pages.

The aim of the bachelor thesis is to get acquainted with the basic methods of time series analysis using Python and to present a brief description of some popular models predicting time series with their characteristics. Python is an increasingly popular language for working with time series. Modeling and forecasting time series data is essential in various practical areas. The thesis is divided into 4 chapters. The first chapter gives basic introduction to time series and Python. The second and third chapters deal with the definition of the work objectives, work methodology and methods of investigation. The fourth chapter deals with the practical part where we use Python on 3 data sets to verify whether the data set is suitable for time prediction and we do this prediction using different models. The last part summarizes and interprets the results of our models.

Key words: time series, Python language, stationarity, Box-Jenkins methodology, linear models, nonlinear models, time series analysis, forecasting, ARIMA, SARIMA, GDP EU, unemployment rates EU, S&P 500 Index

Obsah

ÚVOD

Časové rady sú jedným z najbežnejších typov údajov, s ktorými sa stretávame v každodennom živote. Počasie, spotreba energie v domácnosti, medicínske prístroje na vizualizáciu životných funkcií (EKG, EEG), finančné ceny (výnosy investícií, úrokové sadzby a pod.), to všetko sú príklady údajov, ktoré možno pravidelne zhromažďovať. Analýza časových radov nie je nová štúdia, no dnešná technológia zjednodušuje prístup a umožňuje zhromažďovať veľké množstvo údajov každý deň. Jednoduchšie, ako kedykoľvek predtým je získať dostatok konzistentných údajov na komplexnú analýzu. Takmer každý dátový vedec sa pri svojej každodennej práci stretáva s časovými radmi. Vedieť ich modelovať je dôležitou zručnosťou pre nástroje dátovej vedy.

V tejto bakalárskej práci sa zameriame na vytvorenie správnych modelov, ktoré nám pomôžu predpovedať budúci vývoj nami pozorovaných časových radov s využitím jazyka Python so svojimi knižnicami na analýzu časových radov. Práca obsahuje štyri kapitoly, z ktorých v prvej si zadefinujeme základné štatistické a ekonometrické pojmy, ktoré približujú čitateľovi teoretické poznatky z oblasti časových radov. Vysvetlíme význam stochastických procesov, stacionarity, autokorelačnej a parciálne autokorelačnej funkcie a bieleho šumu. Taktiež si predstavíme programovací jazyk Python, ktorý využívame v praktickej časti. V druhej kapitole uvedieme ciele tejto bakalárskej práce. V tretej kapitole si predstavíme jednorovnicové lineárne modely AR, MA, ARMA, ARIMA a SARIMA pomocou Box-Jenkinsovej metodológie. V záverečnej časti tretej kapitoly sa zameriame na správnu identifikáciu modelu. Následne sa zameriame na verifikáciu modelu, kde uvedieme základné informačné kritéria, ktorými sú AIC, BIC a testovacie kritéria, ktorými sú reziduálna stacionarita, koeficient determinácie, autokorelácia a miery presnosti, ktoré overíme pre správne určenie modelu.

V poslednej kapitole si uvedieme do praxe teoretické poznatky z predchádzajúcich kapitol. Aplikujeme ich na dátach štvrťročných cien HDP EÚ, na mesačných dátach miery nezamestnanosti EÚ a na dátach denných cien S&P 500 Indexu. Pokúsime sa vytvoriť čo najpresnejší model na prognózu budúcich hodnôt. Výsledky nášho výskumu budú zhrnuté v závere tejto práce.

1 Súčasný stav analýzy časových radov

1.1 Časové rady

V tejto podkapitole sme čerpali hlavne z publikácie [7].

Časový rad je chronologicky usporiadaná postupnosť údajových bodov za určité časové obdobie, od minulosti po súčasnosť. Ide o súbor údajov, ktorý sleduje vzorky v čase a je vecne a priestorovo definovaný. Každý časový rad má dve hlavné zložky: čas a hodnotu priradenú zodpovedajúcemu časovému bodu.

Takto určený časový rad budeme zapisovať ako y_t, $t = 1, 2, 3, ..., n$. V časových radoch je čas nezávislou premennou. Cieľom je zvyčajne urobiť predpoveď do budúcnosti. V časovom rade potrebujeme poznať počiatočný a konečný bod, medzi počiatočným bodom a konečným bodom je mnoho ďalších bodov, čo nám umožňuje analyzovať časový rad. Každý bod musí mať rovnaký interval a byť jasne definovaný, čo vedie ku konštantnej frekvencii. Frekvencia je meranie času a môže sa pohybovať od milisekúnd až po desaťročia, ale najčastejšie používame denné, mesačné, štvrťročné a ročné frekvencie.

Časový rad umožňuje najmä vidieť aké faktory ovplyvňujú určité premenné z obdobia na obdobie. Očakáva sa, že vzorce pozorované v časových radoch budú pretrvávať aj v budúcnosti, preto sa často snažíme robiť predpovede analýzou zaznamenaných hodnôt.

Časový rad pozostáva zo zložiek ako je trend, sezónna, cyklická a náhodná zložka. Základný model rozkladu zohľadňuje tieto dve štruktúry:

- aditívne
- multiplikatívne

Časový rad zaznamenávajúci merania jednej premennej sa nazýva jednorozmerný, zatiaľ čo časový rad zaznamenávajúci merania viacerých premenných sa nazýva viacrozmerný. V tejto práci sa zameriame na predpoveď modelov, ktoré sa zaoberajú jednorozmernými časovými radmi.

1.1.1 Typy časových radov

Existuje mnoho klasifikačných kritérií pre časové rady [6]:

- **podľa rozhodujúceho časového hľadiska**

a) diskrétne (okamihové) časové rady – sú merané v pravidelne rozložených časových krokoch (mesačne, štvrťročne, ročne). Tieto časové rady sú veľmi bežné ako napr. sledovanie miery nezamestnanosti, ktorá sa zvyčajne meria každý štvrťrok.

b) spojité (intervalové) časové rady - majú tendenciu byť hustejšie, keďže sú priebežne zaznamenávané počas určitého intervalu a ich hodnota závisí od dĺžky tohto intervalu. Príkladom týchto časových radov je EEG alebo merania senzorov.

- **podľa periodicity sledovania**

a) dlhodobé časové rady – pozorovania sa uskutočňujú za ročné alebo dlhšie obdobie

b) krátkodobé časové rady – pozorovania sa uskutočňujú za obdobie kratšie než jeden rok

- **podľa časovej závislosti** - označuje vplyv minulých hodnôt na novo pozorované hodnoty zaznamenanej premennej:

a) časové rady s dlhou pamäťou - sú tie, ktoré majú funkciu autokorelácie, ktorá pomaly klesá. Opisujú proces, ktorý sa pomaly mení. Stretneme sa s nimi zvyčajne v meteorologických, geologických údajoch.

b) časové rady s krátkou pamäťou - opisujú proces s rýchlym obratom. Majú funkciu autokorelácie, ktorá rýchlo klesá, čím dlhšie sme od súčasnosti, tým je opatrenie menej užitočné pre budúcnosť.

- **podľa stacionarity:**

a) stacionárne časové rady - majú štatistické vlastnosti (stredná hodnota, rozptyl), ktoré nezávisia od času. Majú konštantnú strednú hodnotu aj rozptyl.

b) nestacionárne časové rady – sú rady, ktoré nezodpovedajú vyššie uvedenému popisu. Tieto časové rady sú veľmi bežné vo finančnom a maloobchodnom sektore. Tieto typy časových radov možno ťažko predpovedať bez akéhokoľvek typu predbežného spracovania, preto používame viacero techník na stacionarizáciu týchto procesov, aby sme mohli lepšie predpovedať.

1.1.2 Analýza časových radov

Analýza časových radov sa vykonáva hlavne z dvoch dôvodov [2]:

- Pochopenie správania sa procesu štúdiom jeho minulých hodnôt, aby bolo možné modelovať a identifikovať hlavné parametre, ktoré ovplyvňujú časový rad a identifikovať jeho komponenty.
- Prognóza budúcich hodnôt časového radu pomocou vhodného modelu, podľa minulých hodnôt.

Existujú rôzne metódy analýzy časových radov, najzakladanejšie z nich popíšeme v tejto podkapitole. Výber metódy pri analýze daného časového radu záleží na účele analýzy a type časového radu [5].

Dekompozícia časových radov

Dekompozícia časových radov je štatistická úloha, ktorá rozkladá časový rad na viacero zložiek ako je trend, sezónna, cyklická a náhodná zložka. Každá zložka predstavuje jednu zo základných kategórií vzorov. Analýzou časových radov pomocou metódy dekompozície, sa budeme ďalej v práci zaoberať v podkapitole 1.3.3 [6].

Box-Jenkinsova metodológia

Box-Jenkinsova metodológia považuje za základný prvok konštrukcie modelu časového radu reziduálnu zložku, ktorá musí byť tvorená korelovanými náhodnými veličinami. Kladie sa teda dôraz na korelačnú analýzu. Predpokladom tohto postupu je dlhší časový rad. Základnými modelmi sú autoregresné procesy AR , procesy kĺzavých súčtov MA, zmiešaný model ARMA a integrovaný model ARIMA. Ďalej sa tejto metóde budeme venovať podkapitole 3.1 a 3.2 [7].

1.2 Stochastické procesy a ich vlastnosti

Bežným prístupom pri analýze údajov časových radov je považovať pozorované časové rady za súčasť realizácie stochastického procesu. Pred definovaním stochastických procesov sú potrebné dve zbežné definície [2].

Pravdepodobnostný priestor je trojitý (Ω, S, P), kde:

- Ω je priestor elementárnych udalostí ω

- *S* je σ-algebra podmnožín priestoru *Ω*, t. j. podmnožín priestoru elementárnych udalostí
- P je miera pravdepodobnosti definovaná pre všetkých členov *S*

Skutočná náhodná premenná alebo reálna stochastická premenná na *(Ω,S,P)* je funkcia x: Ω→R, takže inverzný obraz ľubovoľného intervalu *(-∞,∞)* patrí *S*, to je merateľná funkcia.

1.2.1 Stochastické procesy

Stochastický proces je v čase usporiadaná postupnosť náhodných veličín *{y(ω,t), ω∈ S, t∈T}*, kde *S* nazývame výberový priestor a *T* je indexová množina [2].

Usporiadaná postupnosť čísel, teda časový rad je stochastický proces definovaný na indexovej množine T. Podľa indexovej množiny času rozlišujeme dva typy náhodných procesov a to:

- diskrétny stochastický proces
- kontinuálny stochastický proces

Stochastický proces je charakterizovaný n-dimenzionálnou distribučnou funkciou. Konečná dimenzionálna distribúcia stochastického procesu je definovaná ako množina všetkých spoločných distribučných funkcií pre všetky konečné celočíselné množiny *T* ľubovoľnej veľkosti, to je počet pozorovaní, ktorý sa nazýva dĺžka časového radu *n*. Každé reálne číslo $k \in R$, že $t_{j+k} \in T$, $j = 1, 2, 3, \ldots, n$. Pre diskrétny proces je to teda množina:

$$F(y_{t1}, y_{t2}, y_{t3}, \ldots, y_{tn}) = F(y_{t1+k}, y_{t2+k}, y_{t3+k}, \ldots, y_{tn+k})$$

Pre každý stochastický proces platia nasledovne definované charakteristiky:

- Stredná hodnota: $\mu_t = E(y_t)$, pre $t = 1, 2, 3, \ldots, n$
- Rozptyl: $D(y_t) = E[(y_t - \mu_t)^2] = \sigma_t^2$, pre $t = 1, 2, 3, \ldots, n$
- Autokovariancia $\gamma(t_i, t_j) = E[(y_{t_i} - \mu_{t_i})(y_{t_j} - \mu_{t_j})]$

Mnohorozmerný stochastický proces je množina reálnych náhodných premenných *Y={y_t(ω,t) ω∈Ω, t∈T}*, všetky sú definované na rovnakom pravdepodobnostnom priestore *(Ω, S, P)*.

1.2.2 Stacionarita

Teraz si postupne zadefinujeme jednotlivé druhy stacionarity, ktoré rozlišujeme v teórii stochastických procesov. Stacionárnosť znamená, že štatistické vlastnosti procesu generujúceho časový rad sa v priebehu času nemenia [2].

Stacionárny rad je taký, ktorého štatistické vlastnosti ako priemer, rozptyl a autokorelácia sú konštantné a nezávisia od času. Matematicky zapíšeme vlastnosti stacionárneho radu:

- Konštantný priemer: $\mu_t = E(y_t) = \mu$, pre $t = 1, 2, 3, \dots, n$
- Konštatný rozptyl: $\sigma_t^2 = D(y_t) = E((y_t - \mu_t)^2 = \sigma^2$, pre $t = 1, 2, 3, \dots, n$
- Konzistentná kovariancia: $cov(y_n, y_{n+k}) = cov(y_m, y_{m+k})$ – vyžadujeme, aby medzi údajmi bola konzistentná kovariancia, teda inými slovami, chceme mať identickú vzdialenosť od seba napr. $cov(y_1, y_4) = cov(y_3, y_6)$.

Stacionárne údaje sú ploché série bez trendu, majú konštantný rozptyl v čase, konštantnú autokorelačnú štruktúru v čase a žiadne periodické výkyvy založené na sezónnosti.

