https://doi.org/10.37528/FTTE/9788673954165/POSTEL.2022.009

PREDVIĐANJE OBIMA PRIMLJENIH EMS POŠILJAKA NA NIVOU SRBIJE POMOĆU SARIMA, RANDOM FORESTS I ELM MODELA

Ivana D. Rogan, Olivera R. Pronić-Rančić
Univerzitet u Nišu, Elektronski fakultet, Aleksandra Medvedeva 14, 18000 Niš, Srbija,
ivana84p@gmail.com, olivera.pronic@elfak.ni.ac.rs

Rezime: *U radu su prikazani rezultati validacije / predviđanja obima primljenih EMS (Express Mail Service) pošiljaka na nivou Srbije u periodu od 2016 do 2021. godine, zadatih mesečno. Korišćeni su sledeći modeli: sezonski autoregresivni integrisani pokretni prosek (SARIMA), algoritam nadgledanog mašinskog učenja tzv. Slučajnih šuma (RF-random forests) u XLSTAT dodatku za Excel i jedan model veštačke neuronske mreže (ANN) – Ekstremno mašinsko učenje (ELM – extreme learning machine) u Matlab okruženju. U okviru procesa validacije modela, pokazana je izrazita prednost primene RF modela u odnosu na ELM i SARIMA model. Korišćenjem RF modela dobijena je najmanja vrednost RMSE (oko 30% manja od one u ELM modelu i oko 70% manju od SARIMA metoda), kao i najkraće vreme izvršavanja programa.*

Ključne reči: *analiza vremenskih serija, EMS, SARIMA, ELM, RF.*

1. Uvod

EMS pošiljke su registrovane poštanske pošiljke bez označene vrednosti i bez posebnih usluga, mase do 30 kg. EMS pošiljke u međunarodnom saobraćaju mogu sadržati: dokumenta, robu, druge predmete, osim onih za koje važe zabrane Svetskog Poštanskog saveza i zakonodavstava pojedinih zemalja [1].

Proces prognoziranja, u okviru poslovnih fenomena, predstavlja predviđanje budućih ishoda različitih poslova. U okviru usluga brze pošte u međunarodnom saobraćaju, predviđanje obima je interesantno i samo po sebi u okviru ukupnog poslovanja kompanije. Pored periodične poslovne provere zastupljenosti pojedinih usluga kompanije na tržištu, prognoze se koriste za planiranje novih usluga.

U slučaju prognoziranja, uglavnom se koriste određeni matematički modeli, te se na osnovu prostorno-vremenski organizovanih podataka ocenjuju parametri modela i određuju vrednosti budućih podataka. U tom smislu najinteresantniji, i za praksu najvažniji, jeste koncept vremenskih serija. Neki od prvih i osnovnih modela vremenskih

serija jesu oni statistički. U okviru ovih modela koriste se odgovarajući testovi i kriterijumi kojima se verifikuje valjanost ocenjenog modela i njegovog predviđanja.

U ovom radu primenjena je jedna klasa statističkih sezoniziranih autoregresivnih modela integrisanih pokretnih sredina, tzv ARIMA (p,d,q) (P,D,Q) *s* - SARIMA. Model se zasniva na Box i Jenkins metodologiji, [2] na validaciji pomoću RMSE (*Root Mean Square Error*). Kod ove klase linearnih metoda, pretpostavka je da tekuća vrednost člana serije zavisi od vrednosti prethodnih članova serije. Tekuće vrednosti, modelski određenog tipa slučajnog procesa, sa normalnim ili sličnim raspodelama, imaju i periodičnu, sezonsku komponentu. Okruženje u kojem je analiza rađena je Excel – XLSTAT dodatak [3]. On je korišćen u [4-6], kao crna kutija.

Drugi algoritam čija će se prediktivna svojstva ispitivati za istu vremensku seriju jeste Mašina za ekstremno učenje – ELM, [7-12]. ELM je algoritam koji u suštini predstavlja jednoslojnu *feedforward* neuronsku mrežu. Njegova struktura se sastoji, pored ulaza, i od jednog sloja skrivenih čvorova. U njemu su dodeljene vrednosti težina između ulaza i skrivenih čvorova, kao i biasi (vrednosti tresholda- pragova), neophodni za aktivacione - transfer funkcije koje su slučajne veličine. To znači da nije potreban proces učenja za izračunavanje parametara modela koji se nalaze u okviru ulaznih težina. Dakle, vrednosti izlaznih težina koje povezuju skrivene čvorove i izlaze mogu se brzo dobiti računanjem matričnog Mur-Penrouzovog pseudinverza (Moore–Penrose inverse). Pomoću datog ulaznog niza i targetnog niza (trening skupa, tačnih vrednosti), na kraju se dobija izlazni niz. Praktično, testni deo ulaznog niza je neka vrsta targeta koji se realizuje izlazom, te se pomoću njega vrši njihovo upoređivanje sa targetom (validacija) pomoću RMSE i/ili R^2 (koeficijent determinacije). Najveća prednost ELM-a je njegova računarska brzina i jednostavnost. ELM se koristi za klasifikaciju uzoraka, regresiju, kao i za implementaciju raznih online modela.

