# 

Univerzitet u Nišu

Elektronski fakultet

Katedra za računarstvo

Ana Tonić

Izveštaj za projekat iz predmeta tehnike i metode analize podataka

Niš, 2023

Sadržaj

[1](#_Toc136603837)

[1. Istraživanje podataka 3](#_Toc136603838)

[2. Preprocesiranje podataka 4](#_Toc136603839)

[2.1 Uklanjanje visoko korelisanih atributa 4](#_Toc136603840)

[2.2 Q-Q dijagram 5](#_Toc136603841)

[2.3 Detekcija duplikata 6](#_Toc136603842)

[2.4 Vizuelizacija 7](#_Toc136603843)

[2.5 Uklanjanje ouliera 7](#_Toc136603844)

[2.5.1 Uklanjanje outlier-a iz celog skupa podataka odjednom 7](#_Toc136603845)

[2.5.2 Uklanjanje outlier-a za svaki dan zasebno 8](#_Toc136603846)

[2.6 Risemplovanje (na frekvenciji od 1Hz) 8](#_Toc136603847)

[3. Napredna analiza vremenskih serija 8](#_Toc136603848)

[3.1 Testovi stacionarnosti 8](#_Toc136603849)

[3.2 Računanje autokorelacije i parcijalne autokorelacije 8](#_Toc136603850)

[3.3 Atributi 8](#_Toc136603851)

[3.3.1 Active Power A average [kW] 9](#_Toc136603852)

[3.3.2 Demand Total [kW] 12](#_Toc136603853)

[3.3.3 HDI A average [%]' 15](#_Toc136603854)

[3.3.4 'THDU A average [%]' 17](#_Toc136603855)

[3.3.5 'Voltage C average [V] 18](#_Toc136603856)

[3.3.6 'Voltage A average [V]' 20](#_Toc136603857)

[3.3.7 'Voltage B average [V]' 21](#_Toc136603858)

[3.3.8 'Frequency average [Hz]' 23](#_Toc136603859)

[3.3.9 'Power Factor A average' 25](#_Toc136603860)

[3.4 ARIMA model 27](#_Toc136603861)

[4. Evaluacija 28](#_Toc136603862)

[4.1 Korelacija reziduala(autokorelacija) 28](#_Toc136603863)

[4.2 Distribucija reziduala 30](#_Toc136603864)

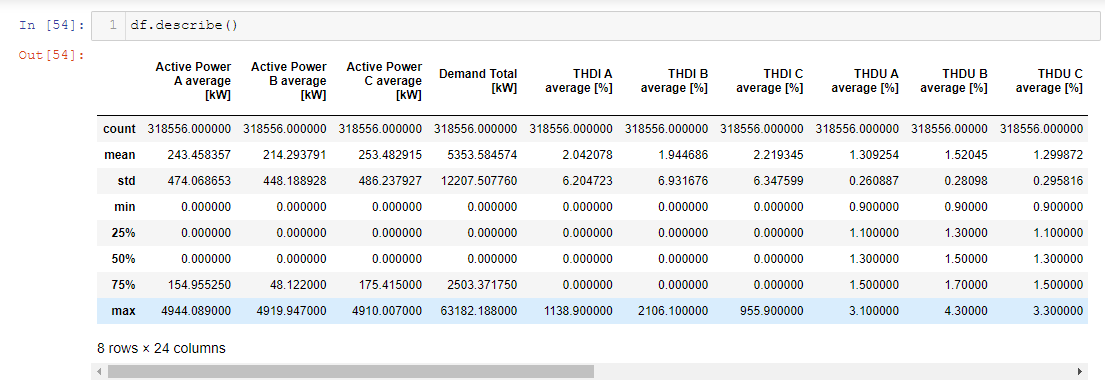
[4.2.1 Histogrami reziduala 30](#_Toc136603865)

[4.2.2 Q-Q plot 32](#_Toc136603866)

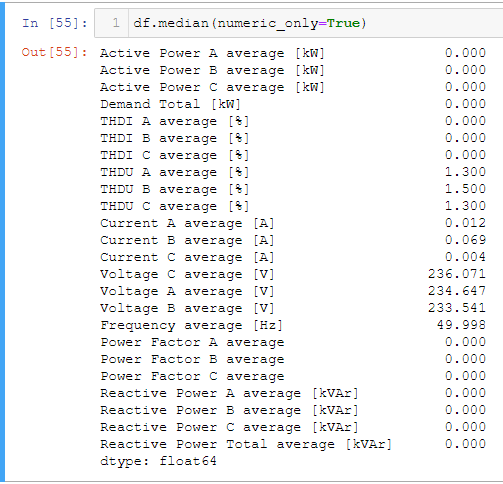
# 

# Istraživanje podataka

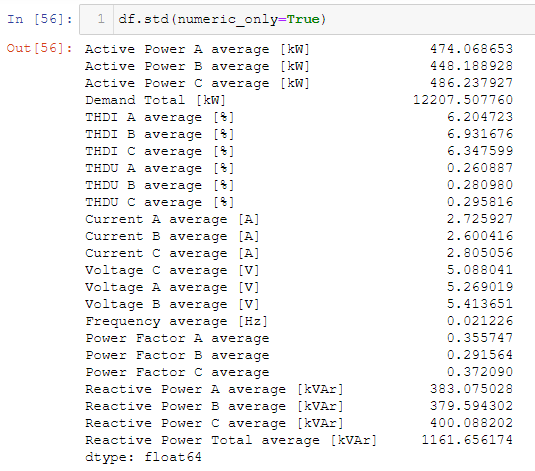
Ovaj skup podataka ima ukupno 318556 redova i 25 kolona. Prvo je urađena provera tipa podataka kao i da li u dataset-u ima null vrednosti. Rezultati pokazuju da su svi podaci tipa float kao i da nema null vrednosti u skupu podataka.Pozivom metode describe() prikazan je broj elemenata u svakoj koloni kao i statistički podaci mean, std, min i percentili.



Zatim je pozvana metoda median() koja je pokazala da je medijana za vecinu kolona jednaka ili blizu nule sem za pojedine kolone. Kolone koje pokazuju THDU vrednost imaju medijanu blizu 1, a kolone koje pokazuju napon na svim fazama imaju median blizu 234.



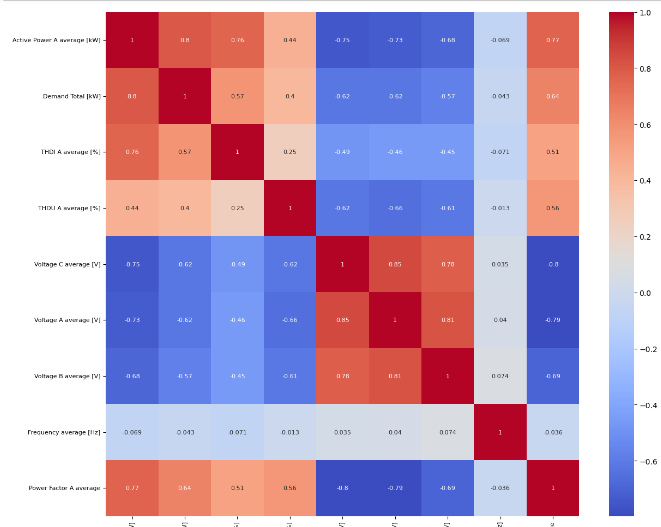
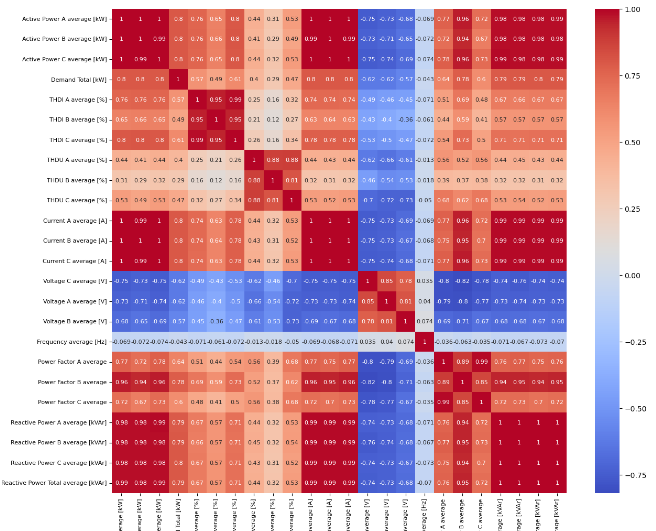
Zatim je izračunata standardna devijacija za sve kolone. Iz prikazanog se vidi da najveću standardnu devijaciju imaju kolone koje pokazuju srednju aktivnu snagu i srednju reaktivnu snagu.