Rozlišujeme:

- silnú stacionaritu - v čase sa nemenia všetky momenty pravdepodobnosti. Proces y_t je silno stacionárny, ak pre každé k, t a n je rozdelenie $y_t,\dots,y_{t+k}$ rovnaké ako pre rozdelenie $y_{t+n}, \dots, y_{t+k+n}$.
- slabá stacionarita – rozptyl, priemer a kovariancia sa v čase nemenia. y_t je slabo stacionárne ak $E(y_t)$, $D(y_t)$ a $cov(y_t, y_{t+k})$ nezávisia od t. Biely šum je významným príkladom slabej formy stacionárneho procesu. Biely šum je postupnosť náhodných premenných ε_t.

Na modelovanie je potrebné mať stacionárne časové rady. Keď dáta sú nestacionárne, musíme ich diferencovať, aby sme ich urobili stacionárnymi. Aby sme zistili, či časový rad je stacionárny, používame testy stacionarity, o ktorých si povieme v kapitole 3.3.3.

1.2.3 Autokorelačná funkcia (ACF) a parciálna autokorelačná funkcia (PACF)

ACF je autokorelačná funkcia, ktorá počíta korelácie, ktoré existujú medzi y_t a y_{t+k}, kde $k = 0,1,2,3,...$ [11].

Výsledkom tejto funkcie je graf s názvom korelogram, tento graf je dôležitým krokom v modelovaní a prognózovaní časových radov, pretože pomáha prísť s vhodným modelom, a taktiež je dôležitou súčasťou overovania parametrov modelu, pretože dokáže odhaliť prítomnosť periodického vzoru, ktorý môže byť skrytým bielym šumom. V grafe ACF os x vyjadruje korelačný koeficient, zatiaľ čo os y uvádza počet oneskorení. ACF je teda symetrická okolo k, nadobúda hodnoty z intervalu $<-1, 1>$.

ACF berie do úvahy všetky komponenty časového radu ako je trend, sezónna, cyklická a reziduálna zložka pri hľadaní korelácií, preto ide o úplný graf autokorelácie.

Výsledkom funkcie normovania autokovariancie príslušnými rozptylmi je autokorelačná funkcia. Pre stacionárne procesy je to funkcia nezávislá od času vyjadrená vzorcom:

$$\rho_k = \frac{\gamma_k}{\sigma_{y,t} \cdot \sigma_{y,t-k}} = \frac{\gamma_k}{\sigma_y^2} = \frac{\gamma_k}{\gamma_0}$$

ACF má vlastnosti:

- $\rho(0) = 1$
- $\rho(k) = \rho(-k)$
- $|\rho(k)| \leq 1$ pre všetky k

Ako ďalšia metóda na určenie časových závislostí medzi nedávnymi a minulými hodnotami používame aj funkciu čiastočnej autokorelácie.

PACF je autokorelačná funkcia, ktorá vysvetľuje čiastočnú koreláciu medzi sériou a oneskorením. Parciálne autokorelácie nám poskytujú informáciu o korelácii premenných y_t a y_{t-k}, ktorá je očistená o vplyv premenných, ktoré ležia medzi nimi $y_{t-1}, y_{t-2}, \ldots y_{t-k+1}$. V literatúre sa parciálna autokorelácia označuje ako ϕ_{kk}. Počítame ju ako podmienenú strednú hodnotu vyjadrenú vzorcom [11]:

$$\phi_{k,k} = E[(y_t - \mu)(y_{t-1} - \mu)(y_{t-2} - \mu) \ldots . (y_{t-k+1} - \mu)]$$

1.2.4 Operátor spätného posunutia

Nech $\{y_t,\ t \in T\}$ je stochastický proces pre operátor spätného posunutia, ktorý sa označuje B, platí [1]:

$$By_t = y_{t-1}$$

Vráti hodnotu procesu posunutú o jedno obdobie dozadu.

Tento operátor môžeme aplikovať niekoľkokrát za sebou. Platí:

$$B^j y_t = y_{t-1}$$

1.2.5 Dekompozícia časového radu

V roku 1919 W.M. Persons ako prvý navrhol dekompozíciu časového radu na niekoľko zložiek [5].

Uviedol, že časový rad sa skladá z týchto 4 zložiek:

a. Trendová zložka T_t: vyjadruje dlhodobú tendenciu vývoja skúmaného javu
b. Sezónna zložka S_t: vyjadruje pravidelné kolísanie okolo trendu v rámci jedného roka
c. Cyklická zložka C_t: vyjadruje kolísanie okolo trendu, jednotlivé cykly sú dlhšie ako jeden rok, a majú nepravidelný charakter
d. Reziduálna zložka I_t: vyjadruje náhodné chyby

Časové rady vždy obsahujú reziduálnu zložku. Trendovú, cyklickú a sezónnu zložku môžu, ale nemusia obsahovať.

Vyššie spomenuté zložky sú usporiadané do matematického modelu dvoma štruktúrami:

1. Aditívny model, ktorý má hodnoty časového radu určené ako súčet hodnôt jednotlivých zložiek:

$$y_t = T_t + S_t + C_t + I_t$$

2. Multiplikatívny model, ktorý má hodnoty časového radu určené ako súčin hodnôt jednotlivých zložiek:

$$y_t = T_t * S_t * C_t * I_t$$

V praxi aditívnu dekompozíciu používame v prípade, že variabilita hodnôt časového radu je v čase približne konštantná, zatiaľ čo multiplikatívnu dekompozíciu zvolíme, pokiaľ

sa variabilita v čase mení. Oba modely sú však ekvivalentné v tom zmysle, že sa na seba dajú previesť logaritmickou transformáciou.

Zložky časového radu:

Trend - vyjadruje dlhodobé zmeny ako je dlhodobý rast alebo dlhodobý pokles. Trend v časových radoch možno vyjadriť pomocou trendových funkcií a kĺzavých priemerov. Najčastejšie používané typy trendov sú: lineárny , kvadratický a exponenciálny trend.

Sezónna zložka - vyjadruje periodické zmeny v časovom rade, ktorého perióda sa rovná určitej štandardnej časovej jednotke, ako sú týždne, mesiace, roky atď. Ide o zmeny, ktoré sa pravidelne opakujú.

Cyklická zložka - je periodická zložka, ktorej perióda nezodpovedá kalendárnym jednotkám. Hovoríme o nepravidelnej fluktuácii okolo trendu, v ktorej sa strieda fáza rastu s fázou poklesu (napr. ekonomický cyklus). Odhaduje sa metódami spektrálnej analýzy.

Reziduálna zložka - sa vždy nachádza v časovom rade. Je tvorená náhodnými pohybmi v priebehu časového radu, ktoré nemajú jasný systematický charakter. Zahŕňa aj chyby v meraní údajov. Predpokladáme, že reziduálna zložka je biely šum - nekorelované náhodné veličiny.

Charakter bieleho šumu znamená, že náhodné veličiny sú navzájom lineárne nezávislé, majú normálne rozdelenie s nulovou strednou hodnotou a konštantným rozptylom. Matematicky to môžeme vyjadriť:

- $\mu(\varepsilon_t) = 0$
- $\mu(\varepsilon^2) = \sigma_\varepsilon^2$ = konštantný
- $cov(\varepsilon_t, \varepsilon_{t-k}) = 0,\ k \neq 0$
- $\varepsilon_t \sim N(0, \sigma^2_\varepsilon)$

ε_t pre $t = 1, 2, ..., T$ je stacionárny proces

1.3 Programovací jazyk Python

Python je vysokoúrovňový, objektovo orientovaný dynamický programovací jazyk. Vytvoril ho holandský programátor Guido van Rossum v roku 1991 [8]. Je to jeden z najpopulárnejších a najpoužívaných programovacích jazykov súčasnosti. Je to tiež jeden z najviac rozvíjajúcich sa programovacích jazykov. Jeho jednoduché používanie, flexibilita,

dobre štruktúrovaný a jasný syntax, čitateľnosť, presnosť a rýchlosť robia z jazyka najlepšiu voľbu. Je to tiež jeden z mála programovacích jazykov, ktorý podporuje matematicky náročné aplikácie. Je široko používaný mnohými veľkými organizáciami. Zdrojový kód Pythonu je čitateľný a veľmi ľahko sa učí a používa na rozdiel od niektorých iných programovacích jazykov, ako sú Java alebo Ruby. Má menej riadkov kódu v ľahkej a jednoduchej forme. Dodáva sa s predinštalovanými knižnicami, ako sú Pandas, Numpy, SciPy a ďalšie s vopred napísanými komponentmi [3].

Používanie Pythonu zahŕňa viacero oblastí, ako je analýza údajov, ML, programovanie webových prehľadávačov, webové servery, webové analyzátory, automatizované testy, počítačová grafika, vývoj aplikácií, vývoj desktopov a ďalšie. Taktiež je ideálnou voľbou pre umelú inteligenciu a strojové učenie.

1.3.1 Syntax jazyka Python

Python má širokú škálu knižníc s otvoreným zdrojovým kódom. Niekoľko takýchto open source knižníc, ktoré budeme používať v nasledujúcich kapitolách, si predstavíme nižšie. Tieto knižnice sú veľmi dôležité na analýzu časových radov. V tejto podkapitole sme pracovali hlavne s publikáciou [3].

Knižnice:

Datetime - Táto knižnica je obzvlášť dôležitá na prácu s časovými radmi. Poskytuje všetky potrebné funkcie dátumu a času na čítanie, formátovanie a manipuláciu s časom.

NumPy - Numerical Python je knižnica používaná na vedecké výpočty. Pracuje na objekte n-dimenzionálneho poľa a poskytuje základné matematické funkcie, ako je veľkosť, tvar, priemer, štandardná odchýlka, minimum, maximum, ako aj niektoré zložitejšie funkcie, ako sú lineárne algebraické funkcie a Fourierova transformácia.

Pandas - Táto knižnica poskytuje vysoko efektívne a ľahko použiteľné dátové štruktúry, ako sú série, dátové rámce a panely. Vylepšila funkčnosť Pythonu od jednoduchého zberu a prípravy údajov až po analýzu údajov.

SciPy - Science Python je knižnica používaná na vedecké a technické výpočty. Poskytuje funkcie pre optimalizáciu, spracovanie signálu a obrazu, integráciu, interpoláciu a lineárnu algebru. Táto knižnica sa hodí pri vykonávaní strojového učenia.

Statsmodels - Táto knižnica sa používa na prieskum štatistických údajov a štatistické modelovanie.

Matplotlib - Táto knižnica sa používa na vizualizáciu údajov v rôznych formátoch, ako je čiarový graf, stĺpcový graf, rozptylové grafy, histogram atď. Obsahuje všetky funkcie súvisiace s grafom, od vykresľovania až po označovanie.

1.3.2 Výhody jazyka Python

- Je všestranný, prehľadný, ľahko sa používa a učí, je čitateľný a dobre štruktúrovaný.
- Je open source, čo znamená, že Python sa dá stiahnuť zadarmo a napísať asynchrónny kód v priebehu niekoľkých minút.
- Má všetky knižnice, aké si dokážeme predstaviť. Vďaka tomu je ideálny pre všetky prípady použitia, ktoré sme si spomenuli vyššie, ako je vývoj mobilných aplikácií, strojové učenie, dátovú analýzu, atď.
- Skvelý pre prototypy. Môžeme urobiť viac s menším množstvom kódu.
- V porovnaní s inými kódovacími jazykmi je Python oveľa produktívnejší jazyk, pretože je dynamicky a výstižne písaný [13].

2 Cieľ práce

Cieľom bakalárskej práce je vysvetliť, akým spôsobom sa dá použiť jazyk Python so svojimi knižnicami na analýzu časových radov. Na ukážku možností jazyka Python v tejto oblasti sme vybrali viacero dôležitých ekonomických premenných, pri ktorých overíme, či dátový set je vhodný na použitie zvolenej metodológie a následne s využitím modelov s najvhodnejšími vlastnosťami robíme predikciu vývoja týchto premenných. Na naplnenie hlavného cieľa práce a lepšiu špecifikáciu našej problematiky si uvedieme čiastkové ciele.

1. Čiastkový cieľ – prehľad skúmanej problematiky - časové rady a ich vlastnosti, jazyk Python
2. Čiastkový cieľ – teoretické spracovanie lineárnych modelov analýzy časových radov a ukazovateľov na identifikáciu, overenie a verifikáciu modelu
3. Čiastkový cieľ – samotná práca v jazyku Python - tvorba modelov s danými dátami
4. Čiastkový cieľ – vyhodnotenie modelov na základe vybraných ukazovateľov

3 Metodika práce a metódy skúmania

3.1 Lineárne modely stacionárnych stochastických procesov

V tejto kapitole si predstavíme základné modely Box-Jenkinsovej metodológie vytvorenej Boxom a Jenkinsom, ktorá bola prvýkrát popísaná v roku 1976 [2]. Základnými modelmi sú autoregresné procesy AR(p) , procesy kĺzavých priemerov MA(q), zmiešaný model ARMA(p,q). Pomocou týchto lineárnych modelov môžeme modelovať lineárne stacionárne stochastické procesy. Za základný prvok konštrukcie modelu sa považuje reziduálna zložka časového radu, ktorá je tvorená korelovanými alebo závislými náhodnými veličinami.

