SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

1/2024
ročník/volume 34

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 7 – 26

Dátum vydania/Publication date: 15. január 2024/January 15, 2024

Peter KNÍŽAT
Štatistický úrad Slovenskej republiky, Ekonomická univerzita v Bratislave

Dagmar CEĽUCHOVÁ BOŠANSKÁ, Martin JANÍK, Filip NGUYEN
Alistiq, s. r. o.

NOVÉ DÁTOVÉ ZDROJE V ŠTATISTIKE: VPLYV INTENZITY NÁKLADNEJ DOPRAVY NA MAKROEKONOMICKÉ UKAZOVATELE

NEW DATA SOURCES IN STATISTICS: THE EFFECT OF FREIGHT INTENSITY ON MACROECONOMIC INDICATORS

ABSTRAKT

Cieľom tohto článku je overiť možnosti využitia modelov SARIMA a SARIMAX na odhad vývoja HDP Slovenska. V modeli SARIMAX boli údaje o najazdených kilometroch nákladnej dopravy využité ako dodatočná exogénna premenná. Po dekompozícii údajov slovenského elektronického mýtneho systému na jednotlivé zložky bola identifikovaná sezónna zložka a zložka rezíduí, ktoré nemožno vysvetliť ani trendom, ani sezónnymi zložkami, čo naznačuje, že časový rad údajov nie je stacionárny. Pri prispôsobovaní modelov SARIMA/SARIMAX sa na kontrolu stacionarity časových radov použil Augmentovaný Dickeyho-Fullerov test (ADF) a na transformáciu časových radov sa použila metóda kĺzavého priemeru. Primeranosť prispôsobených modelov bola potvrdená pozorovaním výsledku Ljung-Box testu. Navyše bol použitý softvér JDemetra+ na sezónnu analýzu časových radov. Podľa testu stacionárnosti časového radu použitím informačného kritéria Akaike (AIC) bol na prognózu HDP Slovenska najvhodnejší prispôsobený model SARIMAX. Výsledky odmocniny zo strednej kvadratickej percentuálnej chyby (RMSPE) a priemernej absolútnej percentuálnej chyby (MAPE) naznačujú nadradenosť modelu SARIMAX, ktorý bral do úvahy sezónne vzorce a exogénne faktory. Model SARIMAX prekonal SARIMA, pričom predpovedal hodnoty v rámci 95 % intervalu spoľahlivosti s hodnotou RMSPE 8,9 %, zatiaľ čo SARIMA mala RMSPE 17,4 %. Závery tohto príspevku nasvedčujú, že nekonvenčné zdroje údajov môžu mať vysoký potenciál využitia na odhad ekonomických indikátorov, ktoré môžu poskytnúť komplexnejší a aktuálnejší pohľad na hospodársku situáciu.

ABSTRACT

The aim of this article is to verify the possibilities of using the SARIMA and SARIMAX models to estimate the development of Slovakia's GDP. In the SARIMAX model, the data on kilometers travelled in freight transport were used as an additional exogenous variable. After decomposing the Slovak electronic toll system data into its component parts, seasonal component and enough residuals component were identified, indicating that the data time series is not stationary. In fitting the SARIMA/SARIMAX models, the Augmented Dickey-Fuller (ADF) test was used to check the time series stationarity and the Moving Average Method was used for the time series transformation. The adequacy of the fitted models was confirmed by observing the Ljung-Box test result. Moreover, the JDemetra+ software was utilized for seasonal analysis of time series. According to the stationarity test of the time series using the Akaike information criterion (AIC), the fitted SARIMAX model was the most suitable to forecast the GDP of Slovakia. The results from the Root Mean Squared Percentage Error (RMSPE) and the Mean Absolute Percentage Error (MAPE) indicate the

superiority of the SARIMAX model, which took into account the seasonality patterns and exogenous factors. The SARIMAX model outperformed the SARIMA, predicting values within the 95 % confidence interval, with a RMSPE value of 8.9 % while the SARIMA had a RMSPE of 17.4 %. The conclusions of this article indicate that non-conventional data sources can have a high potential of use for the estimation of economic indicators, that can provide a more comprehensive and up-to-date view of the economic situation.

KĽÚČOVÉ SLOVÁ
rýchle odhady, hrubý domáci produkt, nákladná doprava, telemetria, SARIMAX

KEY WORDS
flash estimates, gross domestic product, freight transportation, telemetry, SARIMAX

1. ÚVOD

V januári 2010 Slovenská republika uviedla do prevádzky mýtny systém, ktorý sa po rozšírení na všetky triedy ciest stal najdlhšou sieťou spoplatnených ciest nižšej triedy v Európskej únii. Prostredníctvom satelitnej technológie výberu mýta bolo pokrytých 17 600 km vymedzených úsekov ciest Slovenskej republiky [17]. Vďaka tomu má Slovensko unikátnu pozíciu na využívanie údajov vzniknutých počas prevádzky tohto systému na vytváranie ekonomických štatistík, keďže zachytávajú zásadnú časť cestnej prepravy tovarov.

Tieto údaje vznikajú automatizovaným spôsobom ako inherentná funkcia palubných jednotiek inštalovaných do nákladných vozidiel, ktoré využívajú spoplatnené úseky slovenskej cestnej siete patriace do elektronického mýtneho systému. Palubné jednotky zaznamenávajú aktuálne geografické údaje o vymedzených úsekoch ciest podliehajúcich mýtnej povinnosti (tzv. geo model sleduje polohu vozidla pomocou GPS, ktorá je porovnaná s uloženými údajmi v geo modeli. Ak algoritmus palubnej jednotky zistí, že vozidlo použilo vymedzený úsek podliehajúci úhrade mýta, vytvorí sa v súlade s platnou legislatívou príslušný mýtny záznam o tejto skutočnosti (tzv. mýtna udalosť). Mýtna transakcia je elektronický dátový záznam, ktorý vznikne na základe vyhodnotenia a spracovania jednej alebo kombinácie viacerých mýtnych udalostí. Mýtna transakcia obsahuje dátum a čas mýtnej udalosti, identifikáciu mýtneho úseku, identifikáciu vozidla, výšku mýta, platobný režim a ďalšie údaje. Mýtny úsek je definovaný ako súvislá časť vymedzeného úseku ciest, na ktorej sa vykonáva detekcia mýtnej povinnosti prechádzajúcich vozidiel. Mýtny úsek je spravidla vymedzený od hranice križovatky, ktorá tvorí začiatok vymedzeného úseku ciest, po hranicu križovatky, ktorá tvorí koniec vymedzeného úseku ciest, a naopak. To znamená, že každému vymedzenému úseku ciest prislúchajú spravidla dva mýtne úseky – jeden na smer tam a druhý na smer späť. Každý mýtny úsek je označený jednoznačným identifikátorom, začiatkom úseku, koncom úseku a spoplatnenou dĺžkou úseku [15]. Spoplatnená dĺžka úseku je číselný údaj v kilometroch stanovený vo vyhláške Ministerstva dopravy, výstavby a regionálneho rozvoja Slovenskej republiky č. 228/2020 Z. z., ktorou sa vymedzujú úseky diaľnic, rýchlostných ciest, ciest I. triedy a ciest II. triedy s elektronickým výberom mýta v platnom znení pre príslušný vymedzený úsek ciest a uvádza sa s presnosťou na tri desatinné miesta bez ohľadu na skutočnú fyzickú dĺžku mýtneho úseku. Na Slovensku majú povinnosť platiť mýto všetky motorové vozidlá s najväčšou technicky prípustnou celkovou hmotnosťou nad 3,5 tony alebo jazdnými súpravami s najväčšou technicky prípustnou celkovou

hmotnosťou nad 3,5 tony uvedenými v § 4 ods. 2 písm. b) a c) zákona č. 106/2018 Z. z. o prevádzke vozidiel v cestnej premávke (vozidlá kategórie M a N) okrem motorových vozidiel kategórie M1 a okrem jazdných súprav tvorených motorovým vozidlom kategórie M1 a N1.

Hlavným cieľom tohto článku je analýza nového dátového zdroja v štatistike, ktorý poskytuje veľké množstvo dát a pomocou ktorého je možné získať dáta vo veľmi krátkom časovom horizonte daného analyzovaného obdobia. Tento dátový zdroj by mohol mať v budúcnosti potenciálne využitie na rýchle odhady makroekonomických ukazovateľov. Na empirickú analýzu vplyvu týchto dát z nového zdroja bol vybraný hrubý domáci produkt (HDP) SR. Z ekonomického hľadiska môžeme predpokladať, že diaľničná doprava zachytáva len určitú časť z HDP v SR.