Treći algoritam koji se ovde koristi je RF [13]. On, preko šuma (skupova stabala - teorija grafova) preko slučajnih odluka grupno (ansambalski) služi za klasifikaciju, regresiju i druge zadatke koji preko konstruisanja skupa stabala vrši adekvatna odlučivanja. Za zadatke regresije, vraća se srednja vrednost ili prosečno slučajno predviđanje pojedinačnih stabala po nekom kriterijumu, a sve sa tendencijom smanjivanja *RMSE*. Stabla odlučivanja su uobičajeni pojedinačni algoritmi učenja pod nadzorom, koji mogu biti skloni problemima kao što su pristrasnost i prekomerno prilagođavanje. Međutim, kada višestruka stabla odluka formiraju ansambl u algoritmu slučajne šume, ona predviđaju tačnije rezultate, posebno kada pojedina stabla nisu u korelaciji jedno s drugim. Ovde je korišćen Bagging metod u RF, poznat i kao bootstrap agregacija. To je metod učenja ansambla koji se obično koristi za smanjenje varijanse unutar skupa podataka sa šumovima. Mnogo je veći opseg problema koje efikasno rešavaju RF metodi nego ELM. Poznato je da su za ovu klasu problema dosada bili bolji samo ELM predviđanja - validacije od ARIMA [6]. Osnovni razlozi za primenu ova tri algoritma jesu njihova dosadašnja primena u slične svrhe, mogućnosti za njihovo poboljšanje i kombinovanje.

U ovom radu razmatrano je validiranje / predviđanje na osnovu obima primljenih EMS pošiljaka na mesečnom nivou u periodu od 2016. do 2021. godine (72 meseca). Predikcija je urađena u okviru ARIMA za sledeća 24 meseca (73-96). Validacija za poslednjih 12 članova serije izvršena je primenom razvijenih ARIMA, ELM i RF modela. Izvršeno je poređenje tačnosti razvijenih modela, kao i vremena izvršavanja programa.

2. Teorijske osnove SARIMA, ELM i RF modela analize vremenskih serija

Vremenska serija x_t jeste funkcija koja se sastoji od diskretnih vremenskih koraka - vremena x_t: $N \rightarrow R^+ \cup \{0\}$. Njena konkretna realizacija obuhvata konačan broj, u ovom slučaju N.

2.1. Osnove SARIMA modela u okviru dodatka XLSTAT za Excel

Osnove SARIMA modela date su u [2]. Neka je B operator vremenskog pomeraja: $Bx_t = x_{t-1}$, operator njegovog stepena $B^s x_t = x_{t-s}$, $s = 0, 1, 2 \ldots$ Osnovna jednačina, ovog linearnog, specifično nestacionarnog modela (a_t je beli šum, $\nabla x_t = (1-B)\, x_t$) je

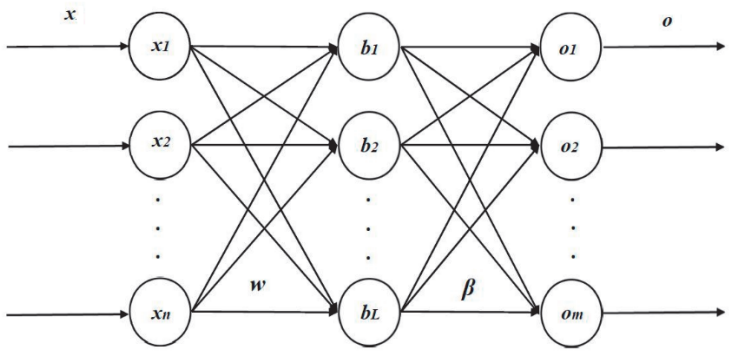
$$\varphi_p(B)\delta_P(B^s)\nabla^d\left(1-B^s\right)^D x_t = \mu + \chi_Q(B^s)\theta_q(B)a_t. \quad (1)$$