3.1.1 Autoregresné procesy (AR)

Autoregresný model, označený ako AR(p) je najjednoduchší model pre časové rady. Matematicky táto formulácia znamená, že súčasná hodnota časového radu y_t, bude lineárnou kombináciou minulých hodnôt $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, vynásobených číselným faktorom $\varphi i \in (1, 2, .., p)$ plus náhodný šum ε_t. Použitím PACF môžeme identifikovať rád AR procesu [2].

Autoregresný model rádu p AR(p) je popísaný rovnicou:

$$y_t = C + \varphi y_{t-1} + \varphi y_{t-2} + \cdots + \varphi y_{t-p} + \varepsilon_t$$

C – je konštantný prienik

$y_{(t-1)}$ - hodnoty počas predchádzajúceho obdobia

p - počet oneskorení

φ - vypočítaná číselná konštanta, ktorou násobíme oneskorenú premennú

ε - reziduál – rozdiel medzi našou predpoveďou a správnou hodnotou – biely šum

3.1.2 Procesy kĺzavých priemerov (MA)

Jedným z ďalších základných modelov analýzy časových radov je model kĺzavého priemeru, označovaný ako MA(q). Model kĺzavého priemeru je podobný autoregresívnemu modelu s tým rozdielom, že priesečník je priemerom najnovších hodnôt q, a namiesto vykonania autoregresného modelu na pozorovaných záznamoch prebieha na rezíduách. Matematická formulácia znamená, že model MA je lineárna kombinácia posledných q

zvyškov s $\theta i \in (1,..,\ q)$ váhou každej korekcie chýb. Použitím ACF môžeme identifikovať rád MA procesu [2].

Matematicky vyjadrujeme proces všeobecného kĺzavého priemeru takto:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

θ_q – číselný koeficient pre hodnotu spojenú s q oneskorením

ε_t – zvyšky za bežné obdobie

ε_{t-1} – zvyšky za minulé obdobie

3.1.3 ARMA procesy

Model ARMA(p,q) je kombináciou vyššie uvedených modelov AR(p) a MA(q), matematicky je model jednoducho súčtom týchto modelov. Bol vytvorený matematikmi Box a Jenkins v roku 1970 [2]. ARMA model je stacionárny, keď AR(p) je stacionárne a je invertibilný, keď MA(q) je invertibilné [4].

Matematicky vyjadrujeme proces všeobecného ARMA (p, q) modelu takto:

$$y_t = \mathrm{C} + \varphi y_{t-1} + \cdots + \varphi y_{t-p} \varepsilon_t + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

3.2 Lineárne modely nestacionárnych procesov

V reálnom živote často narážame na časové rady s nestacionárnými stochastickými procesmi. Nestacionarita procesu môže byť spôsobená buď v čase meniacou sa strednou hodnotou, alebo v čase meniacim sa rozptylom.

3.2.1 ARIMA procesy

Najdôležitejšou charakteristikou modelu ARIMA je, že dokáže spracovať nestacionárne časové rady vďaka parametru d, ktorý definuje počet vykonaných diferencií na dátach. Pri tvorbe modelu ARIMA nemusíme mať stacionárne dáta, ale časový rad musí byť prevoditeľný na stacionárny časový rad prvou alebo vyššou diferenciou. V modeli ARIMA môžeme vyjadriť diferenčný operátor nasledovne [2]:

$$\begin{aligned} \Delta y_t &= y_t - y_{t-1} \\ &= (1 - B) y_t \end{aligned}$$

Model ARIMA sa skladá z troch *(p, d, q)*, ktoré predstavuje:

p - počet autoregresívnych parametrov (AR)

d - diferenciačný operátor

q - počet kĺzavých priemerných parametrov (MA)

Aj keď je model ARIMA lepšie vybavený na spracovanie časových radov, stále mu chýba schopnosť spracovať sezónne údaje.

3.2.2 SARIMA procesy

Model SARIMA(p, d, q)(P, D, Q)^s je zovšeobecnením modelu ARIMA, ktorý je vybavený na spracovanie zložky sezónnosti periodických časových radov. Vo svojom zovšeobecnenom návrhu Box a Jenkins vykonávajú druhú diferenciáciu na sezónnu zložku časového radu. Do modelu teda pridáme ďalšie štyri parametre, ktoré budú *(P, D, Q)*^{s}. Model bude zapísaný nasledovne [2]:

$$y_t - y_{t-s} = (1 - B^S)y_t$$

P – rád sezónneho AR modelu

Q- rád sezónneho MA modelu

D – rád sezónnej diferencie

S – je dĺžka sezónnej periódy

Model SARIMA je zovšeobecnením všetkých vyššie spomenutých modelov, ktoré zvládajú sezónne a nestacionárne časové rady.

Tabuľka č. 1: Metodika pre autoregresné modely

	Stacionárnosť	Nestacionárnosť	Sezónnosť
AR	Áno	Nie	Nie
MA	Áno	Nie	Nie
ARMA	Áno	Nie	Nie
ARIMA	Áno	Áno	Nie
SARIMA	Áno	Áno	Áno

Zdroj dát: [2], vlastné spracovanie

3.3 Informačné kritéria a testovacie hypotézy

Informačné kritérium predstavuje mieru kvality štatistického modelu. Používa sa na určenie vhodného rádu modelu, a porovnanie alternatívnych modelov prispôsobených rovnakému súboru údajov. Informačné kritérium berie do úvahy ako dobre model zodpovedá údajom a komplexnosť modelu. Testovacie hypotézy overujú vlastnosti modelu. Miery presnosti modelu overujú pravosť a dôveryhodnosť modelu.

3.3.1 AIC – Akaikeho informačné kritérium

Akaikeho informačné kritérium AIC je jedno z najpoužívanejších informačných kritérií. Navrhol ho štatistik Hirotugu Akaike v roku 1973 [12]. Toto kritérium je užitočné pri výbere poradia *(p, d, q)* modelu ARIMA.

AIC je matematicky vyjadrené ako:

$$AIC = -2\frac{\ln L}{T} + \frac{2}{T}k$$

L - pravdepodobnosť údajov

k – počet parametrov

T – dĺžka časového radu

V praxi vyberáme model s najnižším AIC v porovnaní s ostatnými modelmi. Je dôležité poznamenať, že AIC nemožno použiť na výber poradia diferenciácie *d*. AIC modelov s rôznymi rádmi diferenciácie, preto nie sú porovnateľné.

3.3.2 BIC – Bayesovo informačné kritérium

Bayesovo informačné kritérium, známe aj ako Schwarz Criterion, je ďalším štatistickým meradlom na porovnávacie hodnotenie medzi modelmi časových radov. Vyvinul ho štatistik Gideon Schwarz a úzko súvisí s AIC. Rozdiel medzi BIC a AIC sa prejaví, keď pridáme niekoľko *k* parametrov, aby sa zvýšila správnosť prispôsobenia modelu. V takom prípade BIC viac (v porovnaní s AIC) penalizuje takéto zvýšenie parametrov [12].

BIC je matematicky vyjadrené ako:

$$BIC = -2\frac{\ln L}{T} + \frac{ln(T)}{T}$$

L - pravdepodobnosť údajov

k – počet parametrov

T – dĺžka časového radu

Podobne ako pri AIC, medzi rôznymi alternatívnymi modelmi sa uprednostňuje model s minimálnou hodnotou BIC.

3.3.3 Test stacionarity dát

Predtým ako začneme modelovať dáta, musíme zistiť, či sú dáta stacionárne. Či je súbor údajov stacionárny, alebo nie zisťujeme pomocou štatistických testov. Najpoužívanejšie sú tieto dva testy [9]:

- Dickey-Fuller test (ADF)
- Test Kwiatkowski-Phillips-Schmidt-Shin (KPSS)

Dickey-Fullerov test bol navrhnutý v roku 1984. Tento test je najpopulárnejší štatistický test s nasledujúcimi predpokladmi.

H0: Časový rad má jednotkový koreň, čo znamená, že je nestacionárny.
H1: Časový rad nemá jednotkový koreň, čo znamená, že je stacionárny.

Na vyhodnotenie sa používa t-test významnosti s tabuľkovými kritickými hodnotami alebo sa používa p-hodnota porovnávajúca s hladinou významnosti α.

p-hodnota > 0,05: Nezamietame H0, dáta majú jednotkový koreň a sú nestacionárne.
p-hodnota < 0,05: Zamietame H0, dáta nemajú jednotkový koreň a sú stacionárne.

3.3.4 Test autokorelácie

Taktiež pred začatím modelovania dát, musíme zistiť, či sa nenachádza v dátach autokorelácia. Prítomnosť autokorelácie zisťujeme pomocou štatistických testov. V tejto podkapitole sme čerpali z publikácie [10].

Najpoužívanejšie sú tieto tri testy:

- Durbinov – Watsonov test
- Ljungovej-Boxov test
- Breuschov-Godfreyov test

Ljungov-Box test vznikol v roku 1978. Test sa aplikuje na rezíduá časových radov po prispôsobení modelu ARMA(p,q) údajom. Ak sú autokorelácie veľmi malé, dôjdeme k

záveru, že model nevykazuje významný nedostatok prispôsobenia. Zatiaľ čo Ljung-Box možno použiť pre akúkoľvek hodnotu oneskorenia. Durbin Watson možno použiť iba pre oneskorenie 1.rádu.

Nulová a Alternatívna hypotéza je pre všetky testy rovnaká:

H0: V modeli nie je prítomná autokorelácia, $\rho_0 = \rho_1 = \rho_2 = 0$.

H1: V modeli je prítomná autokorelácia, existuje aspoň jedna $\rho_i \neq 0$.

Testovacia štatistika *Q-stat*:

$$Qstat = n(n+2)\sum_{k=1}^{m}\frac{\hat{p}_k^2}{n-k}$$

$\hat{p}_k^2$ - odhadovaná autokorelácia série pri oneskorení

k - oneskorenie

m - počet testovaných oneskorení

Na vyhodnotenie sa používa chí-kvadrát tabuľky $\chi^2_{1-\alpha}(h)$ s *h* stupňami voľnosti a hladinou významnosti α. Pretože test je aplikovaný na rezíduá, stupne voľnosti musia zodpovedať odhadovaným parametrom modelu tak, že *h=m-p-q*, kde *p* a *q* označujú počet parametrov z modelu ARMA (p,q) zodpovedajúcich údajov.

Ak *Q-stat* > $\chi^2_{1-\alpha}(h)$, *h=m-p-q*, resp. p-value < α - zamietame H0

Ak *Q-stat* < $\chi^2_{1-\alpha}$ (h), h=m-p-q, resp. p-value > α - nezamietame H0

3.3.5 Koeficient determinácie

Koeficient determinácie označovaný ako R^2 nám vyjadruje podiel celkovej variability závislej premennej, ktorú vysvetľuje model. Jeho hodnota je vyjadrená v percentách je v intervale *<0,1>*, pričom vyššia hodnota signalizuje lepší model. Počíta sa podľa vzťahu [10]:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

3.3.6 Miery presnosti vyrovnania, resp. priemerné charakteristiky rezíduí

Tieto hodnoty nám slúžia na overenie predpovede a na vyhodnotenie výkonnosti modelu [10].

Priemerná absolútna chyba MAE (mean absolute error):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y_t}|$$

Priemerná štvorcová chyba – rozptyl MSE (mean squared error):

$$MSE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y_t}|^2$$

Stredná štvorcová chyba odhadu RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y_t}|^2}$$

Používa sa najčastejšie na vyhodnotenie presnosti modelu.

Priemerná absolútna percentuálna chyba MAPE (mean absolute percentage error):

$$MAPE = \frac{1}{T}\sum_{t=1}^{n}\frac{|y_t - \widehat{y_t}|}{y_t} * 100$$

Prognózy s hodnotou MAPE do 5% sa považujú za veľmi dobré. MAPE nad 10% poukazuje na chybné prognózy.

Keď porovnávame vhodnosť modelu mierami presnosti, predikčnými vlastnosťami, za najvhodnejší model považujeme model s najnižšími hodnotami vyššie uvedených charakteristík. Ak v uvedených mierach nie je významný rozdiel, potom vyberieme model, ktorý je jednoduchší a má menší počet parametrov.

3.4 Tvorba modelu

Box-Jenkinsova metodológia má nasledovný postup:

- identifikácia modelu
- odhad parametrov a odhad štatistickej významnosti
- verifikácia modelu

1. Identifikácia modelu

Prvá fáza tvorby modelu je jednou z najťažších. Táto úloha spočíva v rozhodnutí, aký typ modelu vybrať. Ako prvé treba overiť, či je daný časový rad stacionárny. Ak zistíme, že časový rad je nestacionárny, treba ho zdiferencovať. Po stacionarizácii

pokračujeme skúmaním priebehu ACF a PACF. Podľa vlastností ACF a PACF procesov, vieme približne odhadnúť vhodný model AR(p), MA(q) a ARMA (p,q), a teda určiť parametre p a q. Dobré je rátať aj s alternatívnymi navrhnutými modelmi, ktoré sa neskôr môžu prejaviť ako vhodnejšie.