# Preprocesiranje podataka

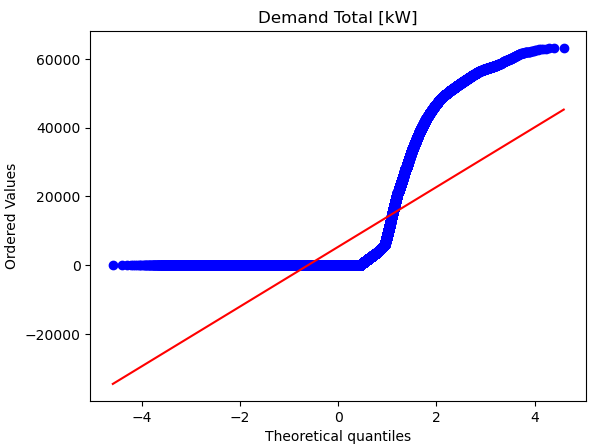
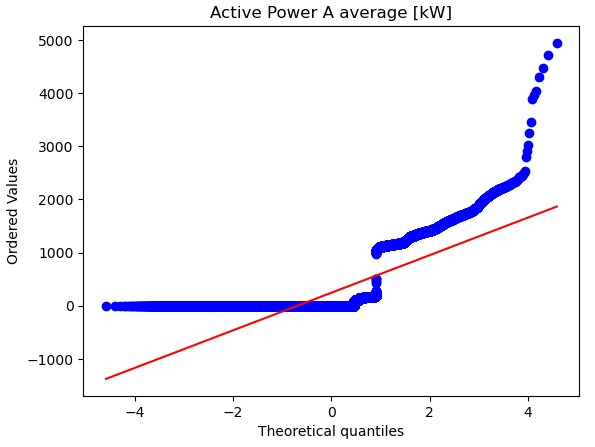
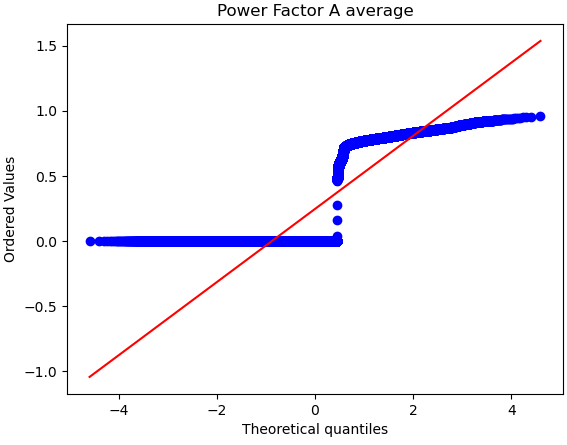
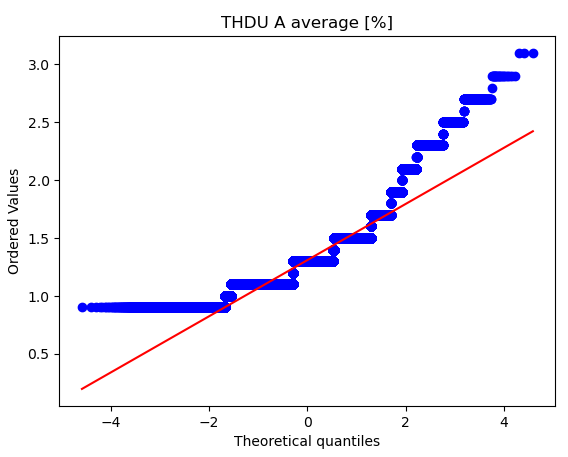
## Uklanjanje visoko korelisanih atributa

Najpre je prikazana matrica korelacije na kojoj se vidi da ima dosta atributa koji su visoko korelisani. Izbačene su kolone tako da ne postoje vise kolone koje su korelisane sa nekom drugom kolonom više od 90%. Na slikama su prikazane matrice korelacije pre i posle izbacivanja visoko korelisanih atributa.

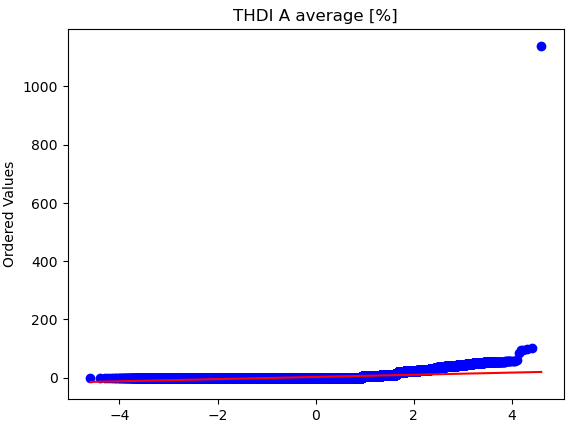
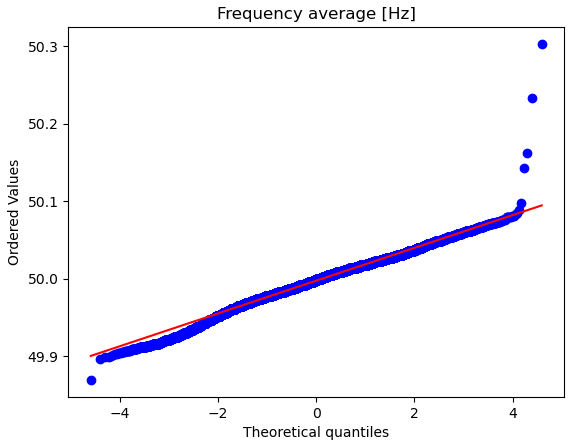