2. POUŽITÉ ÚDAJE

Vstupnými údajmi sú údaje o pohybe nákladných vozidiel v mýtnom systéme za roky 2018 až 2022. Dáta sa zbierajú v reálnom čase, ako sa nákladné vozidlá s hmotnosťou nad 3,5 tony presúvajú po platených úsekoch diaľnic, rýchlostných ciest a ciest I. triedy. Podstatná väčšina (expertný odhad NDS je viac ako 99 %) údajov je dostupná do 10. dňa mesiaca nasledujúceho po mesiaci, v ktorom sa zbierali. Súbor údajov obsahuje:

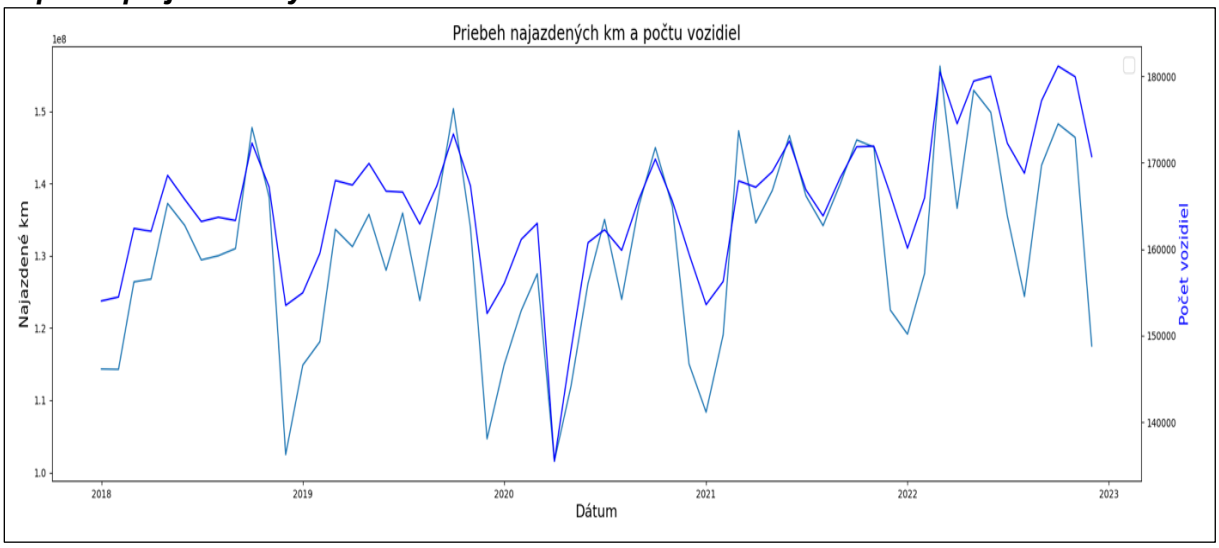
- unikátny identifikátor vozidla, aby bolo možné správne spočítať jeho prejdené kilometre a zrealizované jazdy,
- dátum a čas záznamu polohy vozidla, aby sa kilometre spočítali v správnom časovom okne,
- kategóriu vozidla, aby sa brali do úvahy len vozidlá relevantné pre nákladnú dopravu, a nie napríklad autobusy,
- identifikátor vymedzeného úseku cesty, ktorý určuje polohu vozidla pre daný záznam a spolu so smerom jazdy a ďalšími záznamami o prejdených úsekoch dovoľuje vyskladať celkovú jazdu vozidla od prvého až po posledný záznam v mýtnom systéme,
- prejdená vzdialenosť v kilometroch, ktorú vozidlo prešlo od predchádzajúceho záznamu.

Limitom pre algoritmy bol dátový súbor s odhadovanou premennou HDP v bežných cenách v mil. eur, ktorý vzhľadom na dostupnosť údajov len štvrťročne neobsahoval dostatok dátových bodov v sledovanom období 2018 až 2022. Pri makroekonomických analýzach časových radov je v praxi bežné použitie oveľa dlhšieho časového okna.

3. DÁTOVÁ ANALÝZA A PRIESKUM VSTUPNÝCH ÚDAJOV

Údaje z elektronického mýtneho systému predstavujú súbor dátových bodov zoradených v čase, teda ich možno považovať za časové rady. Obrázok č. 1 ukazuje časový rad najazdených kilometrov a počtu unikátnych vozidiel v nákladnej doprave za jednotlivé mesiace rokov 2018 až 2022. Údaje sú indexované podľa času a agregované na konci každého mesiaca každého roka. Počet najazdených kilometrov a počet vozidiel v priebehu každého roka stúpa a klesá a tento fenomén sa opakuje každý rok, čo predstavuje sezónnosť. Obrázok č. 2 obdobne znázorňuje počet najazdených kilometrov a štvrťročné HDP v bežných cenách v mil. eur, z ktorého však nie je zrejmá podobnosť týchto časových radov, na rozdiel od obrázka č. 1.

Obrázok č. 1: Priebeh počtu najazdených kilometrov a počtu vozidiel v nákladnej doprave po jednotlivých mesiacoch rokov 2018 až 2022



Zdroj: vlastné spracovanie autorov

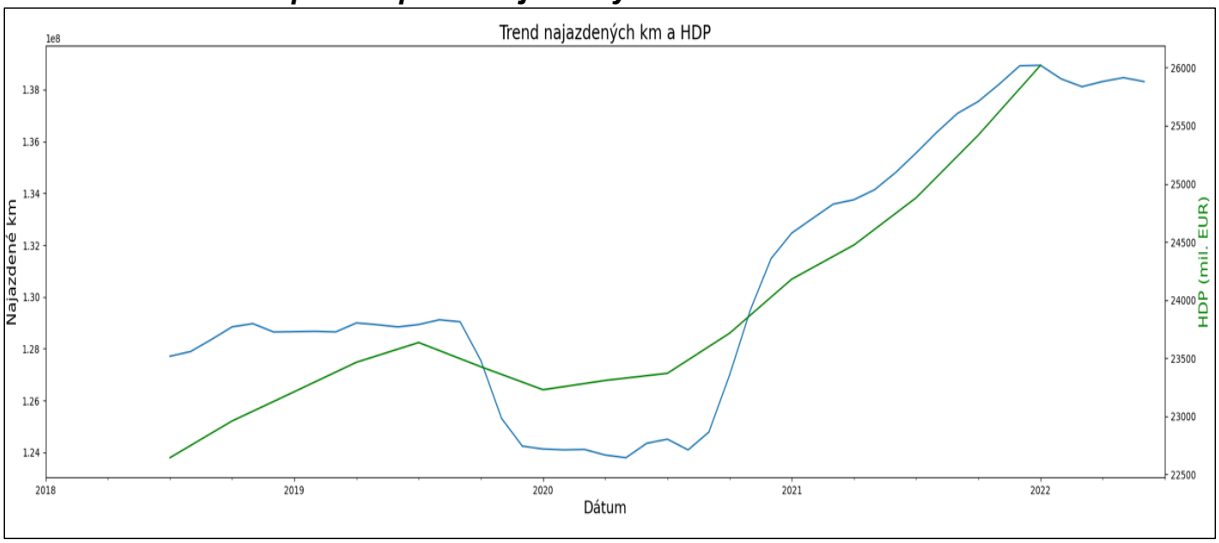
Obrázok č. 2: Priebeh počtu najazdených kilometrov v nákladnej doprave a HDP po jednotlivých mesiacoch a štvrťrokoch rokov 2018 až 2022

Zdroj: vlastné spracovanie autorov podľa údajov [21]

Časové rady možno lepšie pochopiť, keď ich rozložíme na tri zložky: trendovú, sezónnu a reziduálnu. Vizualizácia týchto zložiek časového radu sa nazýva dekompozícia a je definovaná ako štatistická úloha, ktorá rozdeľuje časový rad na jeho jednotlivé zložky [22]. Vizualizácia každej zložky dovoľuje identifikovať trend a sezónny vzorec v údajoch, čo sa nie vždy dá jednoducho rozpoznať len pri pohľade na súbor údajov (obrázok č. 1. a obrázok č. 2).

Dekompozícia mesačného počtu najazdených kilometrov a štvrťročných údajov HDP v bežných cenách v mil. eur, je znázornená na nasledujúcich obrázkoch. Pozorované údaje boli rozdelené na trend (obrázok č. 3), sezónnu zložku (obrázok č. 4) a rezíduá (obrázok č. 5). Výsledkom kombinácie týchto troch zložiek je opäť priebeh počtu najazdených kilometrov a HDP v čase (obrázok č. 2).

Obrázok č. 3: Dekompozícia počtu najazdených kilometrov a HDP – trend



Zdroj: vlastné spracovanie autorov podľa údajov [21]

Obrázok č. 3 znázorňuje celkový pozitívny trend a pokles, ktorý je pravdepodobne spôsobený pandémiou ochorenia COVID-19.

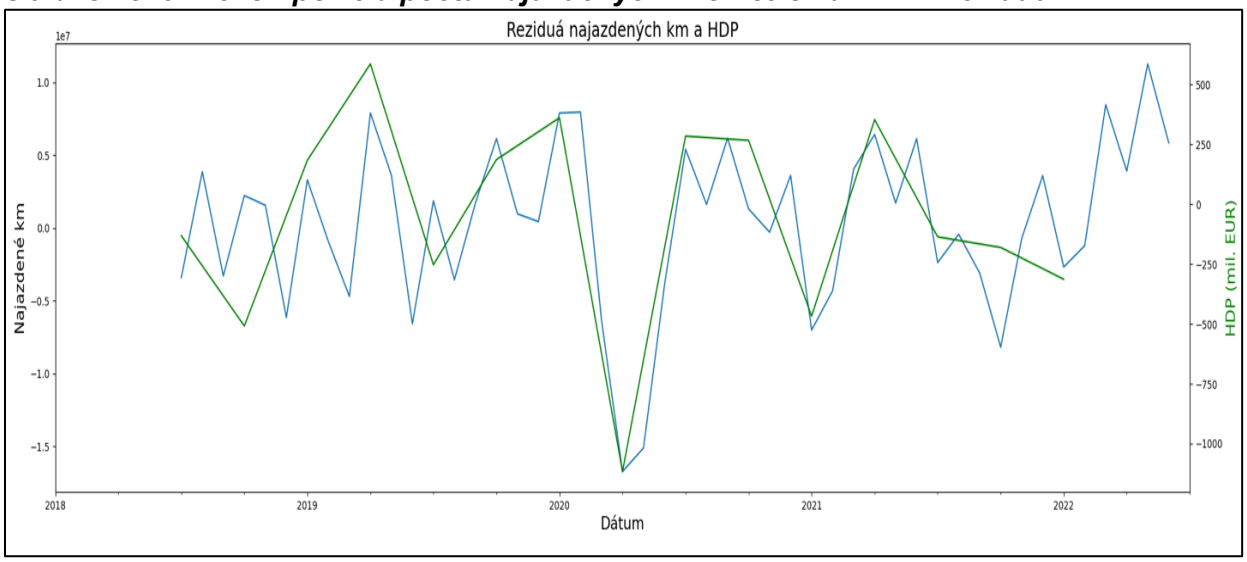
Obrázok č. 4: Dekompozícia počtu najazdených kilometrov a HDP – sezónna zložka

Zdroj: vlastné spracovanie autorov podľa údajov [21]

Sezónna zložka (obrázok č. 4) zachytáva sezónne výkyvy, ktoré predstavujú cyklus, ktorý sa vyskytuje počas kalendárneho roka. V priebehu roka začínajú najazdené kilometre a HDP na nízkej hodnote, následne rastú a na konci roka opäť klesajú.