2. Odhad parametrov

Pre odhad parametrov modelov sa používa väčšinou nelineárna metóda najmenších štvorcov. Metóda najmenších štvorcov je minimalizovať sumy štvorcov (druhých mocnín) reziduálov medzi skutočnou a vyrovnanou hodnotou závislej premennej. Hľadáme také odhady parametrov, ktoré spĺňajú vzťah:

$$MIN \sum_{t=1}^{n} (e_t)^2 = MIN \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

3. Verifikácia modelu

Posledná fáza je verifikácia modelu, ktorá spočíva v overení odhadnutého modelu, či spĺňa všetky verifikačné kritéria a podmienky kvality modelu a či nie je potrebné upraviť odhadnutý model. A to preveríme informačnými kritériami ako je AIC, BIC, mierami presnosti modelu a testovacími hypotézami, ktoré sú vyššie uvedené v tejto bakalárskej práci v podkapitole 3.3.

4 Výsledky práce

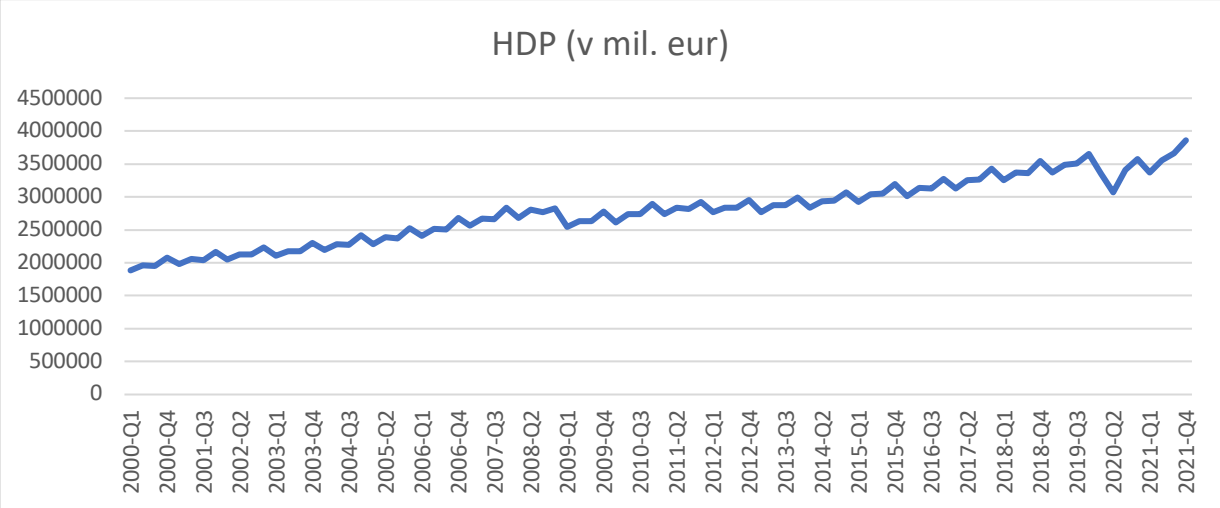
Témou tejto bakalárskej práce sú základy analýzy časových radov s využitím jazyka Python. Pre praktickú časť bakalárskej práce sme sa snažili o výber dát, ktoré sú relevantné z ekonomického hľadiska a existuje k ním dôveryhodný zdroj historických dát za dostatočne dlhé časové obdobie. Analyzované dáta všetkých troch časových radov sú uvedené v Prílohe 1 tejto bakalárskej práce.

4.1 Analyzované dáta

4.1.1 Hrubý domáci produkt HDP Európskej únie

Hrubý domáci produkt (HDP) je najčastejšie používaným meradlom veľkosti ekonomiky. Je dôležitým ukazovateľom trvalo udržateľného rozvoja. Ide o základný makroekonomický ukazovateľ, ktorý meria celkovú ekonomickú aktivitu krajiny. Jeho množstvo závisí od počtu reprodukčných procesov a z hľadiska národného hospodárstva predstavuje kolobeh všetkých výrobných faktorov v krajine a prostredníctvom nich vyrobených tovarov a služieb. Je to hodnota všetkých finálnych výrobkov a služieb, ktoré ročne vyprodukuje rezidentská jednotka krajiny. Zodpovedá takzvanej hrubej pridanej hodnote celej ekonomiky v trhových cenách [17].

Zdrojové dáta sú sezónne neočistené hodnoty HDP Európskej únie. Dáta sú kvartálne a jednotkou je milión eur od dátumu 1.1.2000. Zdrojom dát je databáza štatistického úradu Európskej komisie Eurostat [15].



Obrázok č. 1: Grafické zobrazenie HDP EÚ.

Zdroj dát: Eurostat [15], vlastné spracovanie

Tabuľka č. 2: Základné charakteristiky dát HDP EÚ.

HDP	
Časové obdobie	2000-Q1 – 2021-Q4
Granularita dát	jeden záznam za každý štvrťrok
Počet záznamov	88
Štandardná odchýlka	480 472,1363
Minimálna hodnota	1 884 715,6 (2000-Q1)
Stredná hodnota	2 783 067,15
Maximálna hodnota	3 860 405,1 (2021-Q4)

Zdroj: Eurostat [15], vlastné spracovanie

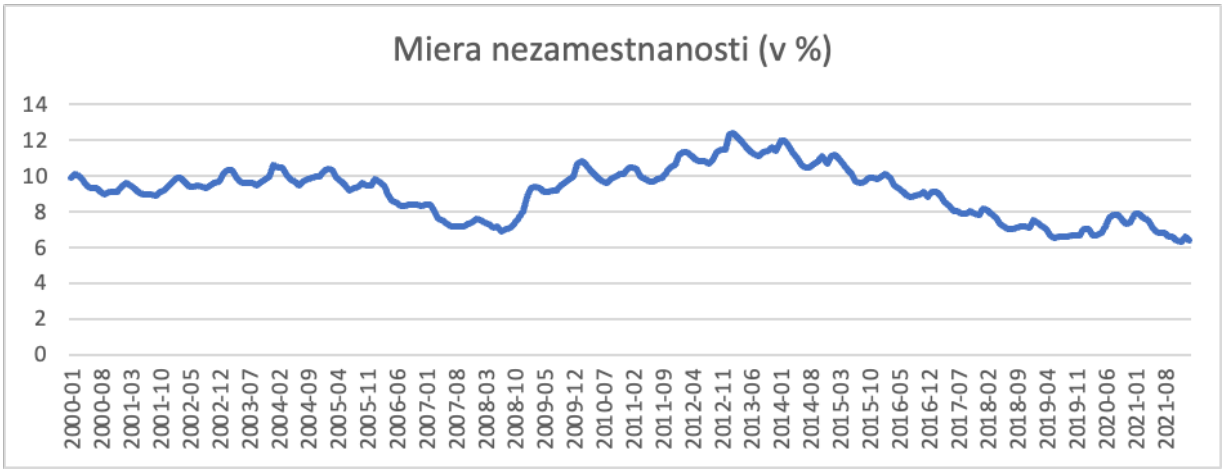
4.1.2 Nezamestnanosť v Európskej únii

Ďalším typom dát, ktoré sme si vybrali na analýzu, je vývoj nezamestnanosti v Európskej únii. Ide o stav, pri ktorom sa časť pracovných síl nezúčastňuje pracovného procesu. Eurostat definuje nezamestnanú osobu podľa pokynov Medzinárodnej organizácie práce (ILO), a to ako všetky práceschopné osoby vo veku od 15 do 74 rokov, ktoré si na trhu práce nemôžu nájsť platené zamestnanie. Nezamestnanosť je následkom nerovnováhy medzi dopytom a ponukou na trhu práce. Štatistiku nezamestnanosti je možné prezentovať v absolútnych číslach, inak povedané ako celkový počet osôb, ktoré sú nezamestnané. Z hľadiska jednoduchšieho porovnávania medzi krajinami, v čase a medzi rôznymi sociálnymi skupinami sa v praxi údaje za nezamestnanosť bežne prezentujú ako miera. Miera nezamestnanosti (MN) je hlavným ukazovateľom zamestnanosti [18].

Miera nezamestnanosti je:

$$MN(v\,\%) = \frac{počet\ nezamestnaných}{počet\ ekonomicky\ aktívneho\ obyvateľstva}\ x\ 100$$

Zdrojové dáta sú sezónne neočistené hodnoty miery nezamestnanosti Európskej únie. Dáta sú mesačné a jednotkou je percento obyvateľstva v pracovnej sile od januára roku 2000. Zdrojom dát je databáza štatistického úradu Európskej komisie Eurostat [16].



Obrázok č. 2: Grafické zobrazenie miery nezamestnananosti.

Zdroj dát: Eurostat [16], vlastné spracovanie

Tabuľka č. 3: Základné charakteristiky dát miery nezamestnanosti.

Miera nezamestnanosti	
Časové obdobie	01.2000 – 02.2022
Granularita dát	jeden záznam za každý mesiac
Počet záznamov	266
Štandardná odchýlka	1,4507
Minimálna hodnota	6,3 (12.2021)
Stredná hodnota	9,1278
Maximálna hodnota	12,4 (02.2013)

Zdroj dát: Eurostat [16], vlastné spracovanie

4.1.3 S&P 500 Index

Standard & Poor's 500 Index je vážený index trhovej kapitalizácie najdôležitejších 500 verejne obchodovateľných firiem v USA. Index neobsahuje exaktne 500 najväčších firiem, pretože index je budovaný aj na základe iných kritérií, ako je napríklad sektor trhu. Zoznam sektorov spolu s počtom zastúpených firiem v rámci indexu S&P 500 ako aj celková váha jednotlivých sektorov je uvedená v nasledovnej tabuľke [14].

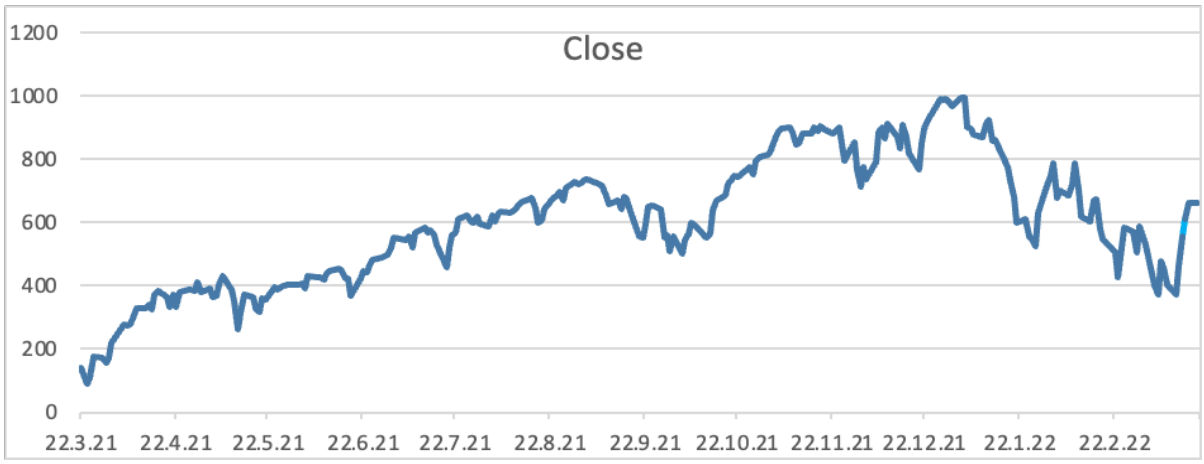
Tabuľka č. 4: Prehľad zloženia S&P 500 Indexu.

Sektor	**Počet firiem**	**Váha**
Information Technology	75	27,3
Health Care	65	13,6
Consumer Discretionary	67	11,6
Communication	60	11,1
Industrials	27	9,3
Consumer Staples	32	6,7
Energy	21	4,4

Utilities	29	2,8
Real Estate	29	2,7
Materials	28	2,5

Zdroj dát:Investopedia [14], vlastné spracovanie

Standard & Poor's prehodnocuje zloženie indexu raz ročne. Index vznikol v roku 1871 za účelom merať výkonnosť amerického trhu. V súčasnosti sa tento index používa hlavne pri pasívnom indexovaní v rámci ETF fondov.



Obrázok č. 3: Grafické zobrazenie S&P 500 Indexu.

Zdroj dát: Marketwatch [19], vlastné spracovanie

SPX je publikovaný v reálnom čase. Pre potreby našej bakalárskej práce sme použili zatváraciu dennú cenu. Zdrojom dát je databáza štatistického úradu Európskej komisie Eurostat, ktorá tieto hodnoty zverejňuje na svojej internetovej stránke [19].

Tabuľka č. 5: Základné charakteristiky dát S&P 500 Indexu.