## Q-Q dijagram

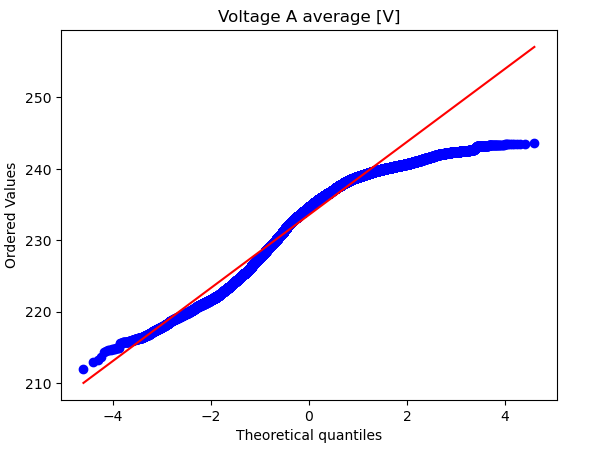
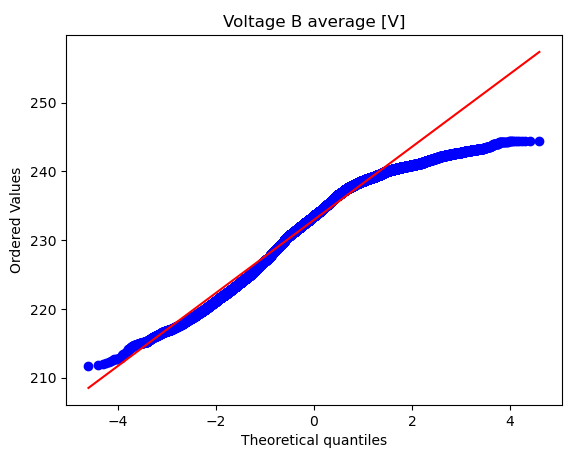
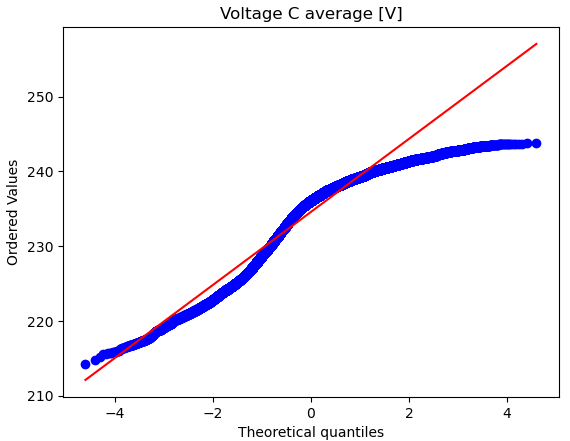
U ovom koraku su prikazani q q plotovi koji treba da pokaze da li atributi imaju dormalnu distribuciju. Kolone Active Pover A average, Demand Total, THDU A average i Power Factor A average ni malo ne prate normalnu raspodelu.

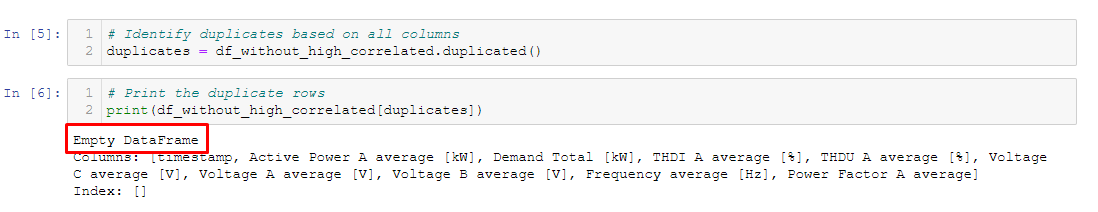
Kolone THDI A average i Frequency average uglavnom prate normalnu raspodelu ali postoje podaci koji odstupaju od toga.

Kolone Voltage A average, Voltage B average i Voltage C average samo delimicno prate normalnu raspodelu atributa.

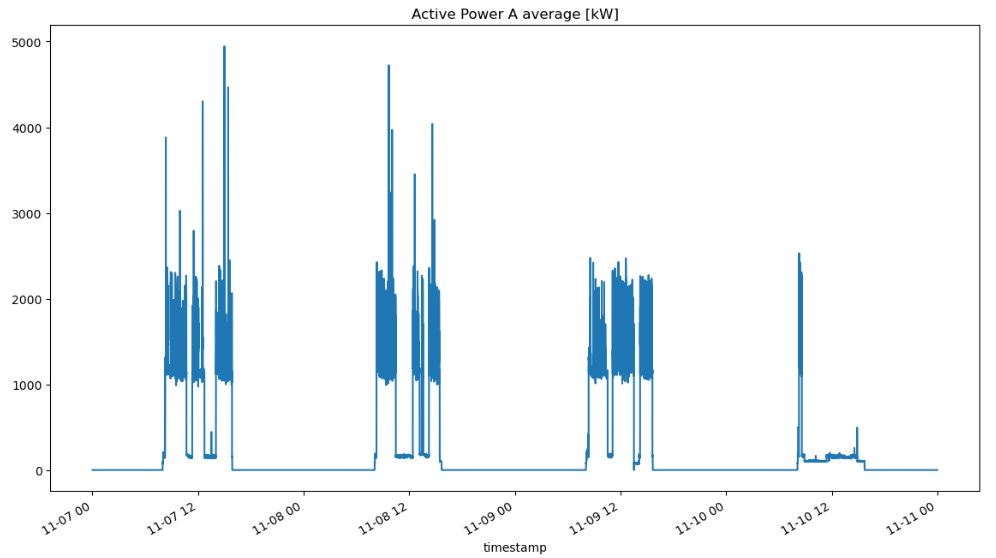
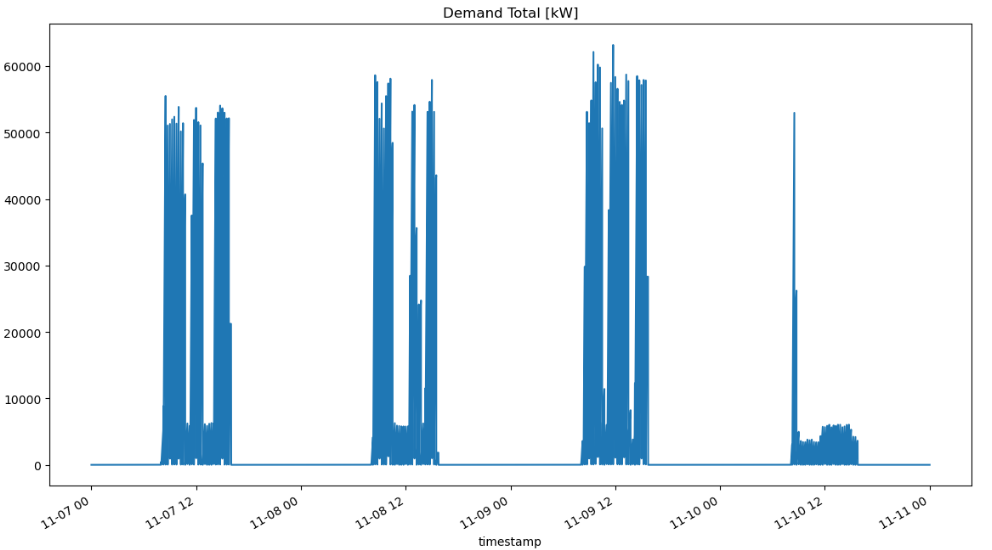
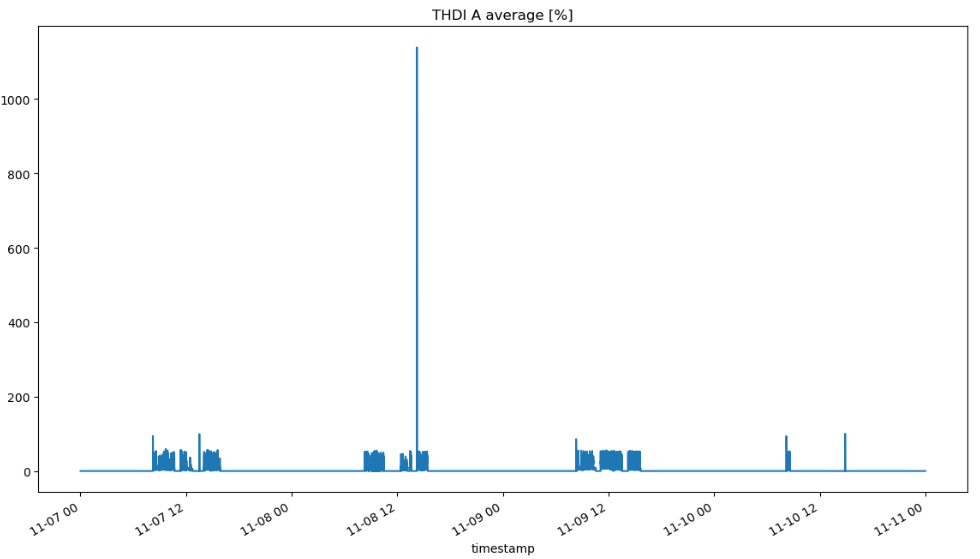
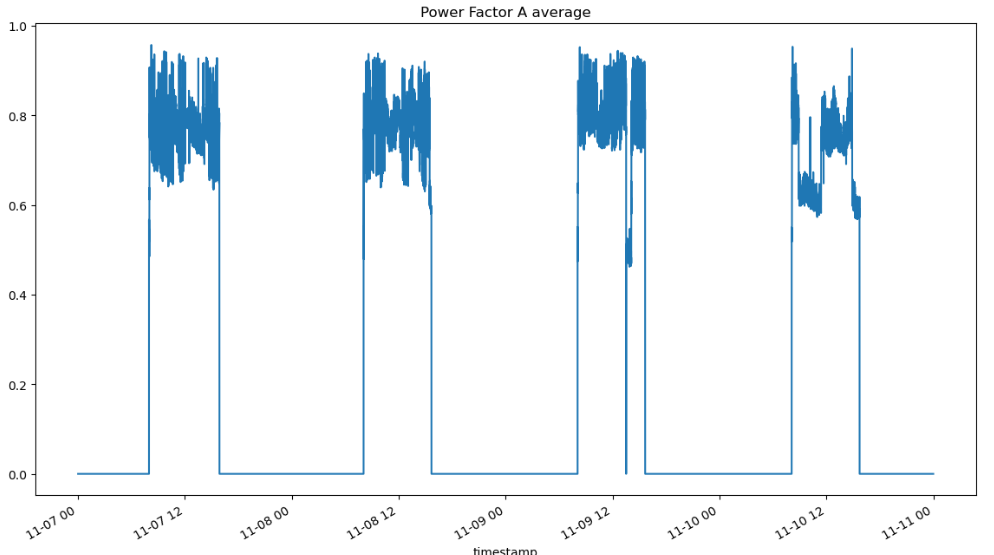
  