U prethodnoj jednačini sa φ, δ, χ, θ, označeni su polinomi odgovarajućih stepena. Konstanta μ opisuje stacionaran trend. S je sezonski indeks, njime se opisuje vremenska periodičnost procesa. Malim latiničnim slovima obeležene se nesezonske komponente modela, dok su velikim latiničnim slovima obeležene sezonske komponente modela. Postoje tri osnovne faze u analiziranju i modelovanju Box-Jenkins modela vremenskih serija: 1) identifikacija modela; 2) estimacija modela; 3) validacija modela (koja ovde nije samo računanje RMSE). Važni pojmovi su: deskriptivna analiza, grafici funkcije autokorelacije (ACF) i parcijalne funkcije autokorelacije (PACF) koji se porede sa teorijskim ponašanjem ovih funkcija. Pomoću njih se vrši indentifikacija modela. Postupak estimacije - fitovanja, podrazumeva nalaženje vrednosti odgovarajućih konstantni modela pomoću metode najmanjih kvadrata ili maksimalne verodostojnosti. Nakon što je odgovarajući model usvojen, i njegovi parametri estimovani, Box-Jenkins metodologija zahteva proveru kvaliteta razlike aktuelnih vrednosti i onih estimovanih, od strane prihvaćenog modela. Najčešći kriterijumi, koji su i ovde korišćeni, su AICC (*Akaike information criterion*) i SBC (*Bayesian information criterion*). U okviru procene količine informacija koje je model izgubio, AICC se bavi kompromisom između dobrog uklapanja modela i jednostavnosti modela (funkcije maksimalne verodostojnosti - najmanjih kvadrata i broja parametara, te obima serije n). SBC kriterijum bolje opisuje navedeni gubitak informacija. On zavisi od proizvoda broja parametara i $\ln(n)$, kada je n mnogo veće od broja parametara. Oba kriterijuma zasnovana su na primeni odgovarajućih formula.

U XLSTAT-u postoje sve navedene opcije za razmatranje SARIMA modela, za $S=0$, $S\neq 0$, automatski i vrlo brzo, uz jednostavni i intuitivni interfejs, do praktično dovoljno velikih vrednosti p, P, q, Q, d, D. Pored vremenskog perioda za validaciju, moguće je menjati i vreme predikcije (npr. koristi se promenljiv interval poverenja za buduće događaje). Pored SBC i AICC dostupno je još 12 kriterijuma za analizu podataka [3], [5]. Po AICC i SBC se vrši odabir najboljih SARIMA modela.

2.2. Osnove ELM modela

Huang i autori su 2004 predložili algoritam „mašine za ekstremno učenje (ELM)" [6-10]. Algoritam predstavlja poseban slučaj ANN-FNN sa jednim skrivenim slojem, takozvanom *Single Hidden Layer Feedforward Neural Network* (SLFN). Na Slici 1 dat je prikaz karakteristične, jednostavne SLFN.



Slika 1. Šema SLFN.

Na Slici 1 prikazan je vektor ulaza $\boldsymbol{x} = [x_1, x_2, \dots, x_n]^T \in R^n$, vektor biasa $\boldsymbol{b} = [b_1, b_2, \dots, b_L]^T \in R^L$, vektor izlaza $\boldsymbol{o} = [o_1, o_2, \dots, o_m]^T \in R^m$, kao i vektori ulaznih i izlaznih težina $\boldsymbol{w}$ i $\boldsymbol{\beta}$.

Algoritam ELM modela opisuje se na sledeći način: razmatra se samo jedan uzorak - vremenska serija, ulaznih veličina x_i, $i \in \{1,2,\dots,N\}$. Pored njih, značajni su vektori targeta (ciljeva) t_i i izlaza o_i. ELM je SLFN sa skrivenim slojem koji se, u okviru ovog razmatranja, sastoji od N' skrivenih čvorova-neurona. Standardna aktivaciona funkcija ovde se pojavljuje u jednom od tri karakteristična oblika (prvi oblik - sigmoidna funkcija, obično daje najbolje rezultate),

$$\begin{aligned} g(x) &= \frac{1}{1+\exp(-x)}; \\ g(x) &= \sin(x); \\ g(x) &= \mathit{hard}\lim(x). \end{aligned} \tag{2}$$

Karakterističan za ELM je vektor targeta $\boldsymbol{T} = [t_1, t_2, \dots, t_N]^T \in R^N$. U okviru modela važi, kada $N \to \infty$,

$$\sum_{i=1}^{N} \|o_i - t_i\| \to 0. \tag{3}$$

Osnovna jednačina ELM modela je

$$\sum_{i=1}^{N'} \beta_i g_i(x_j) = \sum_{i=1}^{N'} \beta_i g(w_i \cdot x_j + b_i) = t_j . \tag{4}$$

Ovde su vektor ulaznih težina $\boldsymbol{w} = [w_1, w_2, \dots, w_{N'}]^T$ i vektor izlaznih težina $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_{N'}]^T$.
U matričnom obliku

$$\boldsymbol{H\beta} = \boldsymbol{T}. \tag{5}$$

Ekcplicitno zapisana matrica $\boldsymbol{H}$ je

$$\mathbf{H} = \begin{bmatrix} g\left(w_1 \cdot x_1 + b_1\right) & \ldots & g\left(w_{N'} \cdot x_1 + b_{N'}\right) \\ \ldots & & \ldots \\ g\left(w_1 \cdot x_N + b_1\right) & \ldots & g\left(w_{N'} \cdot x_N + b_{N'}\right) \end{bmatrix}_{N \times N'}. \quad (6)$$