SPX	
Časové obdobie	22.3.2021 – 21.3.2022
Granularita dát	jeden záznam za každý pracovný deň
Ukazovatele	Close Price
Počet záznamov	253
Štandardná odchýlka	202,0188
Minimálna hodnota	3889,14 (24.3.2021)
Stredná hodnota	4399,23
Maximálna hodnota	4796,56 (3.1.2022)

Zdroj: Marketwatch [19], vlastné spracovanie

4.2 Modelovanie časového radu

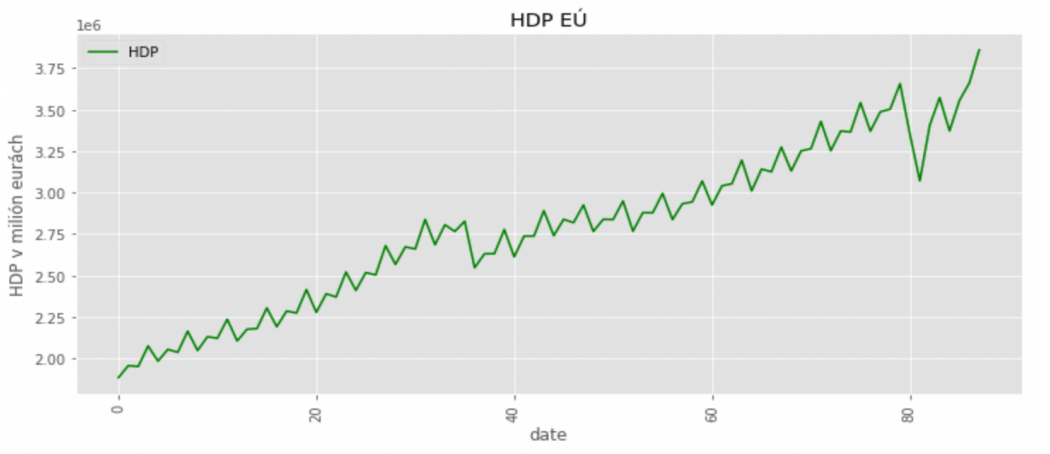
V tejto podkapitole bakalárskej práce uvedieme postup modelovania časových radov, a zameriame sa na výber vhodného modelu pre naše pozorované dáta. Vhodný ARIMA model vyberieme na základe postupu popísaného v podkapitole 3.4. Analýzu vybraných časových radov robíme pomocou programu Python.

Časové rady sú uložené v csv. formáte a obsahujú dva stĺpce - dátum a nameranú hodnotu vrátane hlavičky s názvom časového radu. Po zapnutí programu Python, konkrétne prostredia Jupyter Notebook, ktoré na analýzu využívame, si nainštalujeme stiahnuté knižnice, ktoré budeme potrebovať na prácu s dátami. Ako je Pandas, Numpy, Matplotlib, StatsModels, Seaborn, Pmdarima, atď. Sú popísané v podkapitole 1.3.1 tejto bakalárskej práce.

4.2.1 Modelovanie dát HDP v Európskej únii

Ako prvé budeme modelovať dáta časového radu vývoja HDP Európskej únii. Popisy jednotlivých príkazov sú uvedené v Prílohe 2 tejto bakalárskej práce. Najprv definujeme časový rad s názvom HDP s frekvenciou 4, pretože pracujeme so štvrťročnými dátami. Ďalej pri definícii časového radu zadáme začiatočný moment a to je prvý štvrťrok roku 2000.

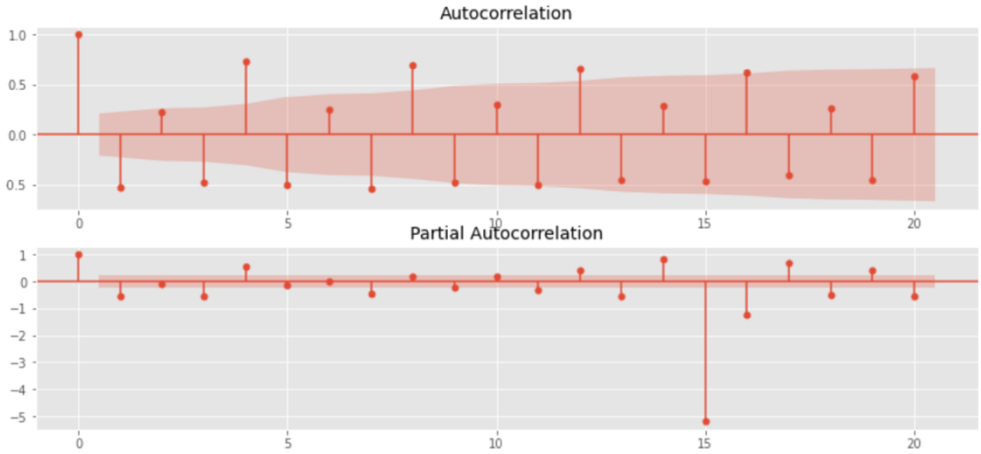
Na nasledovnom obrázku č. 4 môžeme vidieť grafické zobrazenie tohto časového radu, a to štvrťročný vývoj HDP v Európskej únii od 1.štvrťroku roku 2000 do 4.štvrťroku roku 2021.



Obrázok č. 4: Vývoj časového radu hodnôt HDP EÚ.

Ďalším krokom v našej analýze je overenie stacionarity, teda či daný časový rad má jednotkový koreň. Použijeme ADF test, ktorý nám pomôže určiť, či je časový rad stacionárny. Výsledná p-hodnota je rovná 0,8740, čiže na hladine významnosti 0,05 prijímame nulovú hypotézu, pretože p-hodnota je väčšia ako hladina významnosti, čo znamená, že náš časový rad je nestacionárny a má jednotkový koreň. Keďže sú dáta nestacionárne, je potrebné ich diferencovať. Spravíme diferenciu prvého rádu a znova vykonáme test ADF na súbore údajov. Výsledná p-hodnota bola 0,0142, čiže na hladine významnosti 0,05 zamietame nulovú hypotézu, pretože p-hodnota je menšia ako hladina významnosti, čo znamená, že dáta sú stacionárne, teda dáta nemajú jednotkový koreň. Ak by ešte stále dáta neboli stacionárne museli by sme pokračovať v diferencovaní, kým by sme jednotkový koreň neodstránili a náš rad by bol stacionárny. Vo všeobecnosti by diferenciácia nemalo byť väčšie ako dva. Takto sme pripravili časový rad na ďalší krok, a pri identifikácii modelu budeme teda využívať diferencovaný časový rad.

Na určenie parametrov ARMA modelu nám pomôžu funkcie ACF a PACF. Z grafov odhadnutej ACF a PACF sa dá odhadnúť vhodný model pre náš upravený časový rad. V prípade, že pri výbere určitých parametrov proces nie je stacionárny, jazyk Python vypíše hlásenie, a je potrebná iná voľba parametrov. Ako sme si už hovorili vyššie v bakalárskej práci, ACF popisuje, ako dobre súvisí súčasná hodnota série s jej minulými hodnotami, zatiaľ čo PACF nachádza koreláciu zvyškov s ďalšou hodnotou oneskorenia. Z vykresleného korelogramu ACF vieme odhadnúť parameter MA(q), ktorý by mohol byť 1. Z vykresleného korelogramu PACF vieme odhadnúť parameter AR(p), ktorý by mohol byť 3.

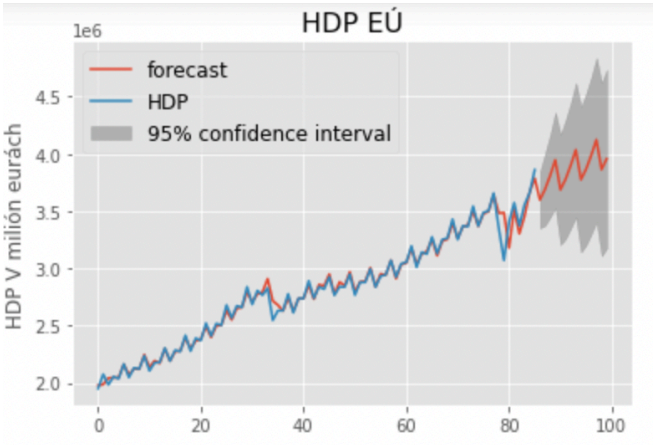


Obrázok č. 5: Vývoj autokorelačnej funkcie a parciálnej autokorelačnej funkcie časového radu hodnôt HDP EÚ.

Podľa doterajších analýz sme odhadli model ARIMA (3,1,1). Teda na prvý pohľad sa môže zdať najvhodnejším modelom pre daný časový rad. Inú voľbu parametrov sledujeme aj na základe informačných kritérií ako je AIC a BIC, ktoré sa snažíme minimalizovať a koeficientu determinácie R^2, ktorý sa snažíme maximalizovať. Model ARIMA (3,1,1) má AIC=2187,812 , BIC=2202,538 a R^2=0,3119. RMSE je 131 006. MAPE je 3,4769. Takto sme pokračovali v ďalších kombináciách parametrov. Najlepším odhadnutým modelom nám vyšiel model ARIMA (4,1,3) s AIC=2199,350, BIC=2201,439 a R^2=0,814, teda v percentách 81,4%. RMSE je 68 453,0431. MAPE je 1,96%, čo je dobré, pretože je menšie ako 5%. Na predikciu HDP Európskej únie použijeme model ARIMA (4,1,3), pretože miery presnosti sú nižšie, informačné kritéria sú nižšie a koeficient determinácie je vyšší.

Vykonáme diagnostický test nezávislosti rezíduí Ljung-Boxovu Q štatistiku. Ktorú sme vypočítali pre lags=20, keďže máme štvrťročné dáta, tak môžeme použiť viac oneskorení. V odhadnutom modeli ARIMA (4,1,3) je výsledná p- hodnota 0,9999 čiže je väčšia ako 0,05, čo znamená, že prijímame nulovú hypotézu nezávislosti rezíduí, takže v časovom rade sa nenachádza autokorelácia. Zatiaľ náš model vyzerá dôveryhodne.

Môžeme začať prognózovať. Najprv však potrebujeme naše dáta rozdeliť na testovacie vzorky a tréningové vzorky. Zoberieme si prvých 95% ako tréningovú vzorku a otestujeme predikciu na zvyšných 5% časové radu čo je testovacia vzorka. Potom spravíme prognózu na najbližšie 3 roky, teda na najbližších 12 období. Ako vidíme na obrázku č. 6 HDP EÚ bude najbližšie 3 roky stúpať.

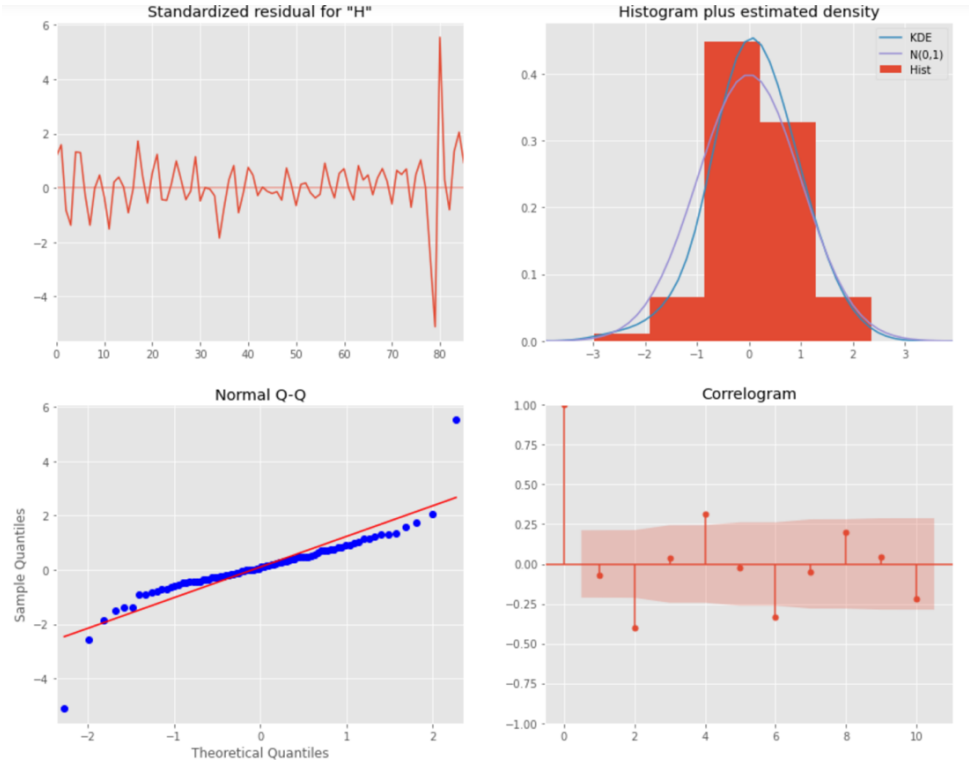


Obrázok č. 6: Prognóza do budúcnosti časového radu hodnôt HDP EÚ pomocou modelu ARIMA(4,1,3).

Po zistení, že náš časový rad je stacionárny, môžeme použiť aj model SARIMA na predpovedanie budúcich hodnôt. Zápis modelu je SARIMA(p, d, q) (P, D, Q) lag. Tieto tri parametre zodpovedajú za sezónnosť, trend a šum v údajoch. Taktiež použijeme informačné kritérium AIC, BIC, miery presnosti ako je RMSE, MAPE a R^2. Ako už vieme, čím nižšia je hodnota AIC, BIC, tým lepšie. R^2 čím je hodnota vyššia , tým lepšie. RMSE, MAPE čím nižšia hodnota, tým lepšie. Po vykonaní viacerých kombinácií modelov naznačuje, že najlepšou kombináciou je SARIMAX(15, 1, 12)x(0, 0, 0) 4 s hodnotou AIC=2233,977 a BIC=2261,634.Vypočítaná hodnota R^2 je 73,9% a RMSE je 81 077,268. MAPE je 1,97%.