## Detekcija duplikata

U ovom koraku je izvršena provera da li postoje duplikati među podacima, i utvrđeno je da ne postoje.  


## Vizuelizacija

U ovom koraku su plotovani svi atributi kako bi se stekao bolji uvid u podatke. Može se primetiti da je za sledeće kolone vrednost veća od 0 samo za vreme radnog vremena: Active Power A average, Demand Total, THDI A average i Power Factor A average. Zbog toga će analiza podataka u ovom radu biti rađena nad skupom podataka koji je dobijen tako što su za svaki radni dan izvučeni podaci prikupljeni u toku radnog vremena a zatim agregirani u jedan skup podataka.

## Uklanjanje ouliera

Sledeći korak je uklanjanje outlier-a. Za ove podatke nije moguće ostvariti empiric rule za detekciju outlier-a zbog toga što atributi nemaju normalnu raspodelu. Druga metoda koja se može primeniti je IQR metoda.Ova metoda takođe pretpostavlja da podaci imaju normalnu raspodelu pa se ne očekuju dobri rezultati primenom ove metode, ali biće provereno eksperimentalno. Pokušaj da se uklone outlier-i izvršen je na dva načina: uklanjanjem outlier-a iz celog dataset-a odjednom i uklanjanjem outlier-a svakog dana pojedinačno pa onda objedinjavanjem podataka.

### Uklanjanje outlier-a iz celog skupa podataka odjednom

Od početnog broja uzoraka 318.556, nakon uklanjanja outlier-a ovom metodom preostaje 252.679 uzoraka, što znači da je izgubljeno 20.6% podataka, što nije prihvatljivo.

### Uklanjanje outlier-a za svaki dan zasebno

Podaci unutar skupa podataka su prvo grupisani po danu, a onda su podeljeni na 4 zasebna skupa podataka. Za svaki dan formiran je novi skup podataka koji obuhvata uzorke samplovane u toku radnog vremena, dakle od 8 ujutru do 16h poslepodne. Uklanjanjem outlier-a za svaki dan posebno a zatim spajanjem ta 4 skupa podataka utvrđeno je da je gubitak podataka 19.5%, što nije prihvatljivo, tako da je dalja analiza vremenskih serija podataka nastavljena bez uklanjanja outlier-a.

## Risemplovanje (na frekvenciji od 1Hz)

U ovom koraku promenjena je frekvencija svih podataka na sekundu, korišćenjem prosečne vrednosti u slučaju da se više podataka nalazi u jednom sekundu. Popunjene su sve praznine koje su nastale ovim procesom, i spojeni su svi podaci u jedan DataFrame.

**Nakon toga, podaci za svaki dan za vreme radnog vremena su spojeni u jedan dataset koji je korišćen na dalje za analizu.**

# Napredna analiza vremenskih serija

## Testovi stacionarnosti

Da bi se ispitalo da li su podaci stacionarni kako bi se nad njima mogli primenjivati modeli za predviđanje vremenskih serija podataka prvo se radi ispitivanje stacionarnosti. Za ispitivanje stacionarnosti koriščeno je nekoliko različitih testova, a to su ADF test, Dickey-Fuller GLS test, Phillips-Perron Test i KPSS test. Svi ovi testovi su dali isti rezultat a to je da su sve kolone slabo stacionarne.