Obrázok č. 5 zobrazuje rezíduá, ktoré nemožno vysvetliť ani trendom, ani sezónnymi zložkami. Rezíduá sa vypočítajú odčítaním grafu pozorovania (obrázok č. 2) od grafov trendu a sezónnej zložky. Rezíduá zodpovedajú náhodným chybám, označovaným aj ako biely šum, a predstavujú informácie, ktoré nevieme modelovať ani predpovedať kvôli ich náhodnosti [11].

Obrázok č. 5: Dekompozícia počtu najazdených kilometrov a HDP – rezíduum



Zdroj: vlastné spracovanie autorov podľa údajov [21]

4. CHARAKTERISTIKY NA POROVNANIE MODELOV

Na základe podobnosti trendu a sezónnosti vývoja HDP a počtu najazdených kilometrov (obrázok č. 2, obrázok č. 3, obrázok č. 4) možno predpokladať, že práve počet najazdených kilometrov predstavuje tie dodatočné poznatky o aktuálnom stave vývoja HDP, ktoré je možné získať s minimálnym oneskorením rádovo v dňoch, oproti údajom na výpočet HDP, ktoré sa získavajú rádovo v mesiacoch. Preto práve s pomocou hodnoty počtu najazdených kilometrov v danom čase je možné odhadnúť aktuálnu hodnotu HDP alebo aspoň jej vývoj (percentuálny rast, pokles alebo stagnáciu).

Výkonnosť odhadov na testovacej množine bola vyhodnotená pomocou priemernej absolútnej percentuálnej chyby (MAPE), ktorá udáva mieru presnosti odhadu pre prognostické metódy. Výhodou MAPE je ľahká interpretovateľnosť a nízka závislosť od rozsahu údajov. MAPE teda vyjadruje percentuálny podiel toho, ako veľmi sa odhadované hodnoty v priemere odchyľujú od pozorovaných alebo skutočných hodnôt, či už bola predpoveď vyššia, alebo nižšia ako pozorované hodnoty a je vypočítaná nasledovnou rovnicou [4]:

$$MAPE = \frac{1}{n} \times \sum_{t=1}^{n} \left(\frac{|e_t|}{A_t} \right) \times 100 \qquad (1)$$

V tejto rovnici je A_t skutočná hodnota v bode t v čase a e_t je chyba predpovede v bode t v čase, definovaná ako A_t - F_t, pričom F_t je predpovedaná hodnota v bode t v čase, n je jednoducho počet predpovedí. V našom prípade, keďže len odhadujeme aktuálnu hodnotu, n = 1.

Odmocnina zo strednej kvadratickej chyby (RMSE) je najbežnejšou metrikou používanou na meranie presnosti prognostického modelu časového radu. Vypočíta sa pomocou druhej odmocniny zo strednej hodnoty štvorcových rozdielov medzi skutočnými hodnotami a predpoveďami modelu. RMSE sa často uvádza v jednotkách chyba na jednotku, čo umožňuje porovnávať rôzne modely na rovnakom základe. Na porovnávanie odhadov HDP bola použitá odmocnina zo strednej kvadratickej

percentuálnej chyby (RMSPE) ako variant RMSE na percentuálne vyjadrenie, ktorá podobne ako MAPE nezávisí od rozsahu hodnôt údajov [3]:

$$RMSPE = \sqrt{\frac{\sum_{t=1}^{n}\left(\frac{e_t}{A_t}\right)^2}{n}} \times 100 \qquad (2)$$

4.1. Stacionárne procesy a procesy „random walk“

Existujú výnimočné situácie, v ktorých sa časový rad dá predpovedať len pomocou jednoduchých metód. Ide o špeciálne prípady, keď sa proces vyvíja náhodne a nedá sa predpovedať pomocou štatistických metód učenia. To znamená, že ide o proces, ktorý sa nazýva random walk a treba ho vedieť rozpoznať, aby sa dalo naplánovať ďalšie analytické modelovanie. Random walk je proces, pri ktorom je rovnaká šanca, že časový rad bude stúpať alebo klesať o náhodné číslo [22].

4.2. Testovanie stacionarity

Stacionárny časový rad je taký, ktorého štatistické vlastnosti sa v čase nemenia – má konštantný priemer, rozptyl a autokoreláciu a tieto vlastnosti sú nezávislé od času [17]. Mnohé prognostické modely predpokladajú stacionaritu a je možné ich použiť len vtedy, ak je overené, že údaje sú skutočne stacionárne. V opačnom prípade je potrebné časový rad transformovať.

Bežným testom stacionarity je takzvaný augmentovaný Dickeyho-Fullerov test (ADF), ktorý overuje nulovú hypotézu H_0: „v časovom rade existuje jednotkový koreň“. Alternatívna hypotéza znie, že jednotkový koreň neexistuje, a preto je časový rad stacionárny. Výsledkom tohto testu je ADF štatistika, ktorou je záporné číslo. Čím je menšie než nula, tým silnejšie je zamietnutie nulovej hypotézy [5] a pri jeho implementácii v štatistických programoch aj *p*-hodnota. Nulová hypotéza sa zamietne, ak je *p*-hodnota menšia ako 0,05. Testovanie stacionarity časového radu HDP sa realizovalo nasledujúcim spôsobom:

- transformácia časového radu do dátovej štruktúry časového radu,
- prevzorkovanie časového radu HDP zo štvrťročnej frekvencie na mesačnú. Pre chýbajúce hodnoty bol použitý takzvaný „forward filling“ – hodnoty pre daný kvartál sa použili pre každý chýbajúci mesiac v kvartáli[1],
- čistenie časového radu od nečíselných hodnôt,
- ADF testovanie.

Výsledná štatistika testu sa rovná -0,091, *p*-hodnota sa rovná približné 0,95 pri počte použitých oneskorení 12, čiže časový rad nie je stacionárny a bolo ho potrebné v ďalšom kroku transformovať.

4.3. Transformácia časového radu

Transformácia je matematická manipulácia s údajmi, ktorá stabilizuje ich strednú hodnotu a rozptyl, čo odstraňuje alebo znižuje vplyv trendu a sezónnosti a tým sa údaje stávajú stacionárnymi. Najjednoduchšou transformáciou, ktorú možno použiť, je diferenciácia, ktorá zahŕňa výpočet zmeny od jedného časového okamihu k druhému.

[1] Pre forward filling „chýbajúcich hodnôt“ bola využitá funkcia dataframe,ffill() v knižnici pandas v jazyku python.

Pri jednorazovom diferencovaní sa použije diferenciácia prvého rádu. Pri druhom použití by išlo o diferencovanie druhého rádu. Na získanie stacionárneho radu často nie je potrebné diferencovať viac ako dvakrát [17]. V prípade rýchleho odhadu HDP bola vybraná metóda na korekciu pri sezónnych dátach cez odčítanie kĺzavého priemeru za kvartál – teda za štyri oneskorenia.

Po aplikácii transformácie na časový rad sa opakovalo testovanie stacionarity pomocou ADF testu na určenie, či je potrebné aplikovať ďalšiu transformáciu, aby sa časový rad stal stacionárnym. Výsledkom testu bolo konštatovanie stacionarity časového radu.

4.4. Autokorelačná funkcia

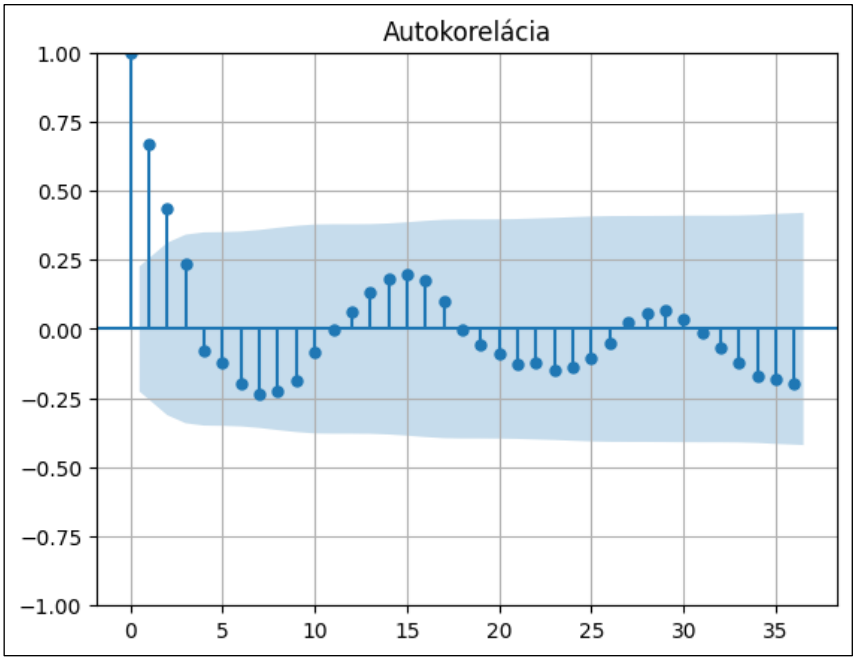
Po potvrdení stacionarity bolo použité vykreslenie autokorelačnej funkcie (ACF) na vylúčenie „random walk“. Autokorelácia meria lineárny vzťah medzi oneskorenými hodnotami časového radu (čiže tej istej premennej v rámci časového radu, ale v inom čase). ACF teda odhaľuje, ako sa korelácia medzi ľubovoľnými dvoma hodnotami mení s rastúcim oneskorením [8].