Obično se slučajno generišu w_i, b_i za date ulaze x_i (trening skup) i date targete, ovde izlaze, t_i. Preko pseudoinverza, prvo se dobijaju vrednosti β_i, $\boldsymbol{\beta}$. Onda se za neki novi test skup x'_i izračunaju novi test izlazi - validacija od t'_i koji se upoređuje sa stvarnim testnim targetom t'_i. Obično se umesto targeta koriste stvarni podaci x. To znači, u jednom slučaju, da se originalan ukupan skup x-ova - ulaza podeli na dva skupa jedan je trening, a drugi je test skup, te su, dakle, ulazi istovremeno i targeti. Razlog za slučajna generisanja datih parametra je u tome da se umesto pretrage celog prostora parametra ($\boldsymbol{w}$,$\boldsymbol{b}$,$\boldsymbol{\beta}$), drastično smanji i pojednostavi navedena procedura. Ispostavlja se da RMSE i/ili R^2 mogu imati, u pojedinim slučajevima prilikom validacije, vrlo loše vrednosti zbog slučajnog generisanja ($\boldsymbol{w}$,$\boldsymbol{b}$), što se kompenzuje izvršenjem velikog broja ciklusa primene ELM. ELM je u ovom radu primenjen u programskom paketu Matlab.

2.2. Osnove RF bagging modela

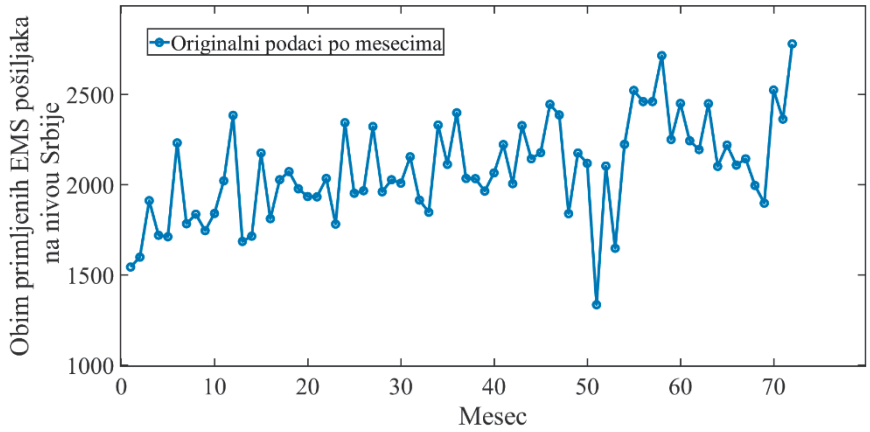
U ovom modelu slučajno se generišu različiti skupovi za obuku uzorkovanjem uz, eventualno, ponavljajuću zamenu iz originalnog ulaznog skupa $\boldsymbol{x}$, [13]. Svaki od tih izbora bi mogao biti primer ulaza. Zatim se pokreće mašinsko učenje za jednu zamenu, ovde prvu hipotezu $h_1(\boldsymbol{x})$ (obično neku funkciju raspodele adekvatno opisanu) uz $\boldsymbol{N}$ takvih računski generisanih vrednosti izbora ulaza. Postupak se ponavlja sve dok se ne iscrpi skup hipoteza, čiji je ukupan broj obeležen sa $\boldsymbol{K}$, koji čine stablo odluka. U problemu regresije izlaz je funkcija raspodele

$$h(\mathbf{x}) = \frac{1}{\mathbf{K}} \sum_{i=1}^{\mathbf{K}} h_i(\mathbf{x}). \quad (7)$$

Uzorci, dakle, čine stablo odluka, a više ovih stabala šumu. Slučajne šume – RF, kao što im ime sugeriše, na slučajan način generiše stabla / šumu. Koreni novih stabala, sa odgovarajućim svojim prvim hipotezama dalje se mogu paralelno procesirati (situacija je idealna za višeprocesorski sistem ili više računara), obrazujući ovakve šume. U praksi, mogu biti različiti tipovi šuma, na primer, centrirane, uniformne. Cilj RF algoritma je da se smanje korelacije među stablima odluka i time smanji ulazna varijansa. Mogu se validirati svi elementi ulaza. Stvarna predikcija se ostvaruje generisanjem novih elemenata pomoću ukupne funkcije raspodele. RF je primenjen ovom radu u okviru Excel dodatka XLSTAT.