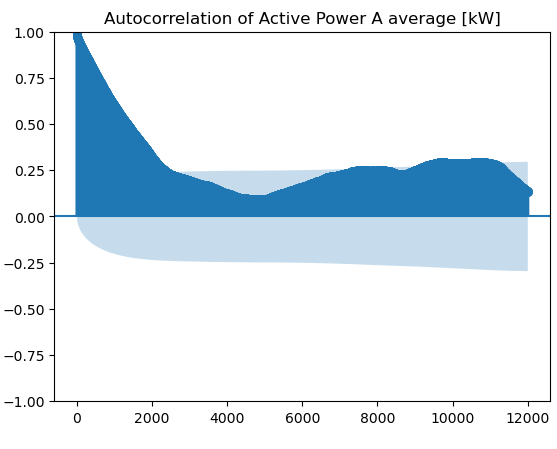
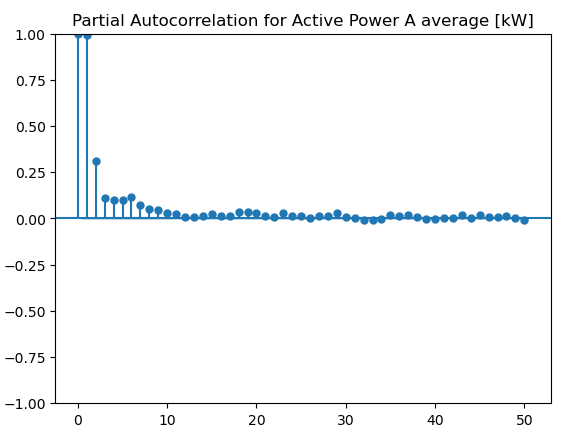
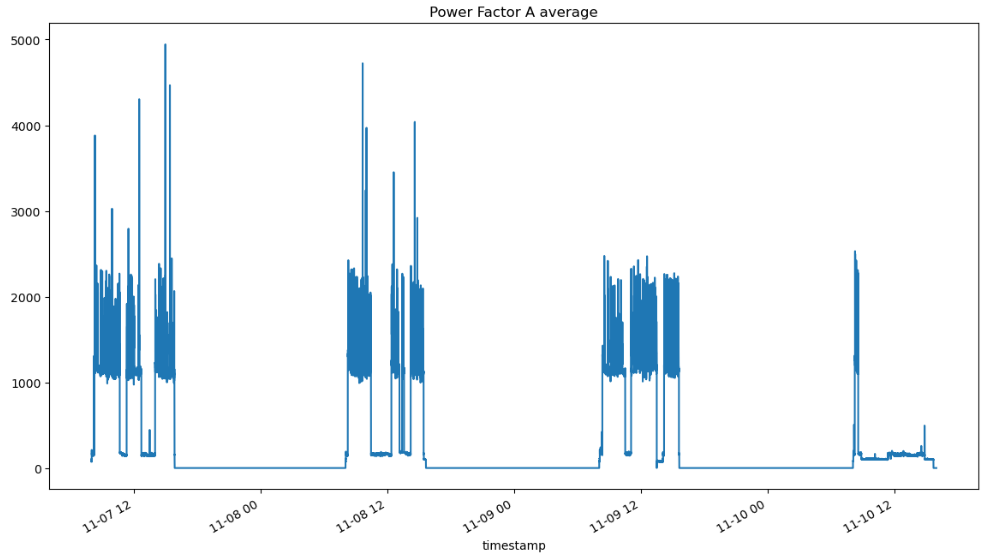
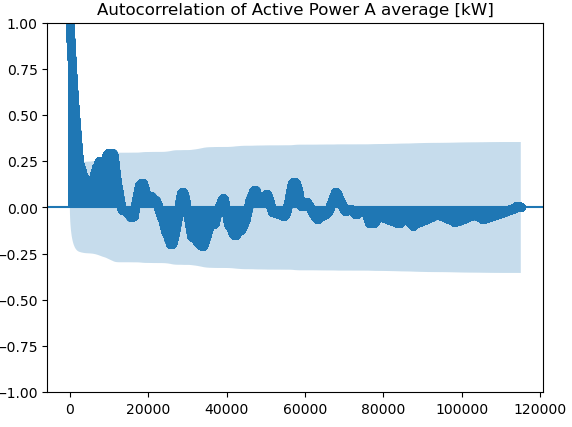
## Računanje autokorelacije i parcijalne autokorelacije

U ovom koraku prikazani su grafici autokorelacije i parcijalne autokorelacije za sve atribute kako bi se stekao uvid u podatke i koliko lagova je izmad significant threshold na oba grafa. Za acf graf za broj lagova je uzet ukupan broj podataka u skupu podataka, a za pacf graf 50 lagova. Za svaki atribut su grafovi posebno razmatrani u sledećem koraku.

## Atributi

U ovom delu analize posmatrana je svaka vremenska serija zasebno kroz grafike na kojima su vizuelizovani sami atributi i prikazani grafici autokorelacije i parcijalne autokorelacije. Iako su svi testovi pokazali da su vremenske serije slabo stacionarne, na osnovu grafika na kojima su vizuelizovane vrednosti atributa vizuelno se može zaključiti o stacionarnosti podataka kao i na osnovu acf i pacf grafa. Ovi grafici su takodje korišćeni kao ukazatelji na najbolje parametre za ARIMA model nad svakom vremenskom serijom.

### Active Power A average [kW]

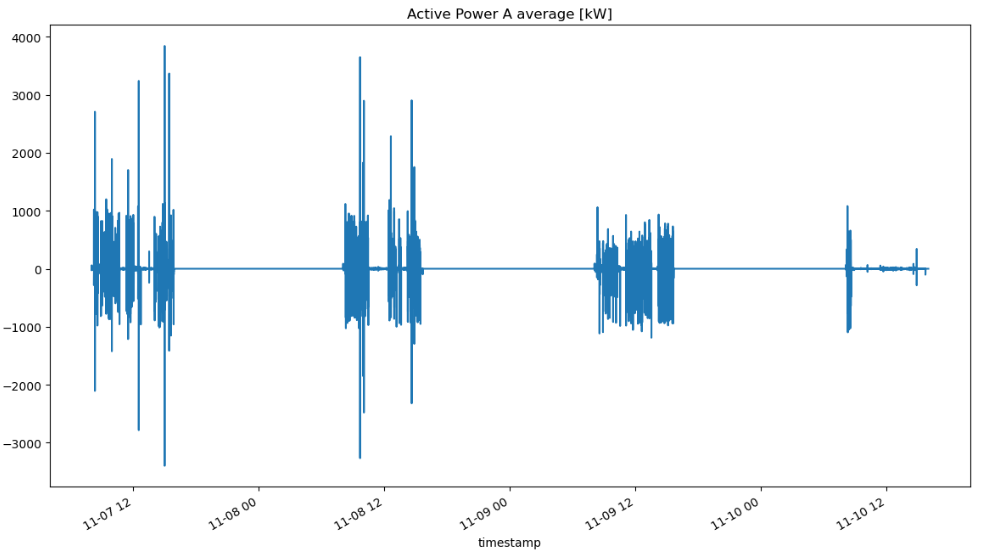
 

ACF graf opada veoma sporo sto je pokazatelj nestacionarnosti u podacima, takodje na većem broju lagova se vidi sinusoidal shape što je drugi pokazatelj nestacionarnosti. Pored toga i testovi daju slične rezutate, pa je potrebno differencirati ovu kolonu.

Na PACF grafu se pokazuje da je significant threshold dosta nisko što moze da bude pokazatelj nestacionarnosti.

Za poboljšanje stacionarnosti vremenske serije može se koristiti više različitih metoda kao što su diferenciranje, korenovanje i logaritmovanje. Sve tri metode su primenjene nad ovom vremenskom serijom.

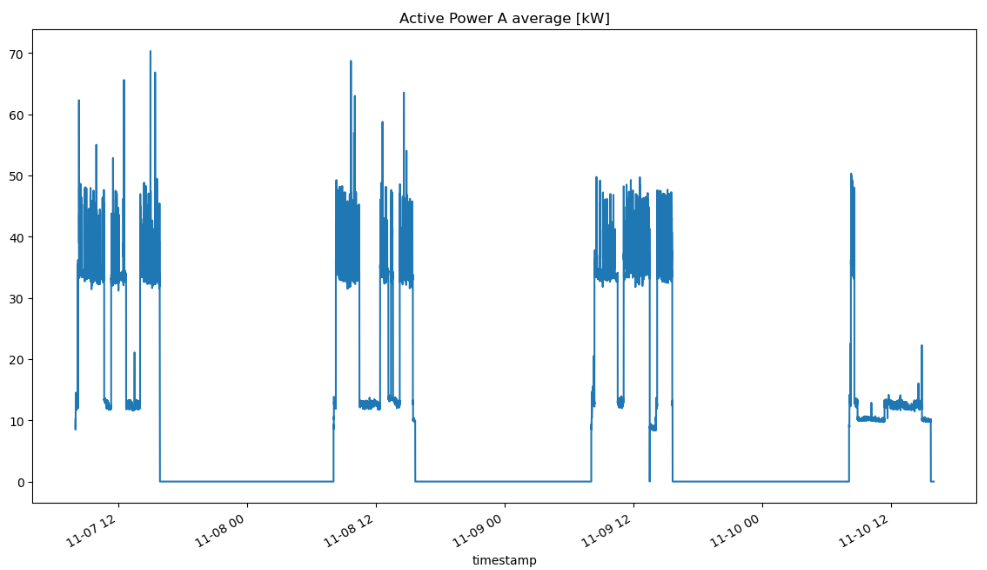
Prva metoda koja je primenjena je diferenciranje:



1Grafik vremenske serije 'Active Power A average [kW]' nakon diferenciranja

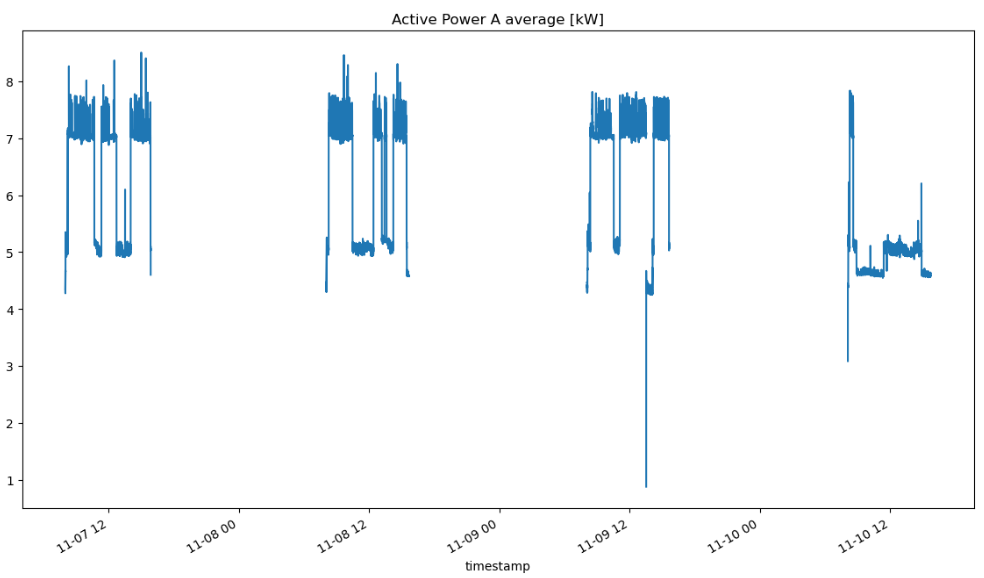
Pošto je izvršena diferencijacija ovog featur-a, d parametar za ARIMA model je d=1.

Druga metoda koja je primenjena radi poboljšanja stacionarnosti nad originalnom vremenskom serijom je korenovanje. Nakon korenovanja grafik vremenske serije izgleda ovako:



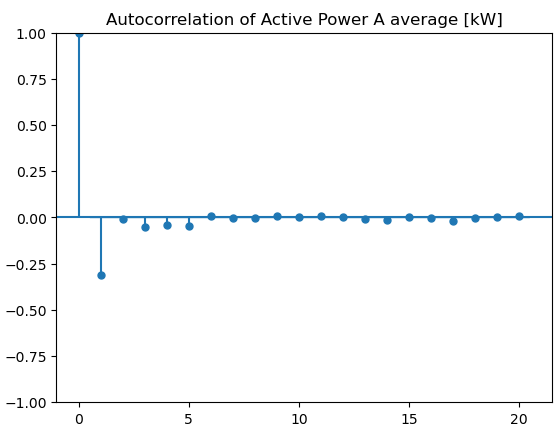
2 Grafik vremenske serije 'Active Power A average [kW]' nakon korenovanja

Nakon korenovanja vremenske serije i plotovanja zaključuje se da korenovanje nije znacajno uticalo na poboljšanje stacionarnosti vremenske serije.

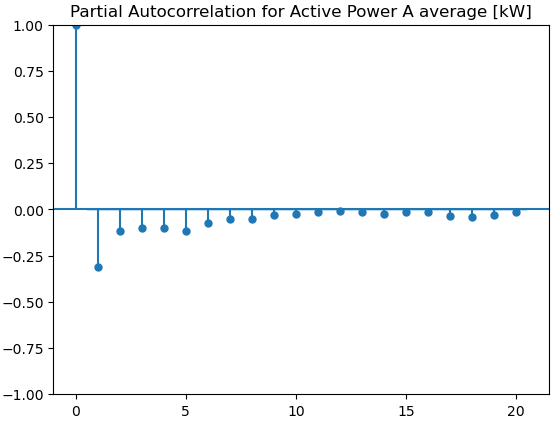
Treća metoda za poboljšanje stacionarnosti koja je primenjena je logaritmovanje. Nakon logaritmovanja grafik vremenske serije izgleda ovako: 

3 Grafik vremenske serije 'Active Power A average [kW]' nakon logaritmovanja

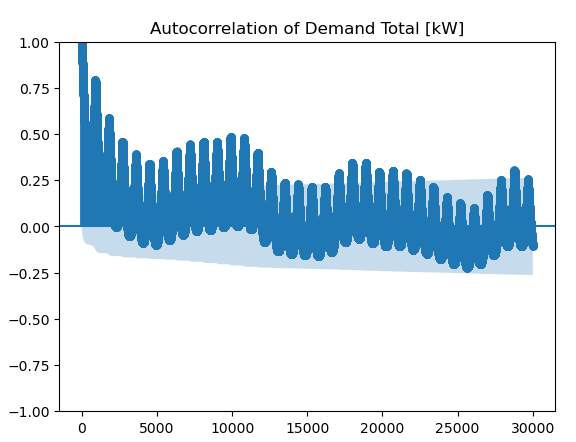
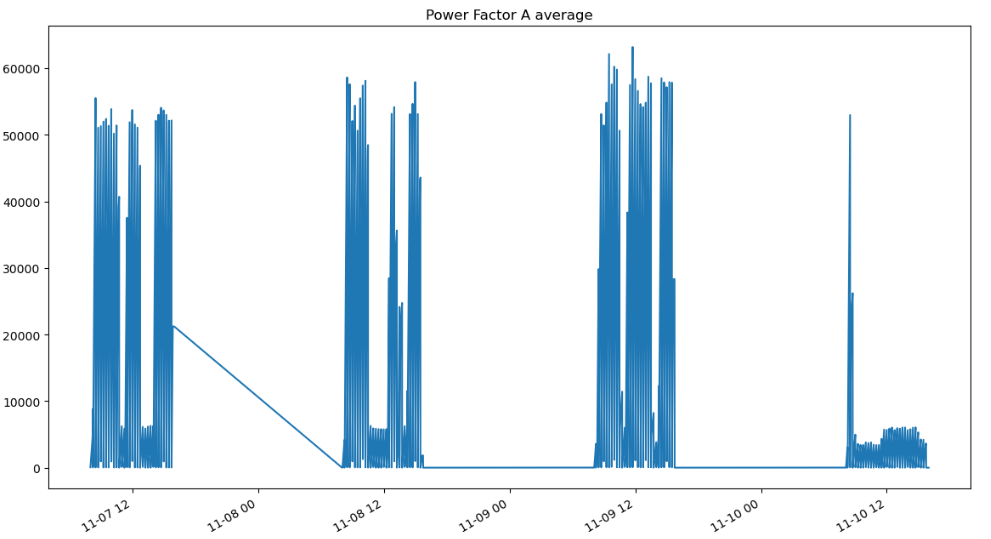
Nakon logaritmovanja, izgleda da su izgubljene neke vrednosti. Grafici ne izgledaju stacionarnije.