V časovom rade HDP je oneskorenie jednoducho počet časových krokov, ktoré delia dve hodnoty. Ak neexistuje autokorelácia, tak ide o časový rad, ktorý zodpovedá random walk.

ACF v prípade prítomnosti trendu ukáže, že koeficienty sú vysoké pri krátkych oneskoreniach a s rastúcim oneskorením lineárne klesajú. Ak sú údaje sezónne, tento graf bude takisto zobrazovať cyklické vzory [8].

V časovom rade HDP existuje autokorelácia (obrázok č. 6), keďže s rastúcim oneskorením sa jej hodnota vyvíja v sínusoide, takže nie je nulová. Z toho vyplýva, že sa nejde o random walk. Vysoký koeficient je vo funkcii len pri nulovom oneskorení, pri vyššom oneskorení sa už nenachádzajú žiadne výrazné koeficienty.

Obrázok č. 6: Autokorelácia časového radu HDP v mil. eur



Zdroj: vlastné spracovanie autorov podľa údajov [21]

5. ANALYTICKÉ MODELY

Z výsledkov analýzy časového radu je zrejmé, že ide o časový rad, ktorý je nestacionárny a ktorý obsahuje sezónnosť. Na podobné prognózy slúži model SARIMA a jeho obmena SARIMAX na zakomponovanie externej premennej, ktorá súvisí s modelovaným časovým radom [9].

Na výber optimálneho modelu sa použilo Akaikeho informačné kritérium (AIC). AIC odhaduje kvalitu modelu v porovnaní s ostatnými modelmi. Vzhľadom na to, že pri prispôsobení modelu dôjde k strate určitej informácie, AIC kvantifikuje relatívne množstvo informácie, ktorú model stratil. Čím menej informácií sa stratí, tým nižšia je hodnota AIC a tým lepší je model. **Výber modelu podľa AIC umožňuje udržať rovnováhu medzi zložitosťou modelu a jeho dobrou zhodou s údajmi.** AIC je z definície funkciou počtu odhadovaných parametrov *k* a maximálnej hodnoty funkcie vierohodnosti modelu, ako je uvedené v rovnici: $AIC = 2k - 2\ln(\hat{L})$.
AIC kvantifikuje kvalitu modelu len vo vzťahu k iným modelom. Je to teda relatívna miera kvality [2].

5.1. Model SARIMA

Autoregresný integrovaný kĺzavý priemer (SARIMA) je kombináciou autoregresného procesu AR(*p*), integrácie I(*d*) a procesu kĺzavého priemeru MA(*q*). Rovnako ako proces ARMA, aj proces ARIMA vychádza z predpokladu, že súčasná hodnota závisí od minulých hodnôt, ktoré pochádzajú z časti AR(*p*), a minulých chýb, ktoré pochádzajú z časti MA(*q*). Namiesto pôvodného radu označeného ako y_t však proces ARIMA používa diferencovaný rad označený ako y'_t, ktorý mohol byť diferencovaný viac ako raz. Podobne ako v procese ARMA, rád *p* určuje, koľko oneskorených hodnôt radu je zahrnutých do modelu, zatiaľ čo rád *q* určuje, koľko oneskorených chybových členov je zahrnutých do modelu. Rád *d* je definovaný ako rád integrácie. Integrácia je jednoducho opačný postup ako diferenciácia. Rád integrácie sa teda rovná počtu diferencií, ktoré boli vykonané, aby sa rad stal stacionárnym. Ak rad diferencujeme raz a stane sa stacionárnym, potom *d = 1*. Ak rad diferencujeme dvakrát, aby sa stal stacionárnym, potom *d = 2*.

O časovom rade, ktorý možno urobiť stacionárnym použitím diferencovania, sa hovorí, že je integrovaným radom. V nestacionárnom integrovanom časovom rade môžeme na tvorbu prognóz alebo odhadov použiť model ARIMA(*p, d, q*). Zjednodušene povedané, model ARIMA je jednoducho model ARMA, ktorý možno použiť na nestacionárne časové rady. Zatiaľ čo model ARMA(*p,q*) vyžaduje, aby bol rad pred prispôsobením modelu ARMA(*p, q*) stacionárny, model ARIMA(*p, d, q*) možno použiť na nestacionárne časové rady. Treba však nájsť rád integrácie *d*, ktorý zodpovedá minimálnemu počtu diferencií, ktoré sa musia vykonať, aby sa rad stal stacionárnym. Keď *d = 0*, je model ekvivalentný modelu ARMA(*p, q*). To tiež znamená, že na to, aby boli rady stacionárne, nebolo potrebné ich diferencovať [9].

Pridaním sezónnych javov v časových radoch ako ďalšej vrstvy zložitosti k modelu ARIMA získame model SARIMA. Keďže SARIMA vnáša do modelu sezónnosť ako parameter, je výrazne výkonnejší ako ARIMA pri predpovedaní komplexných časových radov obsahujúcich sezónne cykly. Význam sezónnosti je celkom zrejmý aj pre časový rad HDP či počet najazdených kilometrov a ARIMA túto informáciu implicitne nezachytáva, čo by sa prejavilo na presnosti tohto modelu. Modely SARIMA dokážu túto informáciu zachytiť a zodpovedajúcim spôsobom upraviť prognózy. Napriek tejto

výhode majú modely SARIMA v porovnaní s modelmi ARIMA aj niektoré nevýhody. Jednou z nich je, že vyžadujú viac parametrov na odhad, čo môže zvýšiť zložitosť a výpočtové náklady modelu. Ďalšou nevýhodou je, že nemusia dobre fungovať, keď údaje majú nesezónne trendy alebo štrukturálne zmeny, ako sú zmeny v správaní spotrebiteľov alebo trhových podmienkach. Modely SARIMA predpokladajú, že sezónne cykly sú stabilné a konzistentné v čase, čo nemusí byť v prípade niektorých údajov reálne [6]. V týchto prípadoch môžu byť modely ARIMA flexibilnejšie a robustnejšie.

Sezónny autoregresný integrovaný kĺzavý priemer je model SARIMA $(p,d,q)(P,D,Q)_m$, ktorý pridáva ďalšiu sadu parametrov umožňujúcich zohľadniť periodické zákonitosti pri prognózovaní časového radu, čo nie je vždy možné pri modeli ARIMA(p, d, q). Ide o štyri nové parametre v modeli, pričom prvé tri P,D,Q majú rovnaký význam ako v modeli ARIMA(p, d, q), ale sú to ich sezónne ekvivalenty. Parameter m znamená frekvenciu. V kontexte časového radu je frekvencia definovaná ako počet pozorovaní za cyklus a dĺžka cyklu závisí od súboru údajov. Pri údajoch, ktoré boli zaznamenané každý rok, štvrťrok, mesiac alebo týždeň, sa za dĺžku cyklu považuje jeden rok. Ak sa údaje zaznamenávali ročne, $m = 1$, pretože za rok je len jedno pozorovanie. Ak sa údaje zaznamenávali štvrťročne, $m = 4$, pretože v roku sú štyri štvrťroky, a teda štyri pozorovania za rok. Samozrejme, ak sa údaje zaznamenávali mesačne, $m = 12$. A napokon pri týždenných údajoch je $m = 52$. P je rád sezónneho procesu AR(P), D je rád sezónnej integrácie a Q je rád sezónneho procesu MA(Q). Model SARIMA$(p, d, q)(0,0,0)_m$ je ekvivalentný modelu ARIMA(p, d, q).

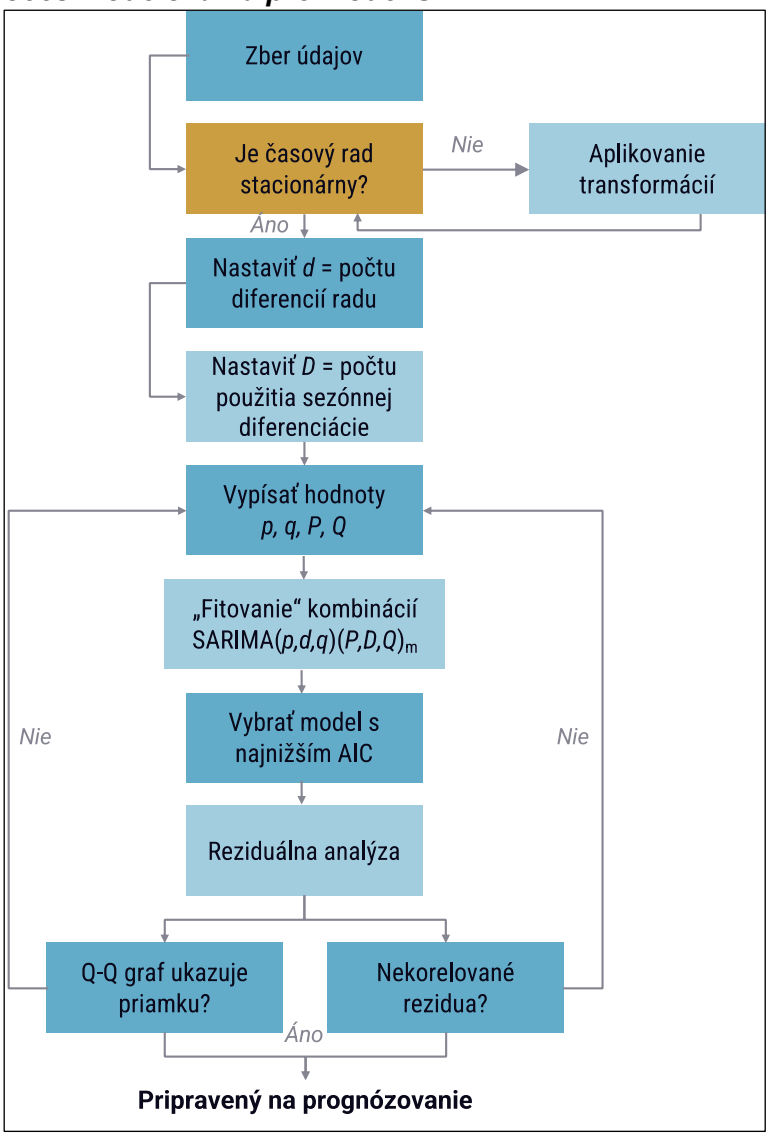
Obrázok č. 7 znázorňuje proces, ktorý je potrebný dodržať pri modelovaní. Prvý krok zberu údajov zostáva nedotknutý. Následne sa kontroluje stacionárnosť a aplikuje transformácia, aby sa stanovil parameter d. Môže sa však vykonať aj sezónna diferenciácia, aby bol rad stacionárny, a D sa bude rovnať minimálnemu počtu aplikovania sezónnej diferenciácie (I v SARIMA).