3. Numerički rezultati

Modeli opisani u prethodnoj sekciji primenjeni su za simulaciju, validiranje ili predviđanje podataka vremenske serija obima 72 – obima mesečno primljenih EMS pošiljaka na nivou Srbije u intervalu od 2016. do 2021. godine. Relevantne karakteristike korišćene računarske konfiguracije su: Intel(R) Core(TM) i5-3570 CPU @ 3.40GHz, 3.40 GHz pri instaliranoj RAM memoriji od 16.0 GB. Originalni podaci koji se odnose na obim primljenih EMS pošiljaka su predstavljeni na Slici 2.



Slika 2. Mesečni obim primljenih EMS pošiljaka u Srbiji.

Primećuje se da obim primljenih EMS pošiljaka na nivou Srbije varira od preko 1000 komada mesečno do preko 2500 komada mesečno, te da postoji trend blagog rasta (nije konstantna veličina, karakteristična za SARIMA metod).

3.1. SARIMA model

Prikazan je deo uobičajene procedure u XLSTAT-u, koja može biti i kompleksnija, u zavisnosti od slučaja, [14]. Standardne setovane vrednosti parametara u okviru ovog okruženja su: convergence value je 10^{-5}; maximum iterations 500000; S=0, 12; $p \leq 4$; $q \leq 3$; d=0,1; $P \leq 2$; D=0,1; $Q \leq 2$; confidence intervals je 90%. Razmatrana je optimizacija i po AICC i po SBC kriterijumu, za navedene različite vrednosti parametara. Optimalan model ima vrednosti parametrara S=12, p=4, q=3, P=2, Q=2, d=D=1. AICC=665.957, broj iteracija 682, SBC=680.954, broj iteracija 354. Razlike u vrednostima parametara opisanih jednačinom (1) su, za navedene kriterijume, velike. Za SBC su dati neki od rezultata, bez grešaka, ali sa vrednostima gornjih i donjih granica, prikazani u Tabelama I i II, u oznakama XLSTAT okruženja.

TABELA I
TREND COMPONENT, μ

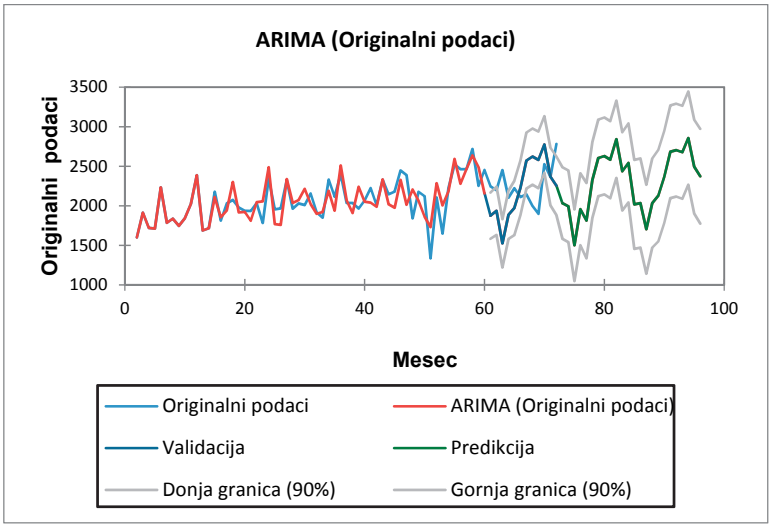
Parameter	Value	Lower bound (90%)	Upper bound (90%)
Constant μ	-0.502	-1.356	0.353

TABELA II
VALUES OF SARIMA MODEL

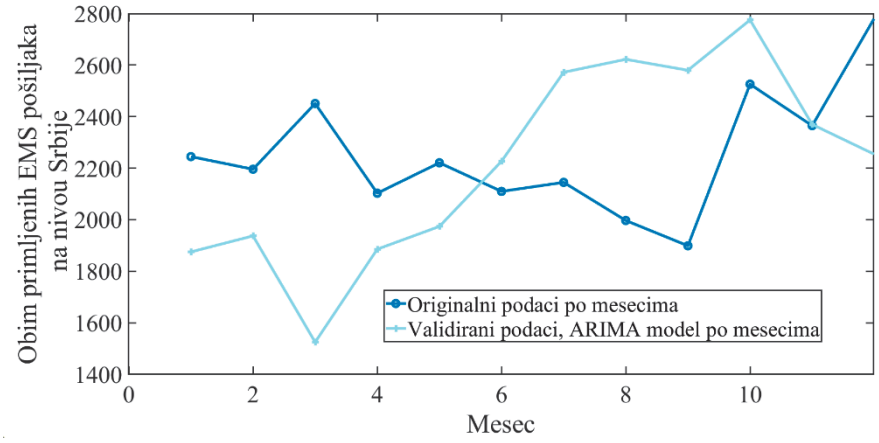
Parameter	Value	Lower bound (90%)	Upper bound (90%)
AR(1)	0.440	0.135	0.746
AR(2)	0.399	-0.206	1.004
AR(3)	0.023	-0.563	0.609
AR(4)	-0.535		
SAR(1)	0.000	-0.065	0.065
SAR(2)	0.000	-0.267	0.267
MA(1)	-1.747	-1.911	-1.583
MA(2)	0.531	-0.269	1.331
MA(3)	0.228	-0.459	0.915
SMA(1)	0.000	-1.007	1.007
SMA(2)	0.000	-0.565	0.565