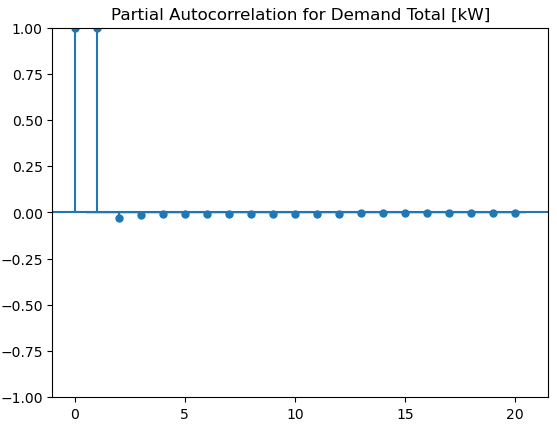


Moving average order (q) se odredjuje na osnovu broja vaznih lagova na ACF grafu. Na osnovu ovog grafa q = 1.

 Na osnovu PACF grafa odredjuje se p parametar u ARIMA modelu koji predstavlja red AR modela. Broj znacajnih lagova na PACF grafu odredjuje red AR modela. Na osnovu ovog grafa p=1.

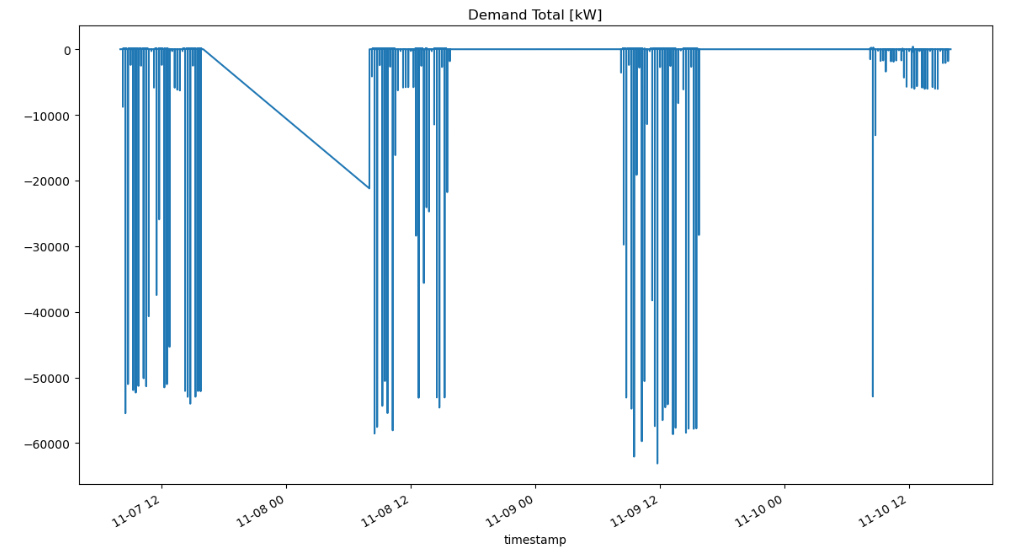
### Demand Total [kW]



 Vizuelno se moze zakljuciti da ova vremenska serije nije stacionarna, javlja se sezonalnost, fluktuacija(std) nije konstantna kao ni srednja vrednost u podacima.

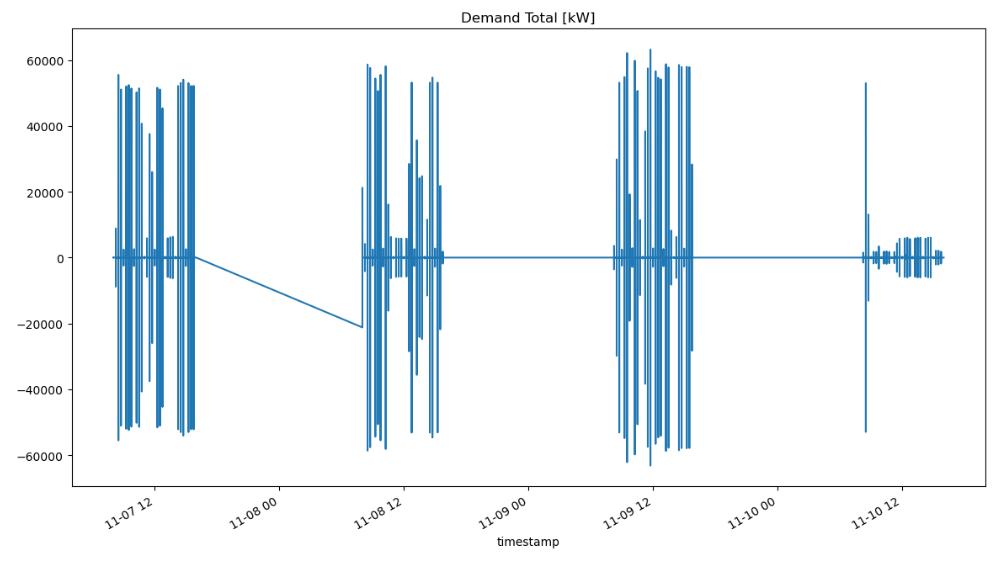
ACF graf podseca na sinusoidalan oblik i prikazuje preveliki broj lagova kao significant. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje, korenovanje i logaritmovanje.

# 



4 Izgled vremenske serije nakon diferenciranja

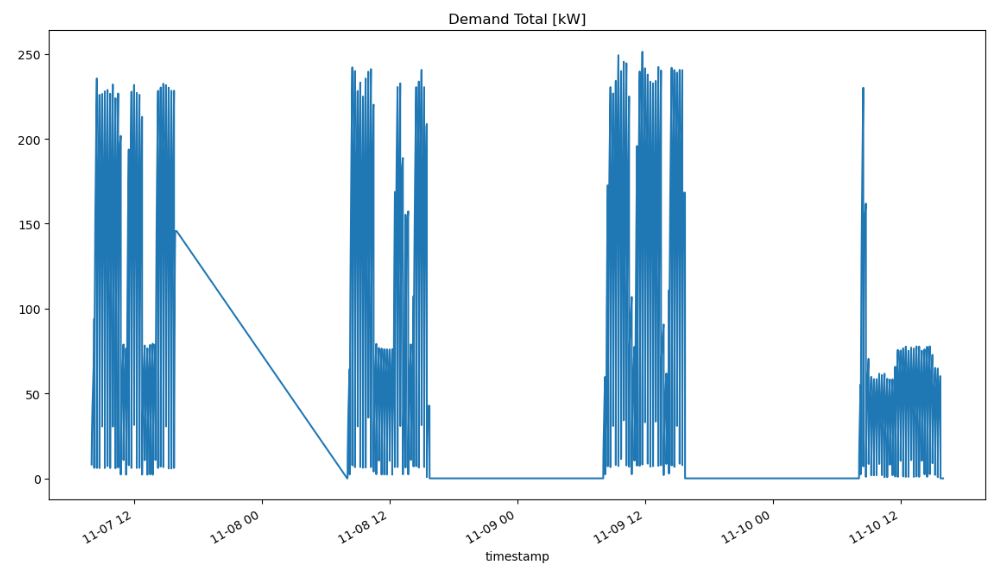
Kako ova vremenska serija nakon diferenciranja i dalje ne izgleda stacionarno, diferenciranje će opet biti izvršeno.



5 Izgled vremenske serije nakon dvostrukog diferencirnja

Kako je dva puta izvrseno diferenciranje parametar d za ARIMA model za ovu vremensku seriju je 2.

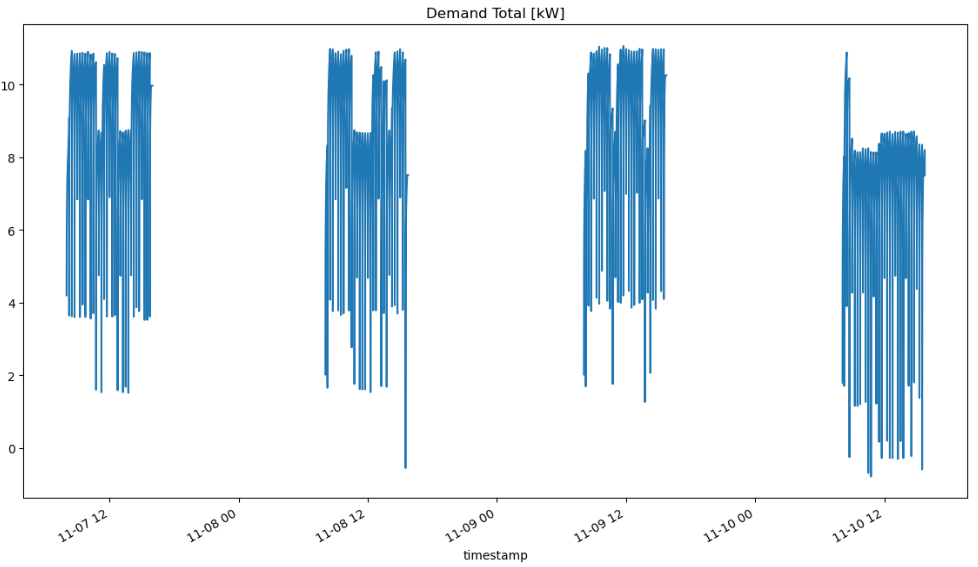
Druga primenjena metoda za poboljšanje stacionarnosti je logaritmovanje.



6 Izgled vremenske serije nakon korenovanja

Nakon korenovanja grafik se nije mnogo promenio i ne izgleda kao da je vremenska serija stacionarnija.

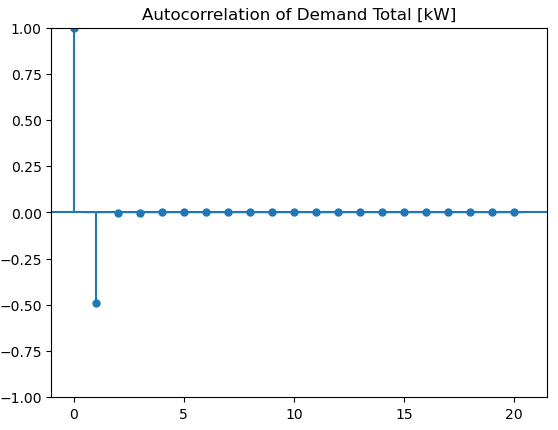
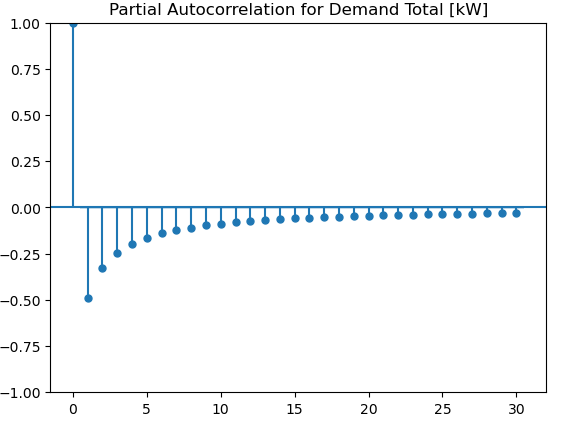
Sledeća metoda za poboljšanje stacionarnosti koja je primenjena je logaritmovanje:



7 Izgled vremenske serije nakon logaritmovanja

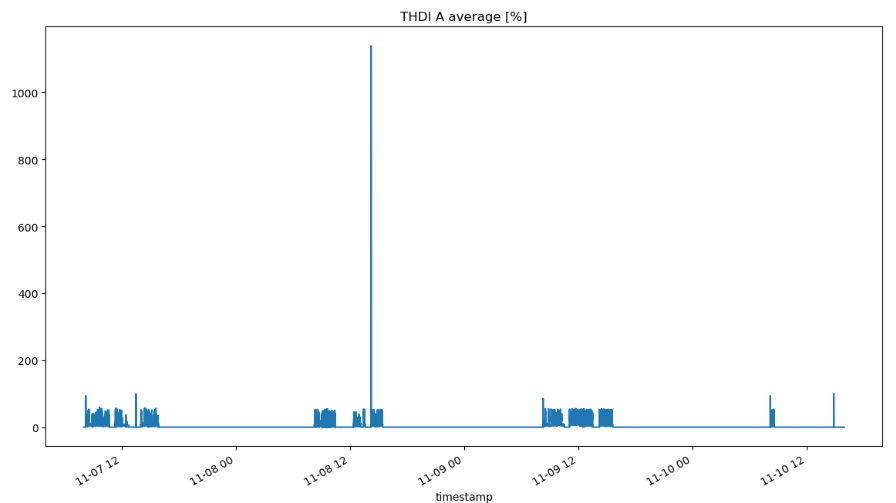
Nakon što su primenjene tri metode za poboljšanje stacionarnosti, moj zaključak je da dvostruko diferenciranje najviše utiče na poboljšanje stacionarnosti.

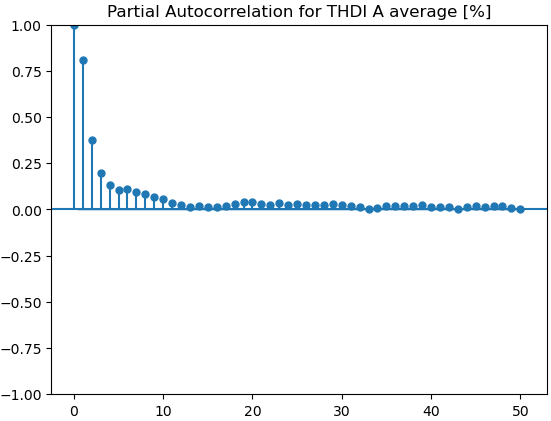
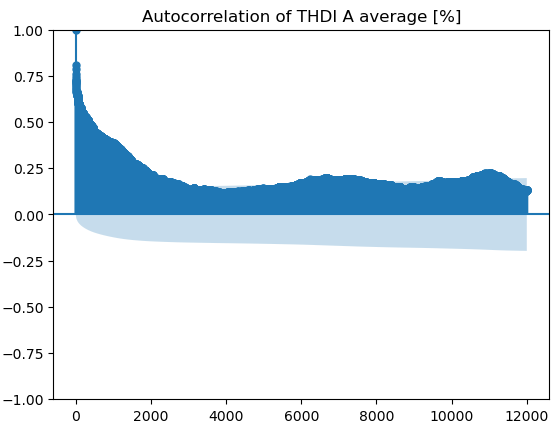
Kako metode za poboljšanje stacionarnosti korenovanje i logaritmovanje nisu dale dobre rezultate za prve dve vremenske serije u nastavku će za svaku vremensku seriju biti primenjivana samo diferencijacija za postizanje.