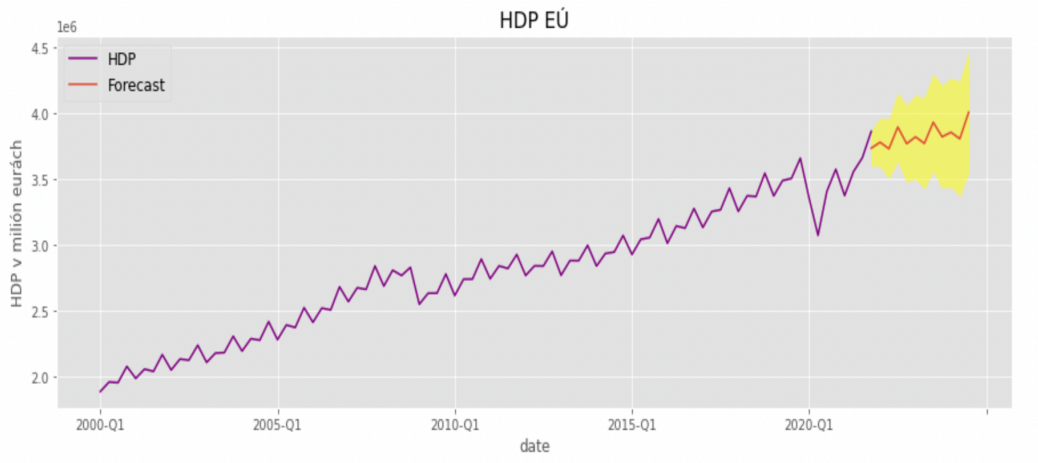
Vykonáme diagnostický test Ljung-Boxovu Q štatistiku, ktorú sme vypočítali pre lags=20. V odhadnutom modeli SARIMAX (15,1,12) x(0, 0, 0) 4 je výsledná p- hodnota 1, čiže je väčšia ako 0,05, čo znamená, že nezamietame nulovú hypotézu nezávislosti rezíduí, takže v časovom rade sa nenachádza autokorelácia. Náš model vyzerá dôveryhodne.

Po prispôsobení modelu údajom skontrolujeme zvyškové grafy, aby sme overili platnosť prispôsobenia modelu. Dobrá prognostická metóda bude mať v rezíduách minimálne informácie. Korelogram vpravo dole naznačuje, že v rezíduách je minimálna autokorelácia, takže ide v skutočnosti o biely šum. Preto sú tieto rezíduá nekorelované a priemer je blízky nule.



Obrázok č. 7: Grafy rezíduí časového radu HDP EÚ.

V ďalšom kroku sa pokúsime predpovedať údaje HDP na nasledujúcich 12 období, teda 3 roky. Nižšie uvedený graf ukazuje dobré prispôsobenie v porovnaní s historickými údajmi. Z grafu na obrázku č. 8 vidíme, že HDP EÚ bude najbližšie 3 roky mierne stúpať.

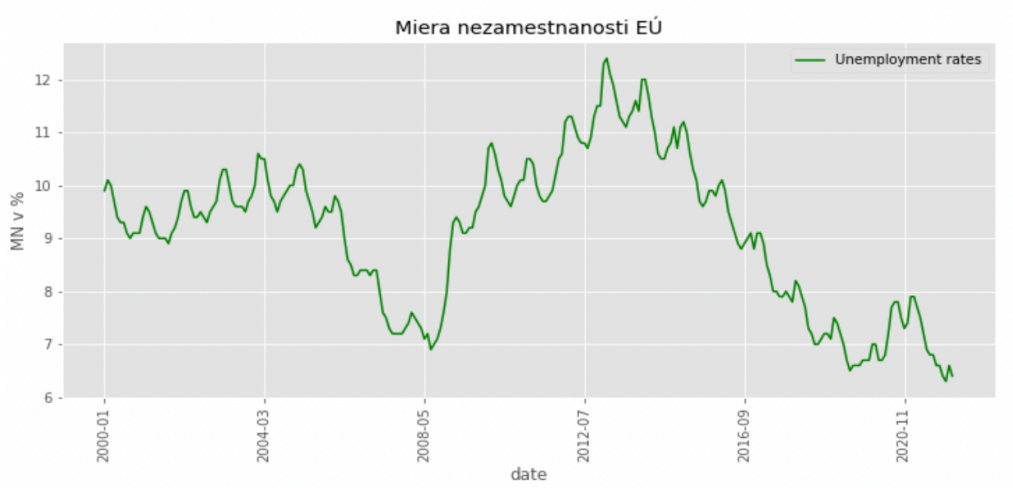


Obrázok č. 8: Prognóza do budúcnosti časového radu hodnôt HDP EÚ pomocou modelu SARIMA(15,1,12)x(0,0,0) 4.

4.2.2 Modelovanie dát miery nezamestnanosti v Európskej únii

Ako ďalšie sa pokúsime namodelovať dáta časového radu vývoja miery nezamestnanosti. Popisy jednotlivých príkazov sú uvedené v Prílohe 2 tejto bakalárskej práce. Definujeme si časový rad s názvom MN s frekvenciou 12, keďže pracujeme s mesačnými dátami. Začiatočným momentom tohto časového radu je január roku 2000. Miera nezamestnanosti je uvedená v percentách.

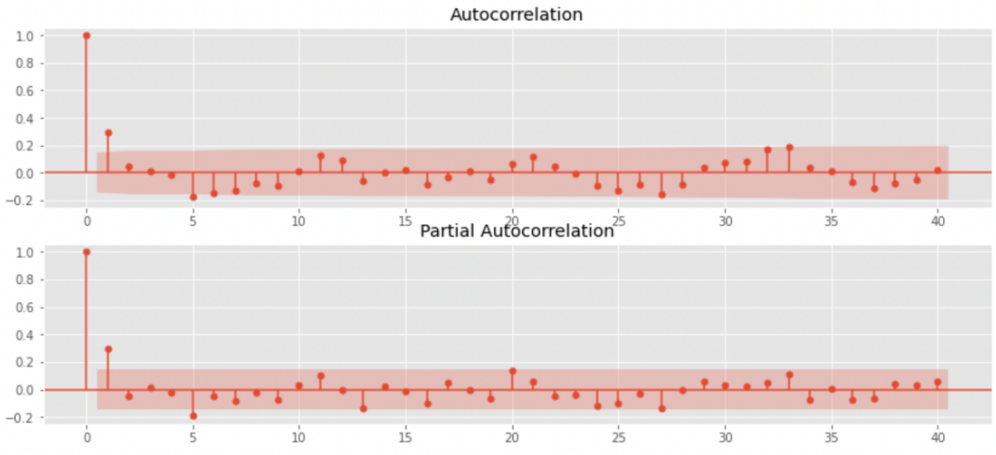
Na nasledovnom obrázku č. 9 je grafické zobrazenie tohto časového radu, a to pravidelný mesačný vývoj inflácie v Európskej únii od januára roku 2000 do februára roku 2022.



Obrázok č. 9: Vývoj hodnôt miery nezamestnanosti v EÚ.

Otestujeme stacionaritu časového radu pomocou ADF testu. Výsledná p-hodnota je rovná 0,5288, čiže na hladine významnosti 0,05 nezamietame nulovú hypotézu, pretože p-hodnota je väčšia ako hladina významnosti, čo znamená, že náš časový rad je nestacionárny a má jednotkový koreň. Musíme spraviť diferenciu prvého rádu a znova vykonáme test ADF na súbore údajov. Výsledná p-hodnota bola 0,1335, čiže naše dáta sú stále nestacionárne. Musíme pokračovať ďalej v diferencovaní, kým neodstránime jednotkový koreň a náš rad bude stacionárny. Spravíme diferenciu druhého rádu a znova vykonáme test ADF. Výsledná p-hodnota je $5{,}2681e^{-10}$, čiže na hladine významnosti 0,05 zamietame nulovú hypotézu, pretože p-hodnota je menšia ako hladina významnosti 0,05, čo znamená, že dáta sú stacionárne, teda dáta nemajú jednotkový koreň. Takto sme pripravili časový rad na ďalší krok, a pri identifikácii modelu budeme teda využívať diferencovaný časový rad druhého radu.

Vykreslíme korelogramy pre znázornenie ACF a PACF, ktoré nám pomôžu určiť parametre ARMA modelu. Z vykresleného korelogramu ACF vieme odhadnúť parameter MA(q), čiže parameter q by mohlo byť 1. Z vykresleného korelogramu PACF vieme odhadnúť parameter AR(p), čiže parameter p by mohol byť 1. Na prvý pohľad môžeme usúdiť, že sa jedná o model ARIMA(1,2,1). No pri tvorbe modelu sme sa riadili hlavne maximalizovaním R^2, minimalizovaním AIC, BIC a minimalizovaním mier presnosti. Pri takomto prístupe sa ako najlepší model javí ARIMA(11,2,11), ktorý má AIC=-240,491, BIC=-154,851 a vypočítaná hodnota R^2 je 0,9937, čiže presnosť modelu je 90,37%. Čo nám hovorí, že náš model by mal byť dôveryhodný. RMSE je 0,1569 a MAPE je 1,62%.



Obrázok č. 10: Vývoj autokorelačnej funkcie a parciálnej autokorelačnej funkcie časového radu hodnôt miery nezamestnanosti v EÚ.

Vykonáme diagnostický test nezávislosti rezíduí Ljung-Boxovu Q štatistiku, ktorú sme vypočítali pre lags=40, keďže máme mesačné dáta tak môžeme použiť viac oneskorení. V odhadnutom modeli ARIMA (11,2,11) je výsledná p- hodnota 0,1041 čiže je väčšia ako 0,05, čo znamená, že nezamietame nulovú hypotézu nezávislosti rezíduí, takže v časovom rade sa nenachádza autokorelácia. Zatiaľ náš model vyzerá dôveryhodne.

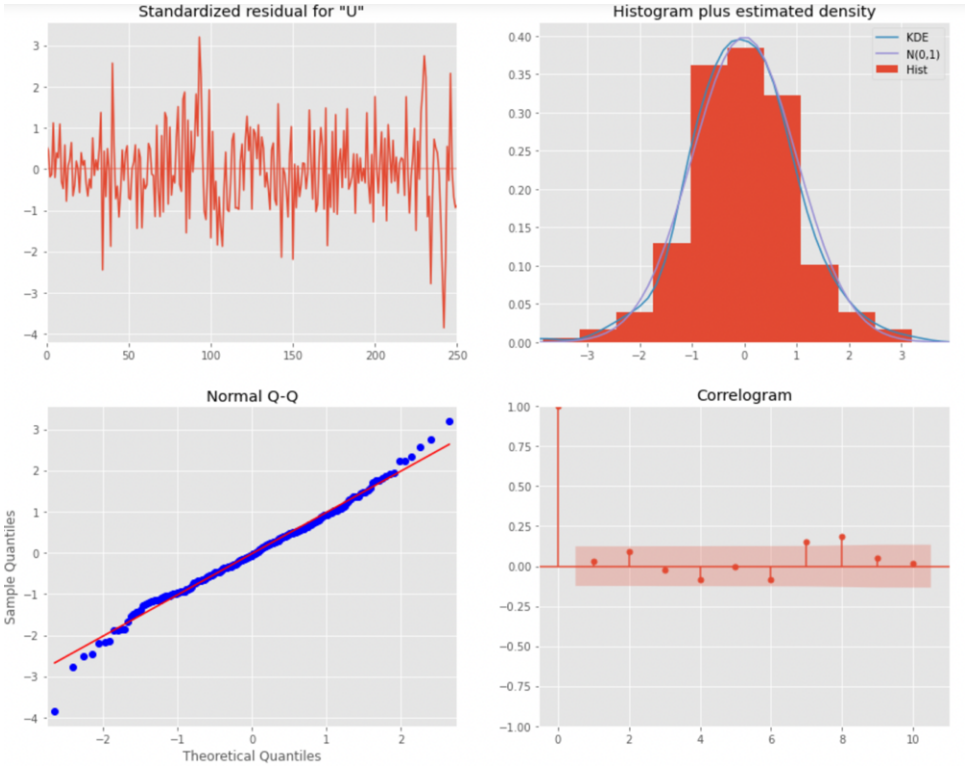
Môžeme začať prognózovať. Rozdelíme si naše údaje na tréningové a testovacie vzorky. Zoberieme si prvých 80% dát ako tréningovú vzorku a otestujeme predikciu na zvyšných 20% časové radu čo je testovacia vzorka. Spravíme prognózu na najbližší rok, teda na najbližších 12 mesiacov. Z grafu na obrázku č. 11 vidíme, že miera nezamestnanosti začne prudko klesať.

Obrázok č. 11: Prognóza do budúcnosti časového radu hodnôt miery nezamestnanosti V EÚ pomocou modelu ARIMA(11,2,11).

Po zistení, že náš časový rad je stacionárny, môžeme použiť aj model SARIMA na predpovedanie budúcich hodnôt. Použijeme informačné kritéria, miery presnosti. Po vykonaní viacerých kombinácií model naznačuje, že najlepšou kombináciou je SARIMAX(5, 2, 5)x(0, 0, 0)12 s hodnotou AIC=-78,236 a BIC=-38,984. Vypočítaná hodnota R^2 je 0,8617, čiže model je presný na 86,17%. RMSE je 0,19 a MAPE je 2,19%.