Na Slici 3 prikazani su rezultati simulacije (validacije i predikcije) nad podacima pomoću XLSTAT-a. Pomoću XLSTAT-a je moguće dobiti i ostale dijagrame, na primer, reziduala, PACF, ACF, propratnih kalkulacija itd. Vrednosti vremena simulacije eksponencijalno rastu sa povećanjem vrednosti parametra *p*, a slično važi i za druge parametre, što je dobijeno u procesu procene vremena neophodnog za obradu podataka.

Na Slici 4 prikazano je uporedo poslednjih 12 oginalnih podataka o obimu EMS pošiljaka i validiranih SARIMA podataka. Uočljivo je izrazito neslaganje originalnih i validacionih podataka. Vrednosti kvaliteta fita su RMSE=462.0147, odnosno R^2=0.0197. Na osnovu obe vrednosti parametara, očigledno je da je ova validacija loša za ovakvu vrstu problema. Samim tim, i predikcija narednih 24 vrednosti nema veći značaj. Interval poverenja od 90% je dovoljno širok da bi se eventualne vrednosti predikcije mogle naći unutar njega. Ovaj interval se značajno širi sa porastom vremena - postupak je divergentan. Vreme izvršenja programa je oko 10 minuta.



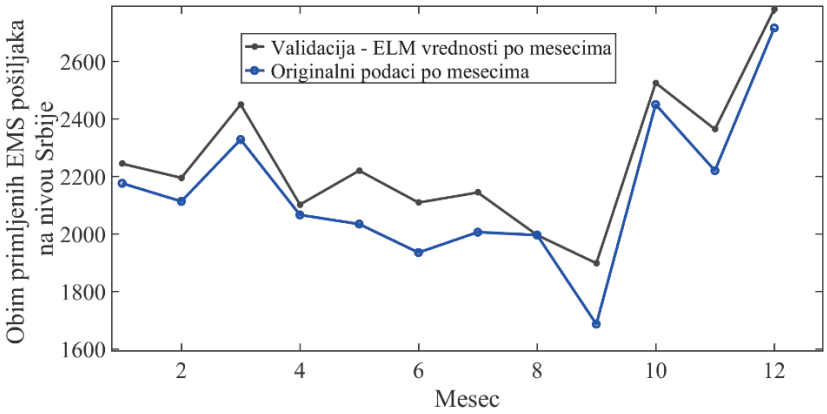
Slika 3. Prikaz podataka i rezultata dobijenih pomoću SARIMA metoda u okruženju XLSTAT.



Slika 4. Poređenje poslednjih 12 originalanih podataka sa (S)ARIMA modelom.

3.2. ELM model

Razvijeni ELM model primenjen je na sledeći način: ukupan broj podataka 72 deli se na dva podskupa. Prvi, trening podskup sadrži prvih 60 elemenata, dok drugi, test skup, preostalih 12. Pošto ELM nema predikcije, postoji samo validacija na ovih 12 elemenata. Broj neurona je N'=60. Taj broj treba da bude, u principu, uporediv sa brojem podataka, na osnovu dosadašnjeg eksperimentisanja na vremenskim serijama sa brojem članova oko stotinu. Rezultati najbolje simulacije u Matlabu su prikazani na Slici 5.



Slika 5. Poređenje poslednjih 12 originalanih podataka sa ELM modelom

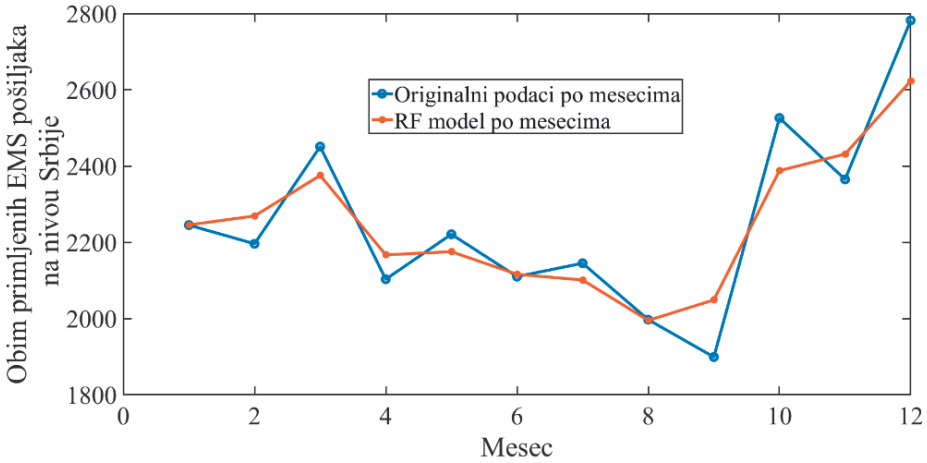
Usled slučajno biranih vrednosti za $\boldsymbol{b}$, $\boldsymbol{w}$, vrednosti za RMSE i R^2 mogu biti vrlo različite. Da bi se one poboljšale, neophodno je naći odgovarajući kriterijum za to. Odabrane su na osnovu eksperimenata i proba, samo one koje imaju ove vrednosti RMSE ispod 125 i R^2 većih od 0.75. Broj ciklusa određen je takođe eksperimentalno; on je jednak 500000. Vrednosti karakterističnih parametara odabrane najbolje simulacije su: RMSE=124.7242 i R^2=0.94367. Dobijena vrednost RMSE je manja za 72%, dok je vrednost R^2 približno 48 puta veća u odnosu na odgovarajuće vrednosti kod SARIMA modela. Vreme izvršenja programa u ovom slučaju je oko 2.5h.

3.2. RF model

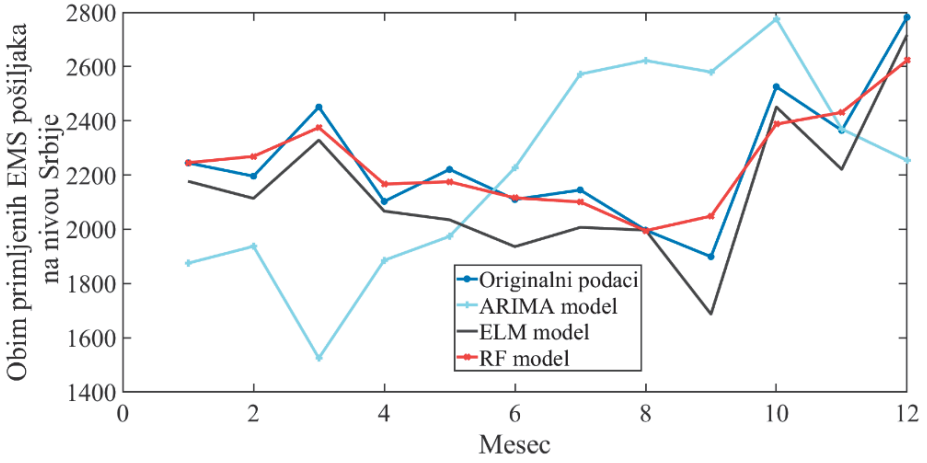
Algoritam RF regresije, u okruženju XLSTAT, u svojoj najjednostavnijoj, bagging varijanti, prisutan je preko svog demoa, [15]. Ako se za datih 72 podataka izvrši navedena RF simulacija, dobiće se novih 72 podataka koji po samoj strukturi algoritma imaju manju varijansu. Broj stabala u RF je 100, on se može menjati. Iako su validirane sve vrednosti, od interesa je validacija poslednjih 12 podataka. Dobijeni rezultati prikazani su na Slici 6. Vrednosti karakterističnih parametara odabrane simulacije su: RMSE=86.7604 i R^2=0.9060. Trajanje procesa rada XLSTAT-a, u ovom slučaju, je desetak sekundi.

Na Slici 7 uporedo su prikazani ukupni rezultati simulacije dobijeni pomoću svih primenjenih modela i originalni podaci koji se odnose na obim primljenih EMS pošiljaka za poslednjih 12 meseci razmatranog perioda. Za datu malu vremensku seriju, direktno programiranje u Matlabu se pokazalo sporijim sredstvom od XLSTAT Excel dodatka. Treba napomenuti da je korišćenje SARIMA u Matlabu (Econometrics Toolbox), daleko sporije od XLSTAT-a [5]. Slično treba očekivati i za RF. ELM, kao relativno nova tehnika, nije izdvojena, kao ostale dve, u okviru ovih okruženja. XLSTAT ima, u principu, bolje organizovano okruženje od Matlaba, sa gotovim optimalizovanim alatima, ali, uz male mogućnosti za programiranje. To se vidi i po vremenima izvršenja algoritama. Skup alata za ispitivanje vremenskih serija mu je ograničen, kao i veličine skupova podataka na koje se može primeniti (big data podaci su mu nedostižni). RF se slično ponaša kao standardni

Savitzky-Golay filter [4], koji je brz, ali je RF u principu, u odnosu na ovaj filter mnogo efikasniji (otežinjeno polinomno usrednjavanje filtera ima tačnu formulu preko koje on radi, dok se pomoću RF formula slučajno modelira, te sa povećanjem nezavisnih stabala odlučivanja može biti dovoljno komplikovanija). U radu [5] korišćen je ANN LSTM (*long short-term memory network*) algoritam. ELM je za praktično isto vreme izvršavanja imao oko 30% manju RMSE od LSTM, slično kao ovde RF u odnosu na ELM. Skraćivanje vremena izvršavanja Matlab programa, pored softverskih inovacija, moguće je ostvariti i u cloud-u [16]. Pored navedenog, za dati problem, od interesa su i drugi jezici i okruženja (Python, C++, Julia, Mathematica, Maple, itd).



Slika 6. Poređenje poslednjih 12 originalanih podataka sa RF modelom.



Slika 7. Poređenje poslednjih 12 originalanih podataka sa predloženim modelima.

4. Zaključak

Istraživanje vremenskih serija obima mesečno primljenih EMS pošiljaka na nivou Srbije u periodu od 2016. do 2021. godine, u okviru procesa validacije u programskim okruženjima XLSTAT-Excel i Matlab, pokazalo je izrazitu prednost RF modela u odnosu na ELM i SARIMA model. Pokazano je da RF model ima za oko 30% manju vrednost RMSE od ELM-a i oko 70% manju od SARIMA metoda, uz mnogo veći R^2. Pored toga, korišćenjem RF modela ostvareno je i najkraće vreme izvršavanja programa.

Literatura

[1]https://www.posta.rs/eng/stanovnistvo/usluga.aspx?usluga=postal-services/express-services-international/ems-express-mail-service .

[2] G.E. Box, G.M. Jenkins, G.C. Reinsel, G. M. Ljung, *Time Series Analysis, Forecasting and Control*. New Jersey, John Wiley and Sons, 2016.

[3] https://www.xlstat.com/en/.

[4] I. D. Rogan and O. R. Pronić-Rančić, "Forecasting the volume of postal services using Savitzky-Golay filter modification," 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), 2021, pp. 123-126, doi: 10.1109/ICEST52640.2021.9483459.

[5] I. D. Rogan and O. R. Pronić-Rančić, "SARIMA and ANN Approaches in Forecasting the Volume of Postal Services," 15th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS), 2021, pp. 193-196, doi: 10.1109/TELSIKS52058.2021.9606429.

[6] Ivana D. Rogan, Olivera Pronić-Rančić, " Combined techniques for forecasting the volume of packages in internal postal traffic of Serbia", Facta Universitatis, Series: Automatic Control and Robotics, 2022, in press.

[7] Jorge Garza-Ulloa, *Applied Biomedical Engineering Using Artificial Intelligence and Cognitive Models*. Academic Press, 2021.

[8] Guang-Bin Huang, Qin-Yu Zhu, Chee-Kheong Siew, "Extreme learning machine: theory and applications," Neurocomputing. 70 (1): 489–501, 2006.

[9] Guang-Bin Huang, Qin-Yu Zhu, K. Z. Mao, Chee-Kheong Siew, P. Saratchandran and N. Sundararajan, "Can threshold networks be trained directly?," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 53, no. 3, pp. 187-191, March 2006, doi: 10.1109/TCSII.2005.857540.

[10] G. -B. Huang, H. Zhou, X. Ding and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 2, pp. 513-529, April 2012, doi: 10.1109/TSMCB.2011.2168604.

[11] Guang-Bin Huang, "What are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle," Cognitive Computation, volume 7, issue 3 2015.

[12] J. Wang, S. Lu, Shui-Hua Wang, Yu-Dong Zhang, " A review on extreme learning machine," Multimed Tools Appl, 2021. [Online]. Available: https://doi.org/10.1007/s11042-021-11007-7 .

[13] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*. Pearson; 4th edition, 2020.
[14] https://help.xlstat.com/6753-fit-arima-model-time-series-excel
[15] https://help.xlstat.com/6551-random-forest-regression-excel-tutorial
[16] https://www.mathworks.com/videos/how-to-run-matlab-in-the-cloud-with-amazon-web-services-1542634996553.html

Abstract: The paper presents the results of validation/prediction of the time series data of the volume of monthly received EMS (Express Mail Service) items at the level of Serbia in the interval 2016 - 2021. The following models were used: a model of seasonal autoregressive integrated moving averages (SARIMA), Random forests (RF) in the XLSTAT add-on for Excel and within the Matlab environment - Extreme Machine Learning (ELM). The validation process demonstrated a distinct advantage of the RF model over the ELM and SARIMA model. It is shown that the RF model has about 30% lower RMSE than ELM and about 70% lower than the SARIMA method, with a much higher R2. In addition, the shortest program execution time was achieved by using the RF model.

Keywords: *time series analysis, EMS, SARIMA, ELM, RF.*

FORECASTING THE VOLUME OF RECEIVED EMS ITEMS IN SERBIA USING SARIMA, RANDOM FORESTS AND ELM MODELS

Ivana D. Rogan, Olivera R. Pronić-Rančić