Potom sa nastaví rozsah možných hodnôt p, q, P a Q, keďže model SARIMA môže zahŕňať aj poradie sezónnych autoregresných a sezónnych kĺzavých priemerov. Pridaním týchto dvoch nových parametrov sa zvýši počet jedinečných kombinácií modelov SARIMA$(p, d, q)(P,D,Q)_m$, ktoré je možné prispôsobovať. Následne sa vyberie model s najnižším AIC a vykoná sa analýza rezíduí pred použitím modelu na prognózovanie [2]. Kvalitatívna časť analýzy rezíduí sa vykonáva pomocou Q-Q grafu. Q-Q graf je graf kvantilov dvoch rozdelení oproti sebe. Pri prognózovaní časových radov sa vykresľuje rozdelenie rezíduí na osi y oproti teoretickému normálnemu rozdeleniu na osi x. Tento grafický nástroj umožňuje posúdiť vhodnosť vybraného modelu. Ak sa rozdelenie rezíduí podobá normálnemu rozdeleniu, Q-Q graf znázorňuje priamku ležiacu na $y = x$. To znamená, že model je dobre prispôsobený, pretože rezíduá sú podobné bielemu šumu. Na druhej strane, ak sa rozdelenie rezíduí líši od normálneho rozdelenia, zobrazí sa na Q-Q grafe zakrivená priamka. Potom je možno konštatovať, že vybraný model nie je dobre prispôsobený, pretože rozdelenie rezíduí sa nepodobá normálnemu rozdeleniu, a preto rezíduá nie sú podobné bielemu šumu [13].

Hoci je Q-Q graf rýchlou metódou na posúdenie kvality vybraného modelu, táto analýza zostáva subjektívna. Preto je vhodné analýzu rezíduí ďalej podporiť

kvantitatívnou metódou použitím Ljungovho-Boxovho testu. Po analýze Q-Q grafu a zistení, že rezíduá sú približne normálne rozdelené, možno použiť Ljungov-Boxov test, aby sa preukázalo, že rezíduá nie sú korelované. Dobrý model má rezíduá, ktoré sú podobné bielemu šumu, takže rezíduá by mali byť normálne rozdelené a nekorelované. Ljungov-Boxov test je štatistický test, ktorý určuje, či sa autokorelácia skupiny údajov významne líši od 0. Pri prognózovaní časových radov je uplatňovaný Ljungov-Boxov test na rezíduá modelu, s cieľom otestovať, či sú podobné bielemu šumu. Nulová hypotéza hovorí, že údaje sú nezávisle rozdelené, čo znamená, že neexistuje autokorelácia. Ak je *p*-hodnota väčšia ako 0,05, nie je možné zamietnuť nulovú hypotézu, čo znamená, že rezíduá sú nezávisle rozdelené. Autokorelácia teda neexistuje, rezíduá sú podobné bielemu šumu a model možno použiť na prognózovanie. Ak je *p*-hodnota menšia ako 0,05, je nulová hypotéza zamietnutá, čo znamená, že rezíduá nie sú nezávisle rozdelené a sú korelované a model nie je možné použiť na prognózovanie [12].

Obrázok č. 7: Proces modelovania pre model SARIMA



Zdroj: vlastné spracovanie autorov

5.2. Model SARIMAX

Model SARIMAX ďalej rozširuje model SARIMA(*p, d, q*)(*P, D, Q*)$_m$ o vplyv exogénnych premenných. V štatistike sa termín exogénny používa na označenie prediktorov alebo vstupných premenných, zatiaľ čo pojem endogénny sa používa na definovanie cieľovej premennej – teda toho, čo sa snažíme predpovedať alebo odhadnúť v prítomnosti (rýchle odhady). To umožňuje modelovať vplyv vonkajších premenných na aktuálnu alebo budúcu hodnotu časového radu. Preto možno súčasnú hodnotu časového radu vyjadriť jednoducho ako model SARIMA(*p, d, q*)(*P, D, Q*)$_m$, ku ktorému pridáme ľubovoľný počet exogénnych premenných, ako je uvedené v nasledujúcej rovnici [16]:

$$y_t = SARIMA(p,d,q)(P,D,Q)_m + \sum_{i=1}^{n} \beta_i X_t^i \quad (3)$$

Po diferencovaní sa časový rad y_t bude označovať y'_t.

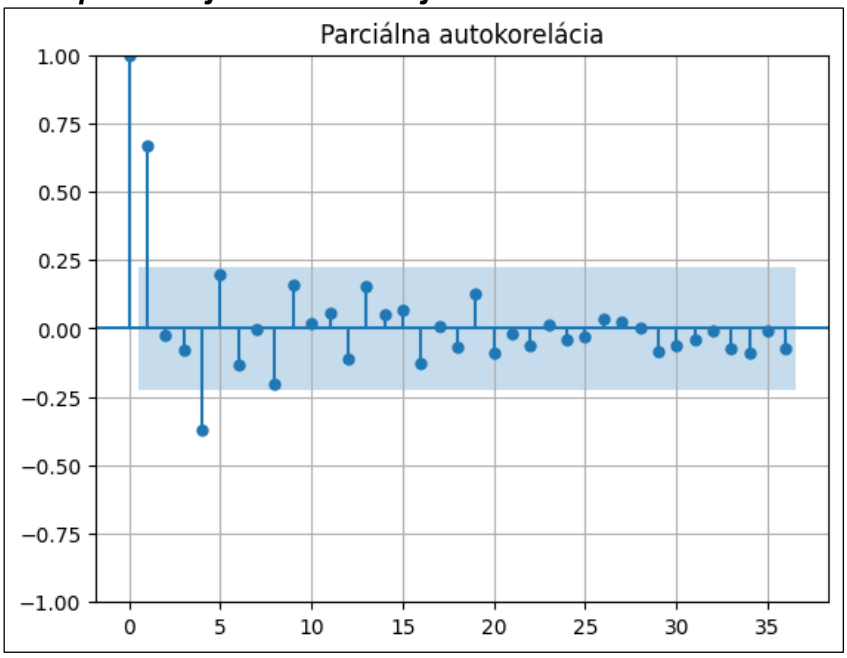
Model SARIMAX je najvšeobecnejší model na predpovedanie časových radov, ktorý umožňuje zohľadniť sezónne vplyvy, autoregresné procesy, nestacionárne časové rady, procesy s kĺzavým priemerom a exogénne premenné v jednom modeli. V dokonalej situácii modelovania sú rezíduá modelu bielym šumom. To znamená, že model zachytil všetky prediktívne informácie a zostala len náhodná fluktuácia, ktorú nemožno modelovať. Rezíduá teda musia byť nekorelované a mať normálne rozdelenie [1].

Analýza rezíduí má dva aspekty: kvalitatívnu analýzu a kvantitatívnu analýzu. Kvalitatívna analýza sa zameriava na štúdium Q-Q grafu, zatiaľ čo kvantitatívna analýza určuje, či sú naše rezíduá nekorelované. Q-Q graf sa vytvorí tak, že sa na os *y* vynesú kvantily rezíduí oproti kvantilom teoretického rozdelenia, v tomto prípade normálneho rozdelenia, na osi *x*. Výsledkom je graf rozptylu. Rozdelenie sa porovná s normálnym rozdelením, pretože sa žiada, aby rezíduá boli podobné bielemu šumu, ktorý je normálne rozdelený. Ak sú obe rozdelenia podobné, čo znamená, že rozdelenie rezíduí je blízke normálnemu rozdeleniu, Q-Q graf zobrazí priamku, ktorá približne leží na *y* = *x*. To zasa znamená, že náš model dobre zodpovedá našim údajom [17].

Parametre výsledného modelu SARIMAX sa vypočítali cez externý nástroj JDemetra+[2], z čoho vyšiel model: **SARIMA(*0, 1, 0*)(*0, 1, 1*)$_4$.** Tieto parametre sme aj čiastočne odvodili zo sezónnych cyklov v dĺžke jedného roka, pričom údaje sa zaznamenávali štvrťročne, teda *m* = *4*. Ďalej bolo potrebné časový rad HDP jedenkrát transformovať na stabilizáciu strednej hodnoty, z čoho vyplýva, že *d* sa rovná 1, a jedenkrát bolo potrebné aplikovať sezónnu diferenciáciu, z čoho takisto vyplýva, že *D* sa rovná 1. Z grafu funkcie autokorelácie (obrázok č. 6) ako aj z grafu parciálnej autokorelačnej funkcie (PACF – obrázok č. 8) vyplýva, že nemožno jednoducho určiť rády *p* a *q*, a treba prispôsobovať model ARMA, ktorý však podľa spomínaného nástroja Jdemetra+ má nulový rád *p* a *q* a pre ekvivalentnú sezónnu zložku ide o autoregresný model prvého rádu, teda *Q* = *1*.

[2] *Zdroj: https://cros-legacy.ec.europa.eu/content/software-jdemetra_en.*

Obrázok č. 8: Graf parciálnej autokorelačnej funkcie časového radu HDP v mil. eur

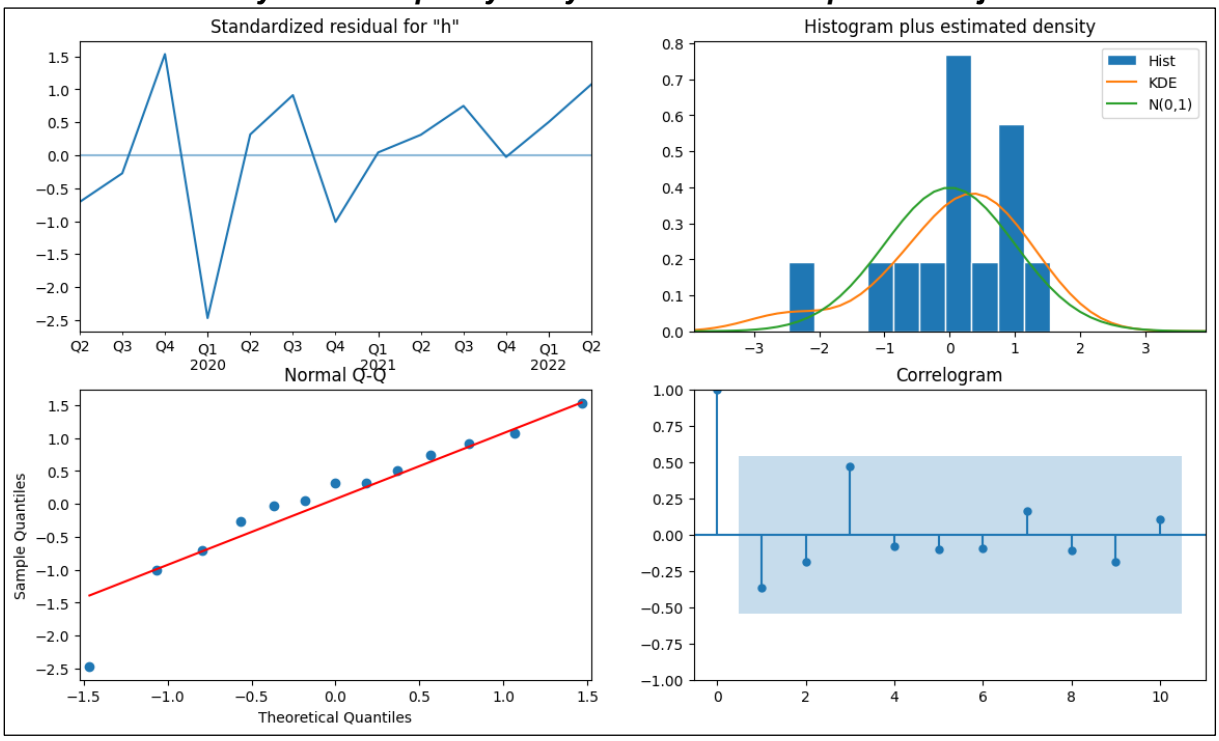


Zdroj: vlastné spracovanie autorov

V tejto analýze možno vidieť *p*-hodnotu spojenú s každým koeficientom každej premennej pre predpoveď v modeli SARIMAX. Často sa *p*-hodnota zneužíva ako spôsob, ako vykonať výber premenných („features“). Mnohí nesprávne interpretujú *p*-hodnotu ako spôsob určenia, či je premenná v modeli korelovaná s cieľom. V skutočnosti *p*-hodnota testuje, či sa koeficient významne líši od 0 alebo nie. Ak je *p*-hodnota menšia ako 0.05, zamietame nulovú hypotézu a usudzujeme, že koeficient je významne odlišný od 0. Neurčuje, či je premenná užitočná na prognózovanie. Preto by sa nemali odstraňovať premenné na základe ich *p*-hodnoty. O tento krok sa postará výber modelu na základe minimalizácie AIC [10].

Q-Q graf pre tento model sa nachádza na obrázku č. 10 v ľavej dolnej časti, kde sú rezíduá normálne rozdelené, s výnimkou nedokonalosti na konci intervalu. Na tomto obrázku možno vidieť, ako `statsmodels` uľahčuje kvalitatívnu analýzu rezíduí. V ľavom hornom grafe sú znázornené rezíduá v celom súbore údajov. Vidno, že neexistuje žiadny trend, hoci priemer sa nezdá stabilný v čase. I keď v rezíduách nie je žiadny trend, zdá sa, že rozptyl nie je konštantný, čo je rozdiel v porovnaní s bielym šumom. Vpravo hore je histogram rezíduí. Teoreticky by mohol byť blízko normálnemu rozdeleniu, avšak nemáme na lepší tvar histogramu dostatok hodnôt. V tejto podobe však naznačuje, že rezíduá nie sú blízke bielemu šumu, keďže biely šum je normálne rozdelený. A napokon, graf vpravo dole ukazuje autokorelačnú funkciu rezíduí. Pri oneskorení 0 je jediný významný vrchol a v ostatných prípadoch je minimum významných koeficientov. To znamená, že rezíduá sú minimálne korelované, čo by mohlo byť ešte vylepšené, ak by sme mali k dispozícii viac hodnôt.

Obrázok č. 9: Analýza rezíduí pre vybraný model SARIMAX podľa zdrojového kódu



Zdroj: vlastné spracovanie autorov

Posledným krokom analýzy rezíduí je použitie Ljungovho-Boxovho testu. Ten umožňuje kvantitatívne posúdiť, či sú rezíduá skutočne nekorelované. Na vykonanie Ljungovho-Boxovho testu na rezíduá bola použitá funkcia acorr_ljungbox zo statsmodels. Funkcia prijíma ako vstup rezíduá, ako aj zoznam oneskorení. V tomto prípade bola vypočítaná Ljungova-Boxova štatistika a *p*-hodnoty pre 10 oneskorení [14].

Výsledný zoznam *p*-hodnôt ukazuje, že každá z nich je vyššia ako 0,05. Preto pri každom oneskorení nie je možné zamietnuť nulovú hypotézu, čo znamená, že rezíduá sú nezávisle rozložené a nekorelované. Z našej analýzy vyvodzujeme záver, že rezíduá sú podobné bielemu šumu.

V ďalšom kroku vyhodnotíme koreláciu medzi exogénnymi premennými, ktoré sú použité v modeli SARIMAX.

Tabuľka č. 1: Korelačná matica exogénnych premenných na odhad premennej hdp_amount modelom SARIMAX

	hdp_ amount	distance	vehicle_ count
hdp_ amount	1,00	0,79	0,78
distance	0,79	1,00	0,91
vehicle_ count	0,78	0,91	1,00

Zdroj: vlastné spracovanie autorov

Je potrebné uviesť, že pri nowcastingu nie je problém multikolinearity medzi exogénnymi premennými významný – respektíve jeho dôsledky model neohrozujú. To okrem iného potvrdzuje aj porovnanie metriky AIC (Tabuľka č. č. 2), ktorá je minimálna práve pre model SARIMAX využívajúci obe exogénne premenné v tabuľke (tabuľka č. 2), ako aj obrázok č. 10 znázorňujúci rôznu kvalitu rýchlych odhadov uvedených

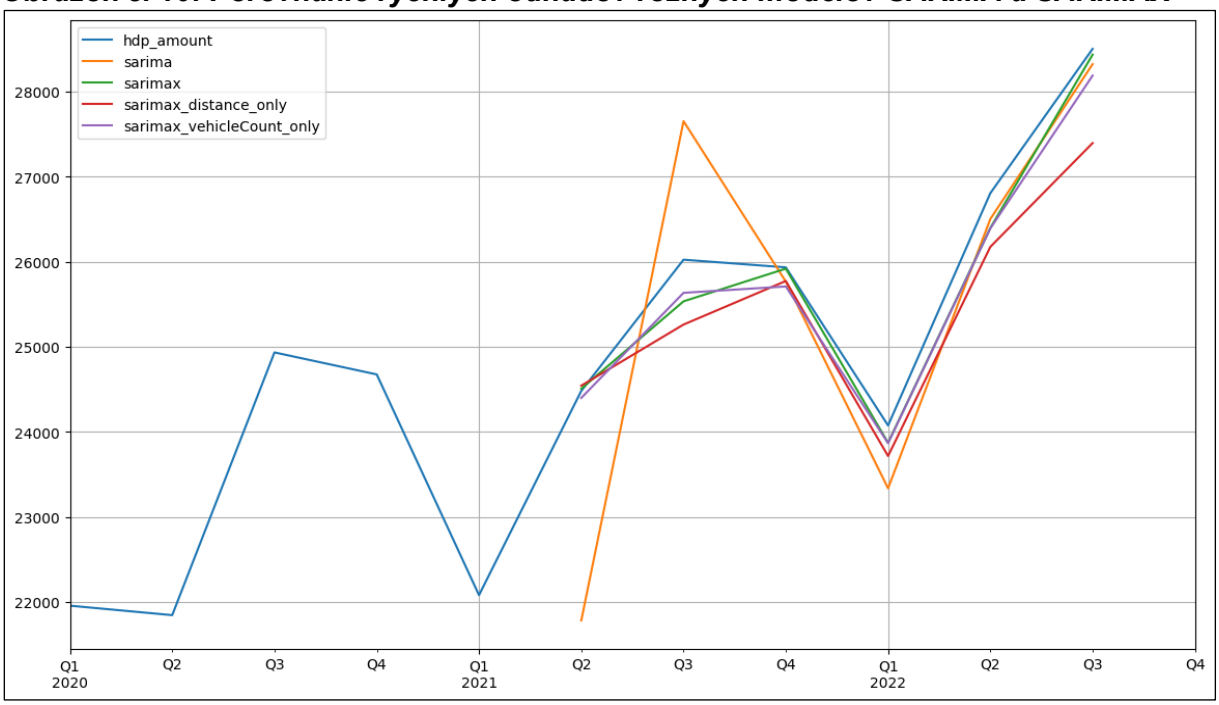
modelov, pričom zelenou je znázornený model SARIMAX s oboma exogénnymi premennými. Na tomto obrázku tiež vidno, že model SARIMA bez exogénnych premenných nedokáže spoľahlivo predpovedať vývoj v roku 2021, ktorý bol ovplyvnený aj pandémiou COVID-19.

Tabuľka č. 2: Porovnanie AIC pre modely SARIMA a SARIMAX

Typ modelu	**AIC**
SARIMA	225,66
SARIMAX s premennou distance	201,72
SARIMAX s premennou vehicle_count	195,13
SARIMAX s premennými distance a vehicle_count	**155,20**

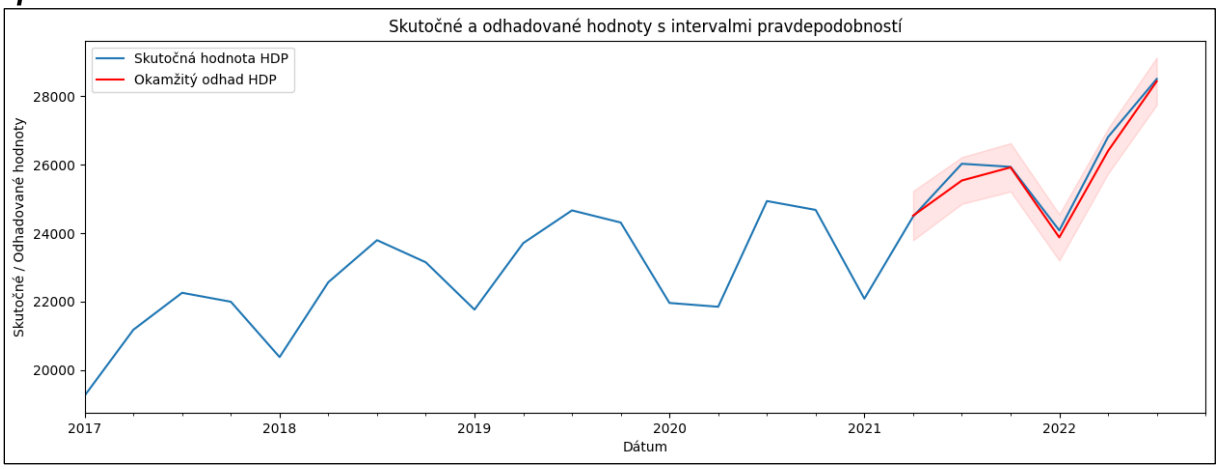
Zdroj: vlastné spracovanie autorov

Obrázok č. 10: Porovnanie rýchlych odhadov rôznych modelov SARIMA a SARIMAX



Zdroj: vlastné spracovanie autorov

Obrázok č. 11: Skutočné a odhadované hodnoty HDP v mil. eur s intervalmi spoľahlivosti



Zdroj: vlastné spracovanie autorov

Navyše obrázok č. 11 znázorňuje veľmi uspokojivý odhad HDP s intervalmi spoľahlivosti.

Pri používaní modelu SARIMAX platí dôležité upozornenie. Zahrnutie externých premenných môže byť potenciálne prospešné, pretože sa dajú nájsť silné prediktory – vstupné premenné do modelu pre cieľovú premennú. Pri prognózovaní viacerých časových krokov do budúcnosti však možno naraziť na problémy. Model SARIMAX používa model SARIMA($p\ d,\ q$)($P,\ D,\ Q$)$_m$ a lineárnu kombináciu exogénnych premenných na predpovedanie jedného časového kroku do budúcnosti. Ale čo ak treba predpovedať dva časové kroky do budúcnosti? Zatiaľ čo s modelom SARIMA je to možné, model SARIMAX vyžaduje, aby sa predpovedali aj exogénne premenné. S týmto problémom sa však pri rýchlych odhadoch nestretneme.

6. ZÁVER

Model SARIMAX nám umožňuje zahrnúť externé premenné, označované aj ako exogénne premenné, do nowcastingu cieľovej premennej – teda aktuálnej hodnoty HDP v mil. eur v bežných cenách. Transformácie sa uplatňujú len na cieľovú premennú, nie na exogénne premenné. Vplyv nového dátového zdroja sa testoval na odhad HDP od Q2 2021 do Q4 2022, kde aj len pozorovaním kriviek na obrázku č. 11 vidno, že model SARIMAX najlepšie kopíruje krivku skutočnej hodnoty HDP v mil. eur v bežných cenách.

Pri aplikácii modelu SARIMAX len na rýchle odhady nemusíme predpovedať viacero časových krokov do budúcnosti, a tak sa nemusia predpovedať ani exogénne premenné. To znamená, že táto vlastnosť modelu SARIMAX nezväčšuje chyby konečného odhadu. Model SARIMAX je vysvetliteľný, nie je príliš komplikovaný ani náročný na výpočet. Využitím modelu prispôsobeného aktuálnej verzii časového radu HDP so všetkými dostupnými hodnotami v modelovanom období minimalizujeme problém chyby prognóz tým, že odhadujeme aktuálnu hodnotu HDP v mil. eur.

Výkonnosť a presnosť modelu SARIMAX potvrdzuje aj tabuľka č. 3, v ktorej sú uvedené pre všetky modely z obrázka č. 11 percentuálne metriky chyby RMSPE a MAPE. Na základe definícií týchto metrík je zrejmé, že čím je nižšia hodnota metriky uvádzaná v percentách, tým o menej percent sa odhad daného modelu líši od skutočnej hodnoty HDP v mil. eur v bežných cenách. Pre model SARIMAX vieme podľa typu metriky dosiahnuť chybu menej ako 9 percent alebo menej ako 0,7 percenta.

Tabuľka č. 3: Porovnanie jednotlivých analytických modelov podľa metrík RMSPE a MAPE

	sarima	sarimax	sarimaxdistance_only	sarimax_vehicle Count _only
RMSPE	17,44	**8,91**	24,92	11,43
MAPE	1,37	**0,67**	2,09	1,10

Zdroj: vlastné spracovanie autorov

Výhodou získaných údajov je ich včasnosť a rozsah. Do systému elektronického mýta sú zahrnuté všetky motorové vozidlá s celkovou hmotnosťou nad 3,5 tony alebo jazdné súpravy s celkovou hmotnosťou nad 3,5 tony. Údaje tiež obsahujú anonymizované identifikátory jednotlivých nákladných vozidiel, z ktorých nie je možné opätovne zistiť či už evidenčné číslo vozidla, majiteľa vozidla alebo identitu vodiča.

Na druhej strane údaje nie sú dostupné priamo vo formáte, ktorý podporuje prelinkovanie, avšak úseky ciest majú svoje identifikátory, pomocou ktorých sa dajú vkresliť do mapy cestnej siete. To možno v budúcnosti využiť napríklad na odhad, aké percento nákladných vozidiel bolo len v tranzite cez Slovenskú republiku. Očistenie dát od tranzitnej nákladnej dopravy by mohlo mať pozitívny vplyv na kvalitu vstupných hodnôt do modelov. Tým by sa mohli eliminovať aj exogénne faktory ovplyvňujúce nákladnú dopravu, ako je geopolitická situácia v okolitých krajinách (vojnový konflikt na Ukrajine) či situácia na hraničných priechodoch (blokáda priechodu Vyšné Nemecké), na ktoré je súčasný model náchylný. Vytvorený model by tiež mohol zefektívniť vypovedaciu schopnosť zahrnutím údajov o preprave tovaru železničnou dopravou, ktorá tvorí okolo 15 % celkovej nákladnej dopravy[3].

LITERATÚRA

[1] ALHARBI, F. R. – CSALA, D.: A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach. In: Inventions, 2022, č. 1, s. 94.
[2] AKAIKE, H.: Factor analysis and AIC. In: Psychometrika, 1987, č. 3, s. 317 – 332.
[3] ALBERTO, A. – DIAMOND, A. – HAINMUELLER, J.: Comparative politics and the synthetic control method. American Journal of Political Science, 2015, č. 2, s. 495–510.
[4] ALLWRIGHT, S. 2023. How to interpret MAPE. [online]. [cit. 11-09-2023]. Dostupné na: https://stephenallwright.com/interpret-mape/.
[5] BAUM, C.: Tests for stationarity of a time series. In: Stata Technical Bulletin, 2000, č. 57, s. 36 – 39.
[6] BOX, G. – JENKINS, M.: Time series analysis: forecasting and control. John Wiley & Sons, 2015.
[7] DURBIN, J. – WATSON, G. S.: Testing for Serial Correlation in Least Squares Regression, II. In: Biometrika, 1951, č. 1 – 2, s. 159 – 179.
[8] HEILBRONNER, R. – Barrett, S. D.: Image Analysis in Earth Sciences: Microstructures and Textures of Earth Materials, Springer-Verlag Berlin Heidelberg, 2014.
[9] HOSSAIN, J.: Comparative Analysis of ARIMA, SARIMAX, and Random Forest Models for Forecasting Future GDP in Relation to Unemployment Rate. 2023.
[10] HYNDMAN, R. Statistical tests for variable selection. [online]. [cit. 11-09-2023]. Dostupné na: https://robjhyndman.com/hyndsight/tests2/.
[11] LANGBEIN, J. – HADLEY J.: Correlated errors in geodetic time series: Implications for time-dependent deformation. In: Journal of Geophysical Research: Solid Earth 102.B1. 1997. s. 591 – 603.
[12] LJUNG, G. M. – BOX, G. E. P.: On a measure of lack of fit in time series models, In: Biometrika, 1978. č. 2, s. 297 – 303.
[13] MEČIAROVÁ, K. Q-Q grafy – Oborový seminár. MFF Univerzita Karlova. 2021. [online]. [cit. 11-09-2023]. Dostupné na: https://www.karlin.mff.cuni.cz/~omelka/Soubory/nmsa401/Q-Q_plots.pdf.
[14] NIST/SEMATECH. Box-Ljung Test. In: e-Handbook of Statistical Methods. [online]. [cit. 11-09-2023]. Dostupné na: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4481.htm.