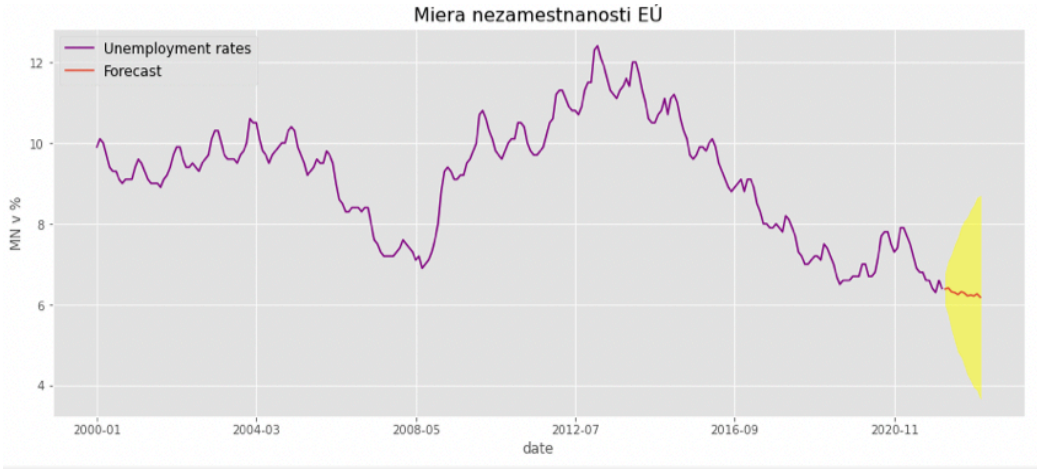
Vykonáme diagnostický test Ljung-Boxovu Q štatistiku, ktorú sme vypočítali pre lags = 40, keďže máme mesačné dáta tak môžeme použiť viac oneskorení. V odhadnutom modeli SARIMAX (5,2,5) x(0, 0, 0, 12) je výsledná p- hodnota 0,5831, čiže je väčšia ako 0,05, čo znamená, že nezamietame nulovú hypotézu nezávislosti rezíduí, takže v časovom rade sa nenachádza autokorelácia. Model vyzerá dôveryhodne.

Po prispôsobení modelu údajom skontrolujeme zvyškové grafy, aby sme overili platnosť prispôsobenia modelu. Dobrá prognostická metóda bude mať v rezíduách minimálne informácie. Korelogram vpravo dole naznačuje, že v rezíduách nie je žiadna autokorelácia, takže ide v skutočnosti o biely šum. Preto sú tieto rezíduá nekorelované a priemer je blízky nule.



Obrázok č. 12: Grafy rezíduí časového radu vývoja miery nezamestnanosti v EÚ.

Spravíme prognózu na najbližší rok, teda na najbližších 12 mesiacov.

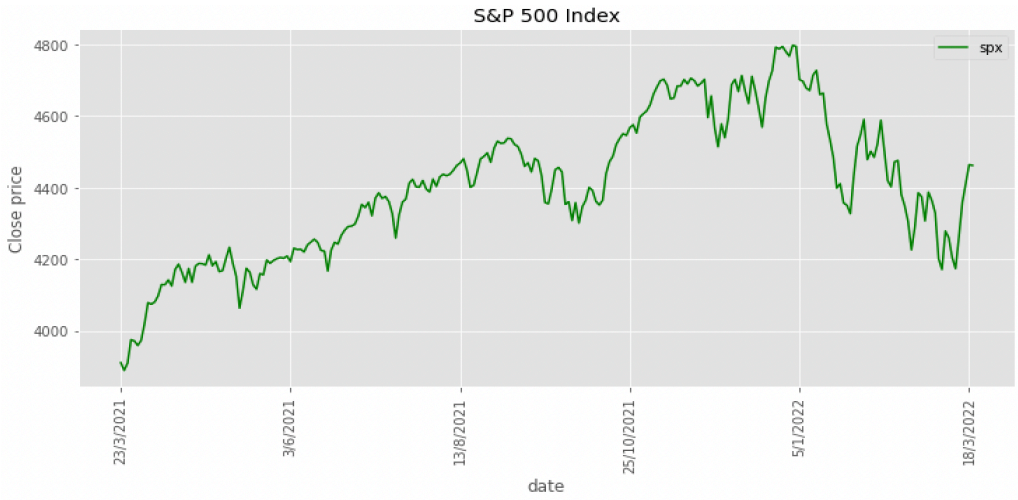


Obrázok č. 13: Prognóza do budúcnosti časového radu hodnôt miery nezamestnanosti v EÚ pomocou modelu SARIMA(5,2,5)(0,0,0)12.

4.2.3 Modelovanie S&P 500 Indexu

Ako posledné sa pokúsime namodelovať dáta časového radu vývoja S&P 500 Indexu. Popisy jednotlivých príkazov sú uvedené v Prílohe 2 tejto bakalárskej práce. Definujeme si časový rad s názvom SP s frekvenciou 365, keďže pracujeme s dennými dátami. Začiatočným momentom tohto časového radu je 22.3.2021.

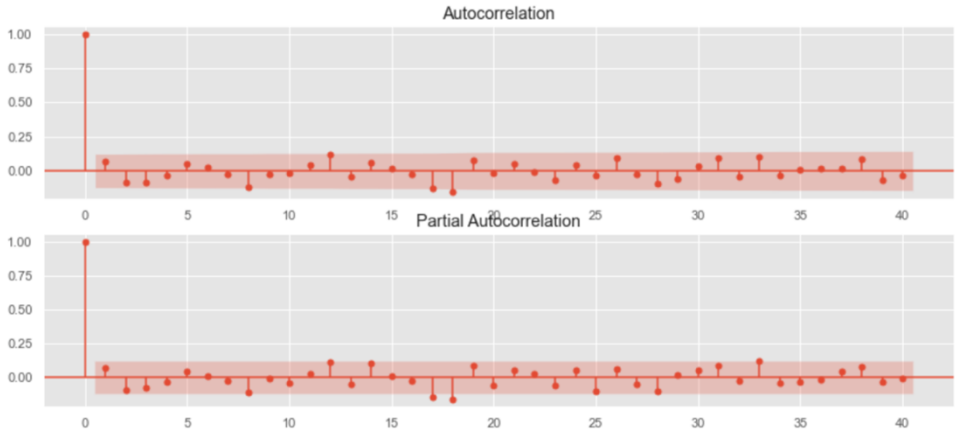
Na nasledovnom obrázku č. 13 môžeme vidieť denný vývoj S&P 500 Indexu v dennom intervale od 22.3.2021 do 21.3.2022.



Obrázok č. 14: Vývoj časového radu hodnôt S&P 500 Indexu.

Otestujeme stacionaritu časového radu pomocou ADF testu. Výsledná p-hodnota je rovná 0,1369, čiže na hladine významnosti 0,05 nezamietame nulovú hypotézu, pretože p-hodnota je väčšia ako hladina významnosti, čo znamená, že náš časový rad je nestacionárny a má jednotkový koreň, čo sme aj očakávali, pretože sa jedná o denné finančné dáta, pre ktoré je charakteristická nestacionarita. Spravíme diferenciu prvého rádu a znova vykonáme test ADF na súbore údajov. Výsledná p-hodnota bola $2{,}5318e^{-27}$, čiže na hladine významnosti 0,05 zamietame nulovú hypotézu, pretože p-hodnota je menšia ako hladina významnosti, čo znamená, že dáta sú stacionárne, teda dáta nemajú jednotkový koreň. Takto sme pripravili časový rad na ďalší krok a pri identifikácii modelu budeme teda využívať diferencovaný časový rad prvého rádu.

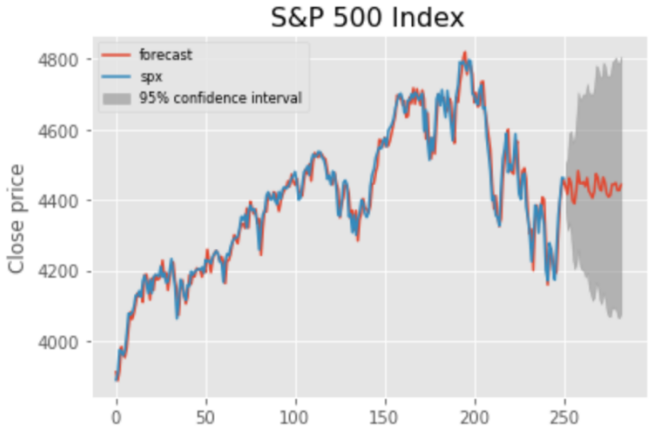
Vykreslíme korelegramy pre ACF a PACF, ktoré nám pomôžu určiť parametre ARMA modelu. Z vykresleného korelogramu ACF môžeme odhadnúť, že náš parameter q by mohlo byť 8 a z korelogramu PACF môžeme odhadnúť, že náš parameter p by mohol byť taktiež 8. Na prvý pohľad môžeme usúdiť, že sa jedná o model ARIMA(8,1,8), no pri tvorbe najvhodnejšieho modelu sme sa znova riadili informačnými kritériami a mierami presnosti. Pri takomto postupe sa ako najlepší model javí ARIMA(17,1,17), ktorý má AIC=2588,497, BIC=2715,413 a R^2=0,8946, teda model je na 89,46% presný. RMSE je 47,0281 a MAPE je 0,8477%. A model ARIMA(8,1,8) má AIC=2586,606, BIC=2650,064 a hodnota R^2=0,8539, takže náš model má presnosť 85,39%, čo nám hovorí že náš model by mal byť dôveryhodný. RMSE je 55,3741 a MAPE je 0,99%. Vyzerá, že model ARIMA(17,1,17) je lepší, síce AIC, BIC má o kúsok vyššie, ale nie výrazne. Miery presnosti má lepšie ako model ARIMA(8,1,8). Môžeme ale vidieť, že teraz pri týchto modeloch nie je jednoznačné, ktorý model je lepší.



Obrázok č. 15: Vývoj autokorelačnej funkcie a parciálnej autokorelačnej funkcie časového radu hodnôt S&P 500 Indexu.

Vykonáme diagnostický test nezávislosti rezíduí Ljung-Boxovu Q štatistiku. Vypočítali sme ju pre lags=40, keďže máme denné dáta tak môžeme použiť viac oneskorení. V odhadnutom modeli ARIMA(8,1,8) je výsledná p-hodnota 0,9723, čiže je väčšia ako 0,05, čo znamená, že nezamietame nulovú hypotézu nezávislosti rezíduí. Zatiaľ model vyzerá dôveryhodne.

Môžeme začať prognózovať. Rozdelíme si naše údaje na tréningové a testovacie vzorky. Zoberieme si prvých 80% ako tréningovú vzorku, a otestujeme predikciu na zvyšných 20% časové radu, čo je testovacia vzorka. Spravíme prognózu na najbližších 30 dní. Na obrázku č. 16 vidíme prognózu S&P 500 Indexu, uzatváracie ceny akcií sa zvýšia. Ale graf nevyzerá úplne dôveryhodne, pretože ARIMA modely nie sú úplne vhodné na denné dáta.

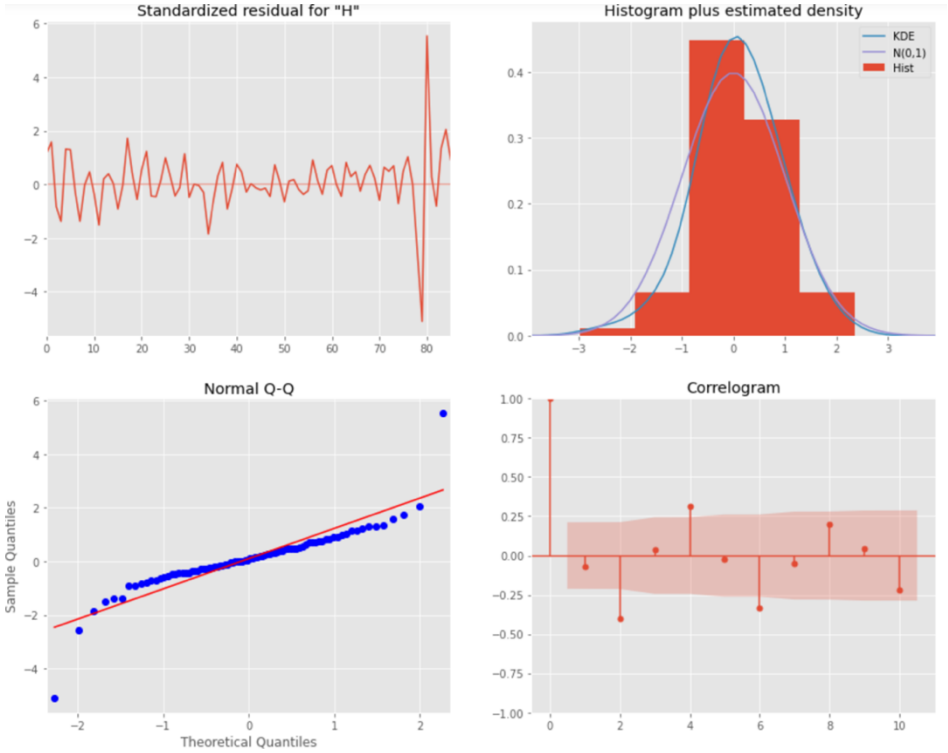


Obrázok č. 16: Prognóza do budúcnosti časového radu hodnôt S&P 500 Indexu pomcou modelu ARIMA(17,1,17).

