

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij matematike
Financijska matematika i statistika

Matej Petrinović

Analiza vremenskih nizova

Seminarski rad

Modeliranje prijedenh putničkih zračnih milja

Voditelj: doc. dr. sc. Danijel Grahovac

Osijek, 2019.

Sadržaj

1	Uvod	1
2	Osnovna statistika varijable	2
3	Milje kroz vrijeme	3
4	Modeliranje	5
4.1	Model 1	5
4.2	Model 2	7
4.3	Model 3	7
5	Predviđanje	9
6	Zaključak	10

1 Uvod

U ovom seminarskom radu obrađena je baza podataka Air Revenue Passenger Miles. Baza podataka sadrži podatke o broju plaćenih zračnih milja korištenih za putovanja zrakoplovom. Revenue passenger miles (RPM) je prijevozna industrijska mjera koja pokazuje broj plaćenih milja koje putnici prijeđu i to je osnovna statistika za zrakoplovni prijevoz. RPM se računa tako da se pomnoži broj putnika koji su platili putnu kartu sa brojem milja koje je zrakoplov prevalio. Na primjer, ako zrakoplov ima 100 putnika i pri tome je prevalio 250 milja njegov RPM će iznositi $RPM = 25\,000$. Baza podataka sadrži podatke i RPM u intervalu od 1.1.2002 do 2.1.2019. godine, te su podaci dani mjesečno u tisućama. Bazu podataka se može preuzeti na <https://fred.stlouisfed.org/series/AIRRPMTSI>. Podatke je prikupio U.S. Department of Transportation, Bureau of Transportation Statistics (BTS).

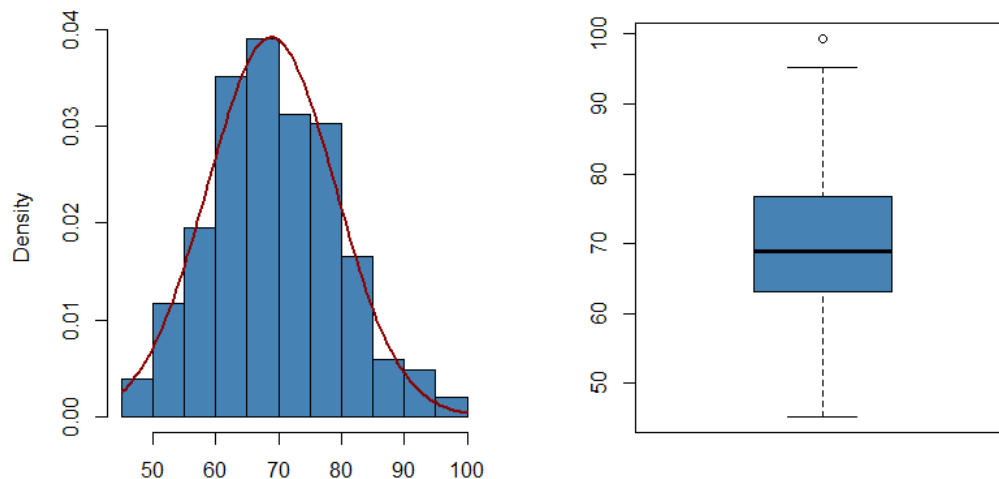
Cilj seminarskog rada je modelirati te prijedene milje nekim slučajnim procesom $\{X_t, t \in \mathbb{Z}\}$. Kako milje promatramo u vremenskom intervalu taj proces ćemo modelirati nekim $(S)ARIMA$ procesom. Prilikom modeliranja izostavit ćemo nekoliko podataka, radi usporedbe u predviđanju.

2 Osnovna statistika varijable

Na početku pogledajmo neku elementarnu statistiku varijable koju modeliramo, tj. milje. Promotrimo tablicu s osnovnim podacima i pripadne grafičke prikaze. Podaci su izraženi u milijunima.

Min	1st Qu.	Medijan	Prosjek	3rd Qu.	Max	SD
45.16	63.02	68.85	69.43	76.68	99.37	10.2

Vidimo iz tablice kako je prosjek prijeđenih milja oko 7 milijuna, pa bi mogli reći kako se mnogo putovalo. Podaci jesu veliki, ali to ni ne čudi jer su izvorno iz američkih zračnih luka. Nadalje promotrimo histogram i kutijasti dijagram.

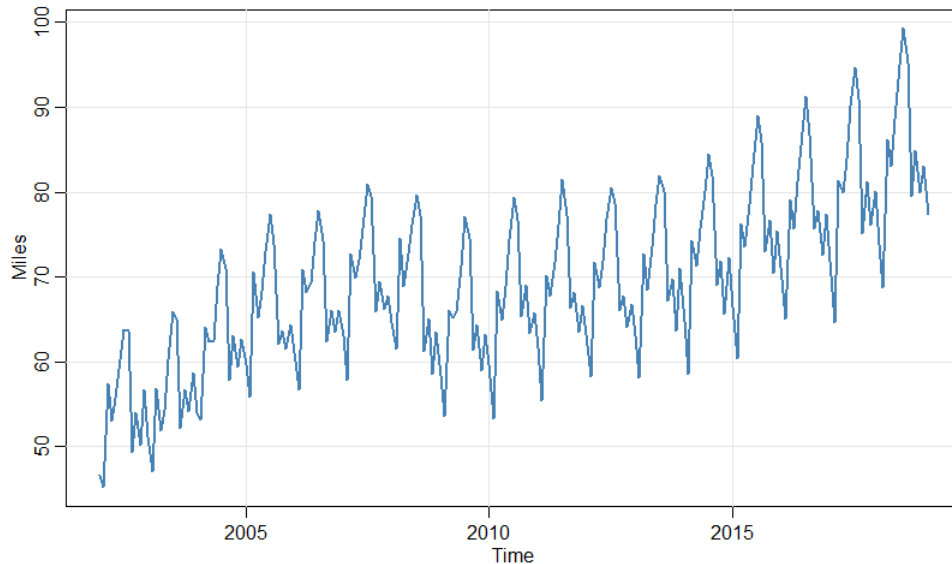


Slika 1: Histogram i kutijasti dijagram

Iz histograma možemo vidjeti kako se najviše prijeđenih milja nalazi između 6.5 i 7 milijuna. Histogram sam po sebi sugerira da se prijeđene milje mogu modelirati normalnom distribucijom. Provođenjem Shapiro-Wilk testa i KS testa dobivamo p -vrijednosti redom 0.628 i 0.7459, pa samim time na razini značajnosti $\alpha = 0.05$ nema razloga sumnjati u istinitost nulte hipoteze. Vidimo kako kutijasti dijagram sadrži jednu stršeću vrijednost. Ta stršeća vrijednost je na dan 7.1.2018, što znači da se u lipnju 2018. godine najviše putovalo.

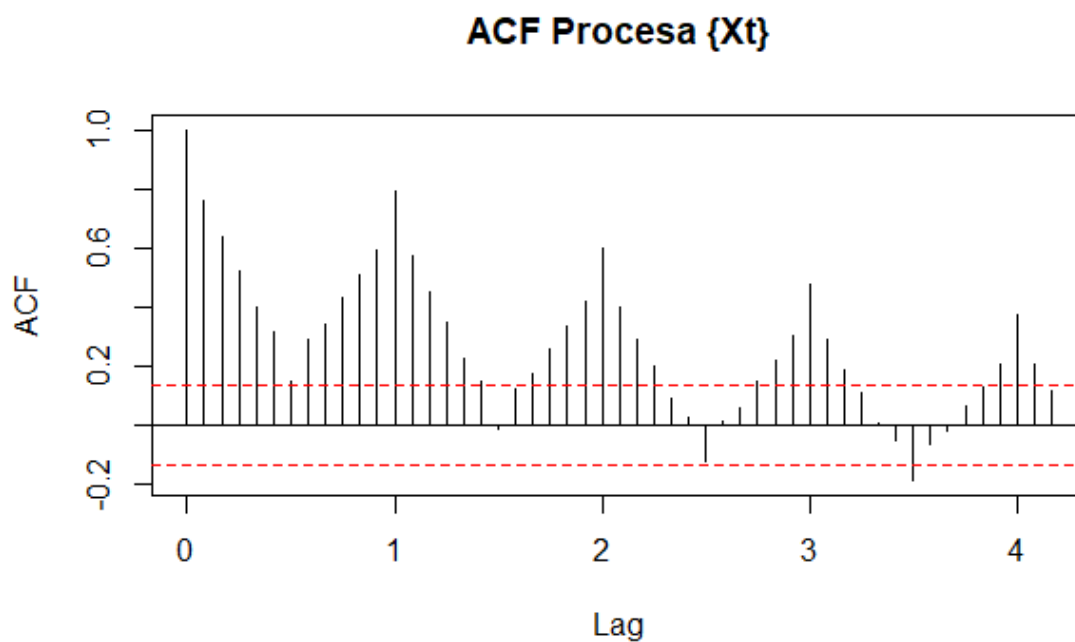
3 Milje kroz vrijeme

U ovom poglavlju promatrat ćemo kretanje broja prijeđenih milja kroz vrijeme i pokušati dobiti uvid u to kakvim modelom ćemo modelirati naš proces $\{X_t\}$. Prvo promotrimo prikaz kretanja kroz vrijeme.



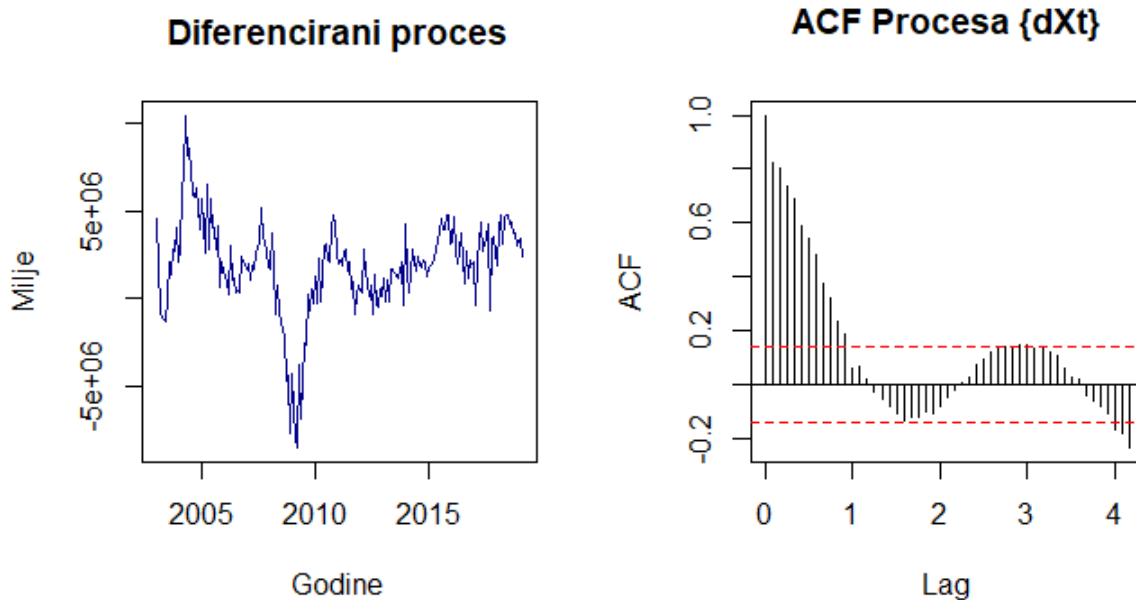
Slika 2: Milje kroz vrijeme

Na prikazu svakako prvo možemo uočiti nekakav rastući trend, tj. kroz vrijeme se broj prijeđenih milja povećava. Također, druga stvar koja se odmah može uočiti jest periodičnost. Tako se najviše putuje u lipnju, a najmanje u siječnju. Ovakav izgled procesa sugerira na nestacionarnost. Pogledajmo funkciju autokorelacija tog procesa.

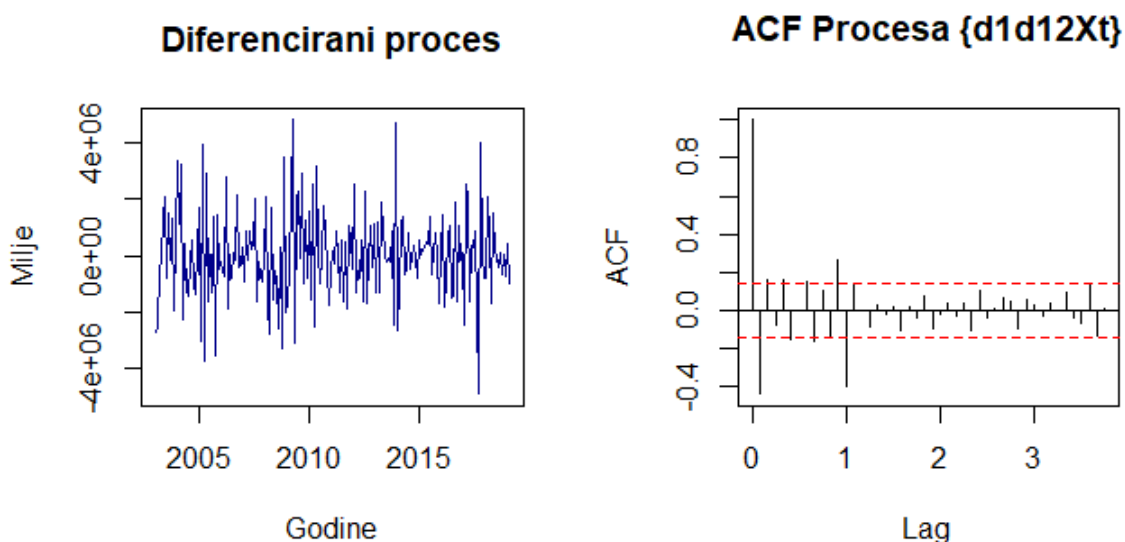


Slika 3: Autokorelacijska funkcija procesa

Iz samog izgleda funkcije autokorelacija procesa $\{X_t\}$ možemo vidjeti nestacionarnost. Isto tako, ukoliko pogledamo korelaciju na koracima 1,2,..., vidimo kako ona eksponencijalno opada u 0, što nam u ovom trenutku sugerira da ćemo proces modelirati nekim *SARIMA* modelom. Kako *adf-test* nije prikladan za sezonalne modele, da bi dosli do stacionarnog procesa diferencirajmo ga prvo na koraku $s = 12$ i pogledajmo izgled proces i njegovu funkciju autokorelacija *ACF*.



Ako sada pogledamo izgled procesa, i dalje vidimo na nekim dijelovima trendove rasta i pada dok, ako pogledamo *ACF*, vidjet ćemo kako se diferenciranjem na koraku $s = 12$ riješi sezonalnosti, ali i dalje postoje nezanemarive korelacije. Stoga ćemo dalje taj proces još jednom diferencirati na koraku 1. Pogledajmo što se dogodi nakon izvršenja te transformacije.



Iz ovih grafičkih prikaza, ako pogledamo sam izgled procesa $\{Y_t\} = \{\Delta\Delta_{12}X_t\}$ on izgleda stacionarno. Vidimo da on nekako "oscilira" oko 0. Ako pogledamo izgled *ACF* vidimo kako smo se riješili većine korelacija, ali ipak postoje nezanemarive korelacije na koracima 1,11 i 12 koje ćemo na neki način uključiti u model. Sve u svemu smatrat ćemo da je proces $\{Y_t\}$ stacionaran i za njega ćemo tražiti model u klasi $SARMA(p, q) \times (P, Q)_{12}$. Početnim diferenciranjem za $\{X_t\}$ smo fiksirali $d = 1$ i $D = 1$.

4 Modeliranje

Kako smo ranije rekli zbog sezonalnosti pojave koju promatramo, proces $\{X_t\}$ modelirati ćemo nekim *SARIMA* modelom, tj. $\{X_t\} \sim \text{SARIMA}(p, d, q) \times (P, D, Q)_{12}$, pri čemu je $d = 1$ i $D = 1$. Model ćemo raditi na dijelu podataka, nećemo uzeti sve, kako bi mogli napraviti kasnije predikciju i usporedbu sa pravim podacima. Dakle, radit ćemo na podacima od 1.1.2002 do 1.12.2017. godine, što nam ostavlja 14 mjeseci za predviđanje i usporedbu. Prije traženje modela i procjene parametara treba i odrediti red modela p, q, P, Q . Redove ćemo odrediti na osnovu informacijskih kriterija *AIC* i *BIC*, te ćemo redove ograničiti $p, q, P, Q \leq 2$. Nakon provedene procedure u *R*-u dobijemo iduće redove (uzimamo prvih 5 najboljih):

<i>AIC</i>				<i>BIC</i>			
p	q	P	Q	p	q	P	Q
1	0	0	1	1	0	0	1
1	0	0	2	0	1	0	1
1	0	1	1	1	0	0	2
1	1	0	1	1	0	1	1
2	0	0	1	1	1	0	1

Vidimo kako *AIC* i *BIC* za najmanje vrijednosti daju iste redove, tj. $p = 1, Q = 1$. Općenito, naš proces $\{X_t\}$ imat će sljedeći oblik

$$\phi(B)\Phi(B^{12})(1-B)^1(1-B^{12})^1X_t = \theta(B)\Theta(B^{12})Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2)$$

Prilikom modeliranja isprobat ćemo više modela i u konačnici odabrati onaj koji nam se čini najbolji. Radi jednostavnosti umjesto diferenciranog procesa $\{X_t\}$ raditi ćemo sa $\{Y_t\}$, kojeg smo definirali kao $\{Y_t\} = \{\Delta\Delta_{12}X_t\}$.

4.1 Model 1

Neka je $\{Y_t\} \sim \text{SARMA}(1, 0) \times (0, 1)_{12}$. Proces $\{Y_t\}$ tada zadovoljava jednadžbu modela:

$$Y_t = \phi Y_{t-1} + Z_t + \vartheta Z_{t-12}$$

pri čemu je $\{Z_t\} \sim WN(0, \sigma^2)$. Parametre ϕ i ϑ treba procijeniti, kao i varijancu bijelog šuma. Kada procedurama u *R*-u procijenimo parametre dobijemo sljedeći model, tj. jednadžbu za Y_t :

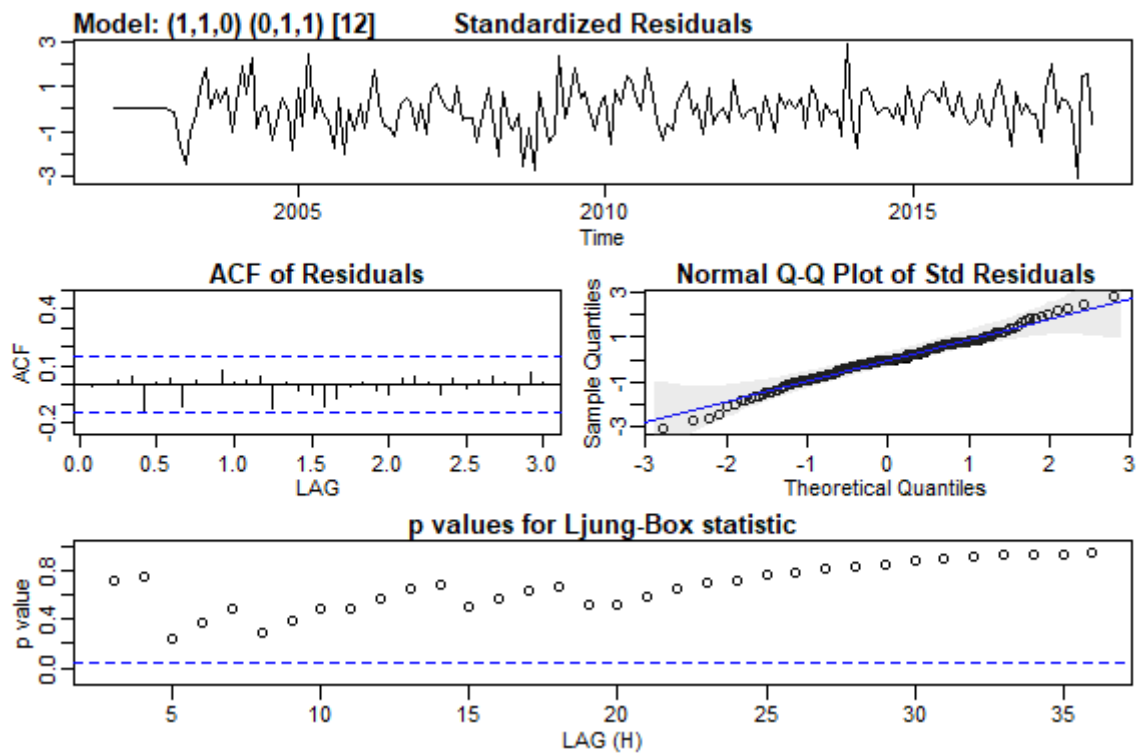
$$Y_t = -0.41Y_{t-1} + Z_t - 0.59Z_{t-12}$$

Također, procjenjena varijanca bijelog šuma iznosi $\sigma^2 = 1.659e + 12$. Pogledajmo dalje značajnost koeficijenata i analizu reziduala.

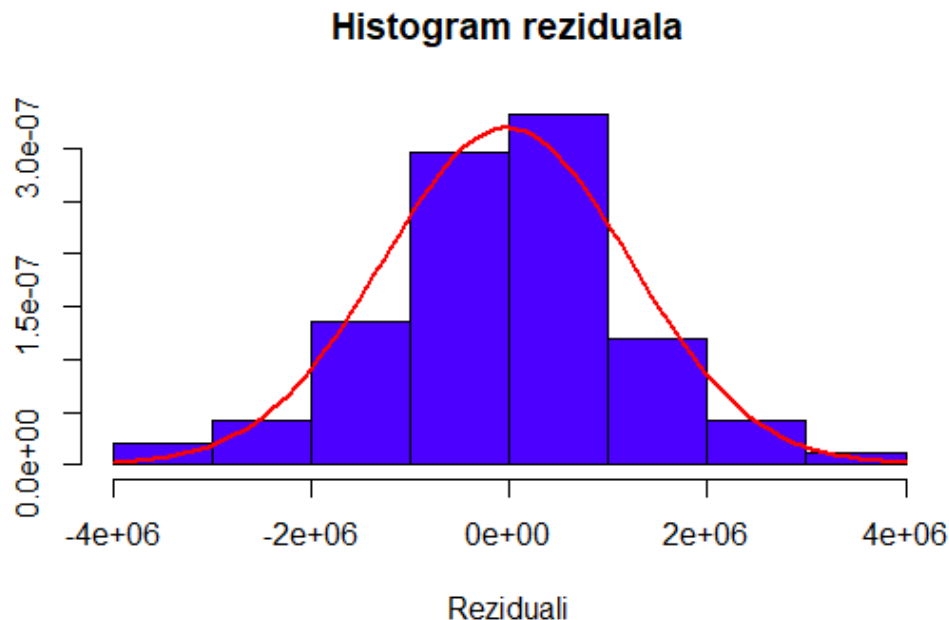
Značajnost koeficijenata provjeriti ćemo tako da pogledamo pouzdane intervale. 95% interval za procjene koeficijente su:

Koef.	Pouzd. int
$\phi = -0.41$	$[-0.5482981, -0.2781085]$
$\vartheta = -0.59$	$[-0.7226117 - 0.4549723]$

Vidimo kako ni jedan pouzdani interval ne sadrži 0, stoga svi koeficijenti su nam statistički značajni. Nadalje, napravimo analizu reziduala.



Iz grafičkih prikaza reziduala možemo vidjeti kako oni dobro stoje što se tiče nekoreliranosti (ako pogledamo p -vrijednosti Ljung-Box testa, sve su iznad 0.05), što je zapravo i najvažnije svojstvo. Iz qqPlot-a možemo možda uvidjeti normalnost istih. Pogledajmo histogram i testirajmo to pripadnim testovima.



Vidmo kao histogram i sam ukazuje na normalnost. Testiranjem normalnosti koristeći Shapiro-Wilk test i Jarque-Berra test dobiti ćemo p -vrijednosti redom 0.123 i 0.1004, stoga na razini značajnosti $\alpha = 0.05$ nema razloga sumnjati u normalnost reziduala.

Vidjeli smo kako ovaj model ima sve parametre statistički značajne, reziduali su nekorelirani i mogli bi reći normalno distribuirani, te bi mogli ovaj model dalje uzeti u razmatranje.

4.2 Model 2

Neka je $\{Y_t\} \sim SARMA(1, 0) \times (0, 2)_{12}$. Proces $\{Y_t\}$ tada zadovoljava jednadžbu modela:

$$Y_t = \phi Y_{t-1} + Z_t + \vartheta_1 Z_{t-12} + \vartheta_2 Z_{t-24}$$

pri čemu je $\{Z_t\} \sim WN(0, \sigma^2)$. Parametre ϕ , ϑ_1 i ϑ_2 treba procijeniti, kao i varijancu bijelog šuma. Kada procedurama u R -u procijenimo parametre dobijemo sljedeći model, tj. jednadžbu za Y_t :

$$Y_t = -0.41Y_{t-1} + Z_t - 0.57Z_{t-12} - 0.03Z_{t-24}$$

Također, procjenjena varijanca bijelog šuma iznosi $\sigma^2 = 1.658e + 12$. Uočimo kako su parametri skoro pa isti kao u početnom modelu. Pogledajmo značajnosti koeficijenata pouzdanim intervalima.

Koef.	Pouzd. int
$\phi = -0.41$	$[-0.5481342 - 0.2777172]$
$\vartheta_1 = -0.57$	$[-0.7395177 - 0.4067280]$
$\vartheta_2 = -0.03$	$[-0.1914653, 0.1406375]$

Iz tablice vidimo kako koeficijent ϑ_2 nije statistički značajan, jer pouzdani interval sadrži 0. Njegovima uklanjanjem bi dobili model (1). Napomenimo samo da je i ovako odabran model dobar, jer su i njegovi reziduali nekorelirani i sugeriraju na normalnu distribuciju.

4.3 Model 3

Neka je $\{Y_t\} \sim SARMA(0, 1) \times (0, 1)_{12}$. Proces $\{Y_t\}$ tada zadovoljava jednadžbu modela:

$$Y_t = Z_t + \theta Z_{t-1} + \vartheta Z_{t-12}$$

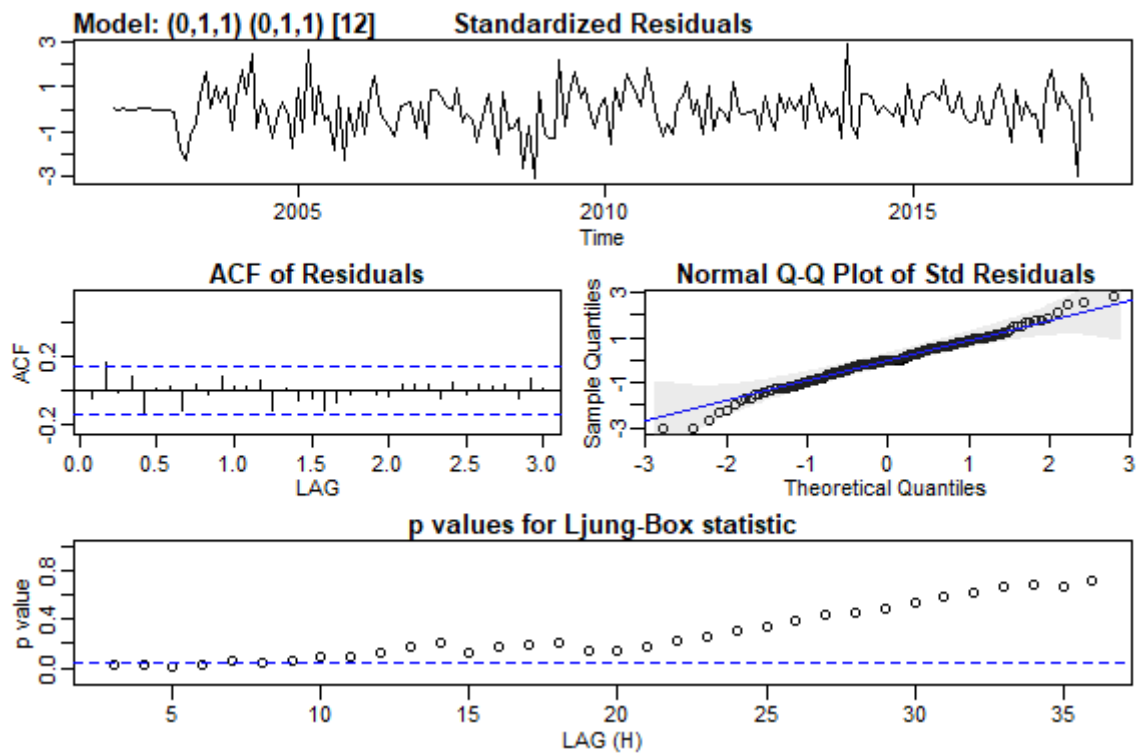
pri čemu je $\{Z_t\} \sim WN(0, \sigma^2)$. Parametre ϕ , θ i ϑ treba procijeniti, kao i varijancu bijelog šuma. Kada procedurama u R -u procijenimo parametre dobijemo sljedeći model, tj. jednadžbu za Y_t :

$$Y_t = Z_t - 0.37Z_{t-1} - 0.6Z_{t-12}$$

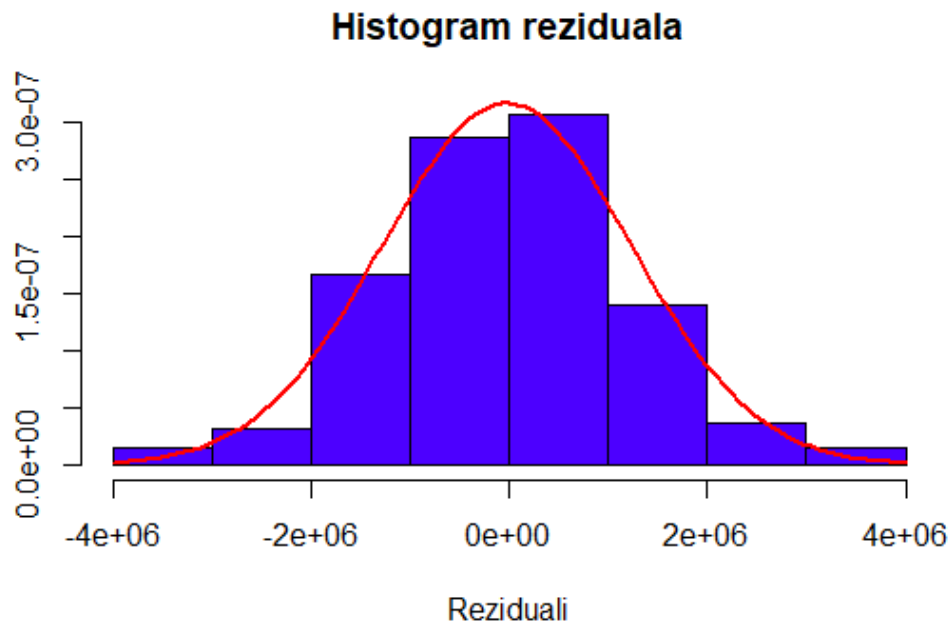
Također, procjenjena varijanca bijelog šuma iznosi $\sigma^2 = 1.699e + 12$. U nastavku pogledajmo značajnost dobivenih koeficijenata i analizu reziduala. 95% intervali za procjenjene koeficijente su:

Koef.	Pouzd. int
$\theta = -0.37$	$[-0.4940394, -0.2458224]$
$\vartheta = -0.6$	$[-0.7275207, -0.4592114]$

Iz tablice vidimo kako nijedan interval ne sadrži 0, pa možemo tada reći da su oba koeficijenta statistički značajna. Nadalje, pogledajmo grafički prikaz analize reziduala.



Iz grafičkih prikaza vidimo da su reziduali do koraka 8 korelirani, p -vrijednost Ljung-Box testa se nalaze ispod linije 0.05, ali dalje nema razloga sumnjati u korelaciju. qqPlot možda sugerira na normalnost pa prikazimo histogram.



Histogram također sugerira na normalnost reziduala. Ako tu tvdrnju testiramo korištenjem Shapiro-Wilk testa i Jarque-Berra testa dobiti ćemo p -vrijednosti redom 0.1249 i 0.05736, stoga na razini značajnosti $\alpha = 0.05$ nema razloga sumnjati u normalnost reziduala.

Na kraju nam treba konačni model. Provedbom ove analize odabrat ćemo za konačni model Model (1), odnosno $\{Y_t\} \sim SARMA(1,0) \times (0,1)_{12}$. Razlog tome je da model (2) je imao jedan neznačajan

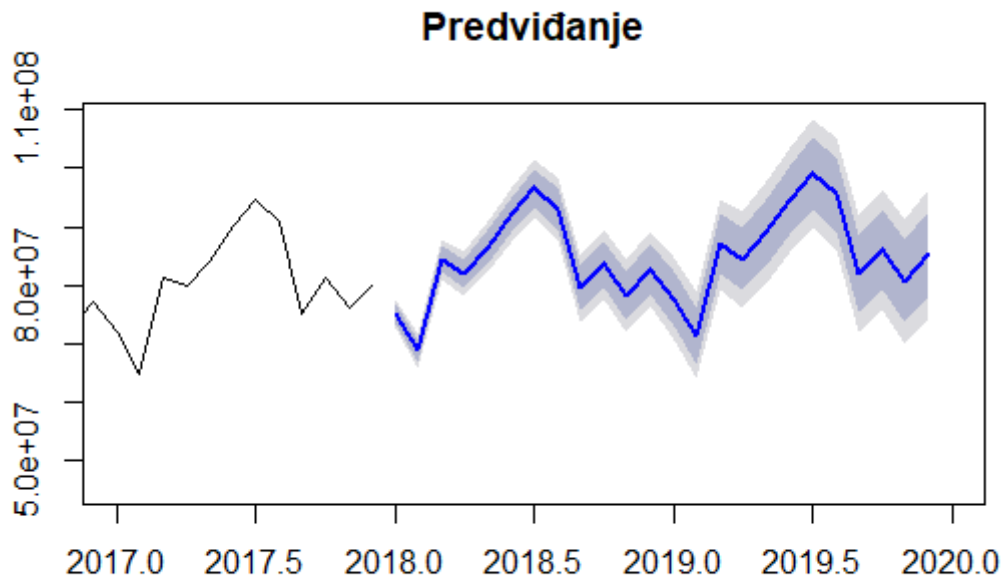
koeficijent te uklanjenjem njega dobili bi model (1), dok model (3) je do koraka 8 imao korelirane rezidualne. Model (1) se čini najprikladnijim zbog *AIC* i *BIC* kriterija jer su njihove vrijednosti za ove parametre najmanje. Koeficijenti su značajni, te su reziduali nekorelirani i normalni. Zapišimo sada konačan model za $\{X_t\}$:

$$X_t = 2.41X_{t-1} + 0.41X_{t-2} + X_{t-12} - 0.59X_{t-13} - 0.14X_{t-14} + Z_t - 0.59Z_{t-12}$$

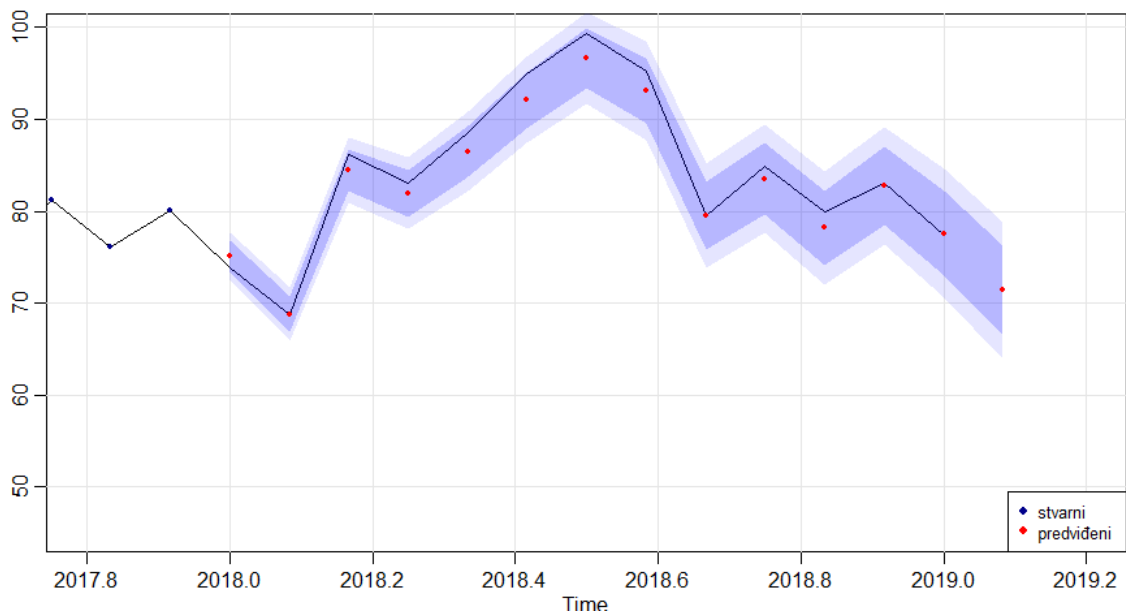
gdje je $\{Z_t\} \sim WN(0, \sigma^2)$, a procjena varijanca bijelog šuma iznosi $\sigma^2 = 1.659$. Srednja greška modela, koju definiramo kao prosjek razlike kvadrata stvarnih i teorijskih vrijednosti, iznosi 1.56.

5 Predviđanje

U ovom dijelu seminarskog rada napraviti ćemo predikciju na osnovu modela koji smo dobili u prethodnom dijelu. Ranije je napomenuto da je model rađen bez posljednjih 14 podataka, pa ćemo to ovdje iskoristiti radi usporedbe stvarnih i predvedenih vrijednosti. Napravimo prvo grafički prikaz predviđenog kretanja procesa za $h = 24$ koraka unaprijed i to prikažimo grafičkim prikazom.



Ako pogledamo plavu liniju koja predstavlja predviđene vrijednosti, možemo vidjeti da model dobro prati sezonalna kretanja i da sam predviđa rastući trend kako je bilo na početnom prikazu. Na idućoj slici pogledajmo usporedni prikaz stvarnih vrijednosti i predviđanih vrijednosti za $h = 14$ koraka.



Osjenčani dijelovi na slikama nam prikazuju 80% i 95% pouzdane intervale predikcije. Ako pogledamo crvene točkice na slici koje predstavljaju predviđenu vrijednost, možemo vidjeti da relativno dobro prate stvarne vrijednosti i sve se nalaze u pozdanom intervalu. Možemo time reći da model ima dobra predikcijska svojstva.

6 Zaključak

Ovim seminarskim radom obrađena je baza Air Revenue Passenger Miles, gdje smo modelirali prijedene putničke milje, za podatke iz 2002 do 2019 godine. Grafičkim prikazima uvidjeli smo samu nestacionarnost procesa, te smo ga prikladnim transformacijama doveli do stacionanarnog, diferenciranje na koraku 12 i zatim na koraku 1. Napravili smo 3 modela od kojih smo uzeli onaj koji se činio najboljim. Takav model nema puno parametara, reziduali su nekorealirani i normalni. Na kraju smo vidjeli kako model dobro predviđa kretanje procesa, tj. kretanje prijeđenih milja, te bi ovakav model mogli koristiti za kvalitetno predviđanje budućih vrijednosti.