[3] Štatistický úrad SR. Súhrnné ukazovatele za dopravu [do1003rs].

[15] Najvyšší kontrolný úrad Slovenskej republiky. Elektronický výber mýta: Záverečná správa. 2019.
[16] PERKTOLD, J. – SEABOLD, S. – TAYLOR, J.: statsmodels.tsa.statespace.sarimax.SARIMAX. [online]. [cit. 11-09-2023]. Dostupné na: https://www.statsmodels.org/stable/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html#statsmodels.tsa.statespace.sarimax.SARIMAX.
[17] RAIMUNDO, B.: Time series Analysis. [online]. [cit. 11-09-2023]. Dostupné na: https://github.com/BernardoRaimundo/Time-Series-Analysis.
[18] SHUMWAY, H. – STOFFER, D.: Time series analysis and its applications. Third edition. New York: Springer, 2011.
[19] Skytoll: Elektronický výber mýta, Slovenská republika. [online]. [cit. 11-09-2023]. Dostupné na: https://www.skytoll.com/elektronicky-mytny-system-sr/.
[20] SONHAO, W.: Multicollinearity in Regression: Why it is a problem? How to check and fix it, In: Towards Data Science. [cit. 11-09-2023]. Dostupné na: https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea.
[21] Štatistický úrad Slovenskej republiky. 2023. Štvrťročné údaje HDP v bežných cenách: [nu0002qs]. [online]. [cit. 11-09-2023]. Dostupné na: http://statdat.statistics.sk/cognosext/cgi-bin/cognos.cgi?b_action=cognosViewer&ui.action=run&ui.object=storeID(%22iA188D5A85FEF4EA9B48BC102C7C66758%22)&ui.name=%C5%A0tvr%C5%A5ro%C4%8Dn%C3%A9%20%C3%BAdaje%20HDP%20v%20be%C5%BEn%C3%BDch%20cen%C3%A1ch%20%5Bnu0002qs%5D&run.outputFormat=&run.prompt=true&cv.header=false&ui.backURL=%2Fcognosext%2Fcps4%2Fportlets%2Fcommon%2Fclose.html&run.outputLocale=sk.
[22] THEODOSIOU, M.: Forecasting monthly and quarterly time series using STL decomposition. In: International Journal of Forecasting, 2011, č. 4, s. 1178 – 1195.

RESUMÉ

V rámci projektu Sociálno-ekonomické aspekty Big Data (SEABD) bol vedený experiment s cieľom preskúmať využitie údajov z elektronického mýtneho systému na vplyv vývoja hrubého domáceho produktu (HDP) ako typ prognózovania. Prognózovanie je postup, ktorý využíva historické údaje ako vstupy na vytvorenie informovaných odhadov, ktoré predpovedajú budúce trendy v dlhšom časovom horizonte. Rýchly odhad oproti tomu predstavuje postup odhadu veľmi nedávnej minulosti, súčasnosti alebo veľmi blízkej budúcnosti stavu ekonomických ukazovateľov. Cieľom projektu bolo na základe údajov o najazdených kilometroch nákladných vozidiel vytvoriť ukazovateľ nákladnej dopravy, ktorý je možné porovnať so štvrťročným ukazovateľom HDP z produkcie Štatistického Úradu SR. Elektronický mýtny systém zbiera údaje automatizovaným systémom s jasne zdokumentovanou metodikou na úsekoch diaľnic podľa platnej vyhlášky. Zozbierané údaje predstavujú agregované hodnoty počtu najazdených kilometrov a počtu unikátnych nákladných vozidiel po mesiacoch a štvrťrokoch. Výhoda tohto súboru údajov je jeho detailnosť, keďže každý najazdený kilometer môže predstavovať konkrétnu ekonomickú transakciu či činnosť. Na hodnotenie presnosti odhadu v rámci nowcastingu sa použili metriky priemernej absolútnej percentuálnej chyby (MAPE) a ich smerodajná odchýlka (RMSPE), ako aj porovnanie so skutočnými hodnotami HDP v danom kvartáli. Výsledky modelu boli veľmi uspokojivé s hodnotami pre MAPE 0,67 % a RMSPE 8,7 %. Možný budúci vývoj a význam pokračujúceho výskumu využívania údajov o najazdených kilometroch nákladných vozidiel na odhady makroekonomických

ukazovateľov spočíva v presnejšom a rýchlejšom určovaní hospodárskej aktivity. Tieto údaje môžu poskytnúť informácie v reálnom čase o pohybe tovaru a obchodu, čo je kritické pre ekonomické rozhodovanie a plánovanie. S rozvojom technológií IT a senzorov v nákladných vozidlách je možné získať ešte presnejšie a aktuálnejšie údaje.

RESUME

As part of the Socio-Economic Aspects of Big Data (SEABD) project, an experiment was conducted to investigate the use of electronic toll system data for the impact of development the gross domestic product (GDP) as a type of forecasting. Forecasting is a practice using historical data as inputs to make informed estimates that predict future trends over a longer time horizon. A flash estimate, on the other hand, represents a process of estimating the very recent past, present or very near future of the state of economic indicators. The goal of the project was to create an indicator of freight transport based on truck mileage data, which can be compared with the quarterly GDP indicator produced by the Statistical Office of the Slovak Republic. The Electronic toll system collects data using an automated system with a clearly documented methodology on sections of motorways. The collected data represents the aggregated values of the number of kilometres travelled and the number of unique trucks by month and quarter. The advantage of this dataset is its detail, as each kilometer driven can represent a specific economic transaction or activity. The mean absolute percentage error (MAPE) and their standard deviation (RMSPE) metrics were used to assess the accuracy of the nowcasting estimate, as well as a comparison with the current GDP values in the given quarter. The results of the model were satisfactory with values for MAPE 0.67% and RMSPE 8.7%. A possible future development and importance of a continued research into the use of truck mileage data for estimation of macroeconomic indicators is to more accurately and rapidly determine economic activity. This data can provide a real-time information on the movement of goods and trade, which is critical for economic decision-making and planning. With the development of the IT technologies and sensors in trucks, even more accurate and up-to-date data can be obtained.

PROFESIJNÝ ŽIVOTOPIS

Peter Knížat, MSc., *je externým študentom doktorandského štúdia na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave. Vyučuje praktické cvičenia: podpora rozhodovacích procesov a viac atribútové rozhodovanie vrátane aplikácie v štatistickom softvéri R. Pracuje ako dátový analytik v sekcii všeobecnej metodiky, registrov a koordinácie národného štatistického systému Štatistického úradu SR, kde je zodpovedný za návrh štatistickej metodiky v cenových štatistikách s využitím webscrapovaných údajov. Predtým pôsobil v medzinárodnej banke ako senior risk a portfóliový manažér, kde viedol vývoj interných modelov používaných v procesoch hodnotenia kreditného rizika.*

Dipl. Ing. Dagmar Ceľuchová Bošanská *je zakladateľkou spoločnosti Alistiq s. r. o. a expertkou na inovácie a digitálnu transformáciu s dlhoročnými skúsenosťami. V roku 2008 absolvovala inžinierske štúdium pre informačné technológie, mobilné komunikácie a štatistické spracovanie signálov na Viedenskej technickej univerzite, kde pôsobila vo vedeckom tíme na vývoji simulátorov technológií pre bezdrôtové siete štvrtej generácie. Od roku 2015 sa venuje vývoju riešenia a návrhu opatrení na zvyšovanie kvality a efektivity využívania údajov vrátane Big Data na sekundárne účely, predovšetkým vo verejnej správe. Aktuálne od roku 2020 pôsobí ako doktorand na Českom vysokom učení technickom v Prahe, kde sa venuje výskumu grafových údajov generovaných z elektronických zdravotných záznamov a ich analýze s využitím strojového učenia a veľkých jazykových modelov.*

Ing. Martin Janík *absolvoval inžinierske štúdium na Fakulte elektrotechniky a informatiky Slovenskej Technickej Univerzity v Bratislave v odbore telekomunikácie, špecializácia bezpečnosť (2008). Už počas štúdia na vysokej škole začal pracovať v súkromnom sektore v oblasti IT, spočiatku v oblasti webových neskôr mobilných technológii ako programátor, analytik a následne softvérový architekt softvérových produktov v oblasti mobilných technológií. Od roku 2022 sa venuje dátovej vede so zameraním na analýzu a návrh grafových dátových štruktúr vo verejnom sektore. Spolupracuje na Centrálnom modeli údajov SR a na analýze a spracovaní Big Data.*

Mgr. Filip Nguyen *absolvoval magisterské štúdium v Ústave pedagogiky a sociálnych štúdií Univerzity Palackého v Olomouci (CZ) v odbore pedagogika – veřejná správa (2019). Od roku 2018 pôsobí v poradenskej spoločnosti Alistiq, s. r. o. ako poradca v oblasti verejného obstarávania a verejných inovačných projektov. Jeho práca sa zameriava na návrh digitálnych služieb štátnej správy a aplikáciu osvedčených postupov PRINCE2 a Agile metodík v projektoch.*

KONTAKT

peter.knizat@statistics.sk
dagmar.bosanska@alistiq.com
martin.janik@alistiq.com
filip.nguyen@alistiq.com