Po zistení, že náš časový rad je stacionárny, môžeme použiť aj model SARIMA na predpovedanie budúcich hodnôt. Pri odhade nášho modelu postupujeme úplne rovnako, ako pri vyššie uvedených dátach. Po vykonaní viacerých kombinácií model naznačuje, že najlepšou kombináciou je SARIMAX(17, 1, 17)x(0, 1, 0) 7 s hodnotou AIC=2567,59 a BIC = 2689,99. R^2 je 86,18%. RMSE je 53,8624 a MAPE je 1,001%

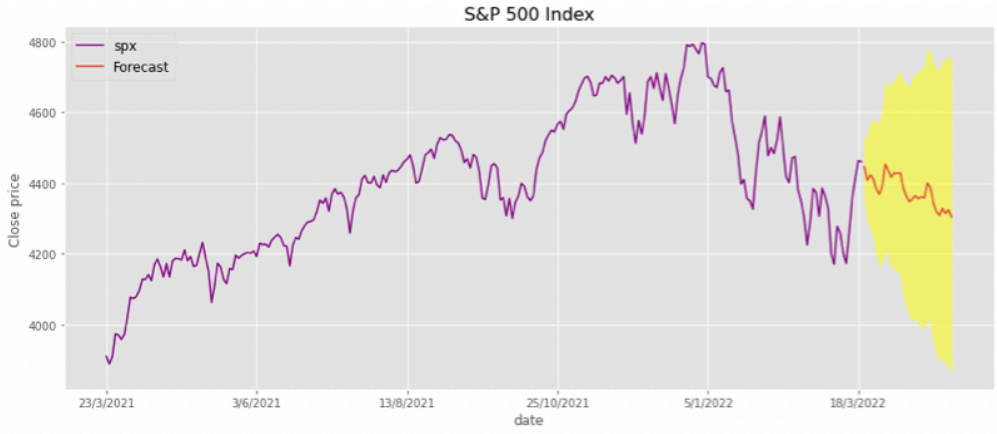
Vykonáme diagnostické test Ljung-Boxovu Q štatistiku, ktorú sme vypočítali pre lags=40. V odhadnutom modeli SARIMAX(17,1,17)x(0,1,0)7 je výsledná p-hodnota 0,4998, čiže je väčšia, ako 0,05, čo znamená, že nezamietame nulovú hypotézu nezávislosti rezíduí, takže v časovom rade sa nenachádza autokorelácia. Model vyzerá dôveryhodne.

Po prispôsobení modelu údajom skontrolujeme zvyškové grafy, aby sme overili platnosť prispôsobenia modelu. Korelogram vpravo dole naznačuje, že v rezíduách vyzerá jemná autokorelácia.



Obrázok č. 17: Grafy rezíduí časového radu S&P 500 Indexu.

Spravíme prognózu pomocou modelu SARIMA(17,1,17)x(0,1,0) 7 na najbližších 30 dní. Na obrázku č. 19 sa javí, že uzatváracie ceny S&P 500 Indexu začnú klesať. Môžeme spozorovať, že naše prognózy ARIMA a SARIMA sa nezhodujú.



Obrázok č. 18: Prognóza do budúcnosti časového radu hodnôt S&P 500 Indexu pomocou modelu SARIMA(17,1,17)x(0,1,0) 7.

4.3 Vyhodnotenie výsledkov modelovania časových radov

Hlavným kritériom hodnotenia modelu sú informačné kritéria a predikčné vlastnosti. Informačné kritéria modelu sú AIC a BIC . Lepší je model s menším AIC, BIC. Kritériom hodnotenia predikčných vlastností modelu je stredná štvorcová chyba odhadu RMSE, koeficient determinácie R^2 a priemerná absolútna percentuálna chyba MAPE. Ďalším kritériom hodnotenia modelu je test nezávislosti rezíduí Ljung-Boxov test, ktorý musí mať väčšiu p-hodnotu, ako hladina významnosti 0,05, pretože potrebujeme prijať nulovú hypotézu. V modeli nemôže byť prítomná autokorelácia. Hodnoty týchto ukazovateľov sme zapísali do tabuliek pre vybrané kombinácie modelov.

V tabuľke je najlepší model označený zelenou farbou. Červenou farbou sú označené hodnoty, ktoré boli zlé v daných modeloch, a kvôli ktorým sme museli daný model zamietnuť.

4.3.1 Výsledky modelovania HDP EÚ

Tabuľka č. 6: Výsledky modelov HDP EÚ na základe jednotlivých ukazovateľov.

Model/kritéria	AIC	BIC	R2	RMSE	MAPE	Ljungbox test
ARIMA (0,1,1)	2 247,502	2 254,865	-0,572	198 992,456	4,916	4,5139e-29
ARIMA (1,1,0)	2 240,003	2 247,396	-0,092	165 830,137	4,113	2,5386e-25
ARIMA (0,1,0)	2 274,112	2 279,021	-0,07	164 158,023	4,407	2,257e-90
ARIMA (4,1,3)	2 199,350	2 201,439	0,814	68 453,043	1,69	0,999
SARIMA (4,1,3)x(0,0,0)4	2 243,852	2 254,486	0,427	120 131,971	3,101	0,999
SARIMA (1,1,0)x(0,0,0)4	2 272,107	2 277,016	-0,304	181 239,529	4,944	0,023
SARIMA (1,1,0)x(0,0,0)4	2 271,856	2 276,764	-0,273	179 038,323	4,885	0,008
SARIMA (15,1,12)x(0,0,0)4	2 233,977	2 302,698	0,739	81 077,267	1,97	1

Zdroj dát: vlastné spracovanie

4.3.2 Výsledky modelovania miery nezamestnanosti EÚ

Tabuľka č. 7: Výsledky modelov miery nezamestnanosti EÚ na základe jednotlivých ukazovateľov.

Model/kritéria	AIC	BIC	R2	RMSE	MAPE	Ljungbox test
ARIMA (1,2,1)	-64,432	-50,129	0,844	0,202	2,234	4.301850e-25
ARIMA (1,2,0)	-2,174	8,553	0,779	0,240	2,704	8.056197e-26
ARIMA (0,2,1)	-12,464	-1,736	-0,797	0,230	2,674	5.345212e-35
ARIMA (11,2,11)	-243,835	-158,012	0,906	0,156	1,62	0,105
SARIMA (11,2,11)x(0,0,0)12	-238,545	-156,298	0,898	0,163	1,747	0,654
SARIMA (1,2,1)x(0,0,0)12	-65,989	-55,262	0,844	0,202	2,259	0,536
SARIMA (1,2,0)x(0,0,0)12	-4,169	2,983	0,779	0,240	2,705	0,429
SARIMA (5,2,5)x(0,0,0)12	-84,122	-44,786	0,862	0,190	2,187	0,583

Zdroj dát: vlastné spracovanie

4.3.3 Výsledky modelovania S&P 500 Indexu

Tabuľka č. 8: Výsledky modelov miery S&P 500 Indexu na základe jednotlivých ukazovateľov.

Model/kritéria	AIC	BIC	R2	RMSE	MAPE	Ljungbox test
ARIMA (0,1,0)	2 580,602	2 587,653	0,822	61,067	1,159	0,160
ARIMA (1,1,0)	2 581,467	2 592,043	0,826	60,396	1,143	0,168
ARIMA (8,1,8)	2 586,606	2 650,064	0,854	55,374	0,993	0,972
ARIMA (17,1,17)	2 588,497	2 715,413	0,894	47,0281	0,848	0,999
SARIMA (1,1,0)x(0,0,0)7	2 580,101	2 587,152	0,827	60,199	1,319	0,494
SARIMA (0,1,0)(0,1,0)7	2 679,316	2 682,813	0,658	84,758	1,507	0,340
SARIMA (17,1,17)(0,1,0)7	2 567,593	2 689,994	0,862	53,862	1,001	0,500
SARIMA (8,1,8)x(0,1,0)7	2 555,898	2 615,349	0,834	58,944	1,101	1,000

Zdroj dát: vlastné spracovanie

ZÁVER

Cieľom tejto bakalárskej práce bolo priblížiť čitateľovi základné charakteristiky časových radov, a ich využitie pri tvorbe lineárnych modelov pomocou jazyka Python. Zamerali sme sa na jednorovnicové ARIMA modely. Analýzu časových radov sme pozorovali na ekonomických dátach s rôznou frekvenciou, ako na štvrťročných dátach HDP Európskej únie, mesačných dátach miery nezamestnanosti Európskej únie a denných dátach S&P 500 Indexu. Na základe identifikácie modelu, overení a verifikácii modelu sme určili najvhodnejší model, ktorým sme mohli robiť prognózy do budúcnosti s najmenšou chybou predikcie. Jazyk Python bol pri tom veľmi účeľný.

Zistili sme, že modely sú vhodnejšie a presnejšie pre štvrťročné a mesačné dáta. Pri denných dátach boli menšie odchýlky v závere prognózy pomocou ARIMA a SARIMA modelu.

Aj tak môžeme zhrnúť, že naše modely ARIMA a SARIMA prinášajú pozoruhodné výsledky. Tieto modely ponúkajú dobrú presnosť predpovedí, a sú relatívne rýchle v porovnaní s inými alternatívami, pokiaľ ide o čas a zložitosť modelovania. Rozhodne by sme mali pristupovať ku všetkým prognózam skepticky, pretože budúcnosť je vo svojej podstate neistá, a žiadne množstvo výpočtovej techniky, alebo údajov túto skutočnosť nikdy nezmení. Ale pozorovaním a analýzou trendov histórie môžeme aspoň malú časť tejto neistoty odhaliť.

Zoznam použitej literatúry

Knižné zdroje:

[1] ANDERSON, O. Time series analysis and forecasting: the Box-Jenkins approach. London: Butterworth, 1976. ISBN 0408706759.

[2] BOX, G.E.P. – JENKINS, G.M. – REINSEL, G.C. – LJUNG, G.M. *Time Series Analysis. Forecasting and Control, Fifth Edition.* 2016. ISBN 978-1-118-67502-1.

[3] BROWNLEE, J. *Introduction to Time Series Forecasting with Python.* 2017. (eBook)

[4] DICKEY, D. A. – FULLER, W. A.: Distributions of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of American Statistical Association* 74, 1979.

[5] FISCHER, B. Decompositions of Time Series. Comparing Dierent Methods in Theory and Practice. Luxembourg: Eurostat, 1995.

[6] HUDEC, O. a kol. *Štatistické metódy v ekonomických vedách*. Košice: Elfa, 2007. ISBN 978-80-8086-059-2.

[7] CHATFIELD, C. *The Analysis of Time Series: An Introduction, Sixth Edition*. Taylor&Francis, 2003. ISBN 1584883170.

[8] CHOLLET, F. *Deep learning v jazyku Python*. Grada. 2019. ISBN 9788024731001.

[9] KIRCHGSSNER, G. – WOLTERS, J. *Intorduction to Modern Time Series Analysis.* 2007. ISBN 978-3-540-73290-7.

[10] LUKÁČIKOVÁ, A. – LUKÁČIK, M.: *Ekonometrické modelovanie s aplikáciami*. Bratislava: EKONÓM, 2008. ISBN 978-80-225-2614-2.

[11] VERBEEK, M. A guide to modern econometrics. New Jearsey: Wiley, 2008. ISBN 0-470-51769-7.

Internetové zdroje:

[12] Claeskens G. Statistical model choice. Faculty of Economics and Business Ku Leuven. Dostupné na internete: <https://feb.kuleuven.be/public/u0043181/papers/KBI_1521.pdf>.

[13] Stránka Python [online]. Dostupné na internete: < https://www.python.org >.

[14] Stránka Invsetopedia, The S&P 500 Index: Standard & Poor's 500 Index [online]. Dostupné na internete: <https://www.investopedia.com/terms/s/sp500.asp >.

[15] Stránka databázy štatistického úradu Európskej komisie Eurostat [online]. Dostupné na internete: <https://ec.europa.eu/eurostat/databrowser/view/namq_10_gdp/default/table?lang=en >.

[16] Stránka štatistického úradu Európskej komisie Eurostat [online]. Dostupné na internete: <https://ec.europa.eu/eurostat/databrowser/view/ei_lmhr_m/default/table?lang=en>.

[17] Stránka štatistického úradu Európskej komisie Eurostat– Čo je to HDP? [online]. ISSN 2443-8219. Dostupné na internete: <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:GDP_What_is_gross_domestic_product_(GDP)%3F/sk#C4.8Co_spad.C3.A1_pod_HDP.3F>.

[18] Stránka štatistického úradu Európskej komisie Eurostat – Trh práce – nezamestnanosť[online]. ISSN 2443-8219. Dostupné na internete: <https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Beginners:Labour_market_-_unemployment/sk>.

[19] Stránka marketwatch [online]. Dostupné na internete: <https:_/www.marketwatch.com/investing/index/spx/download-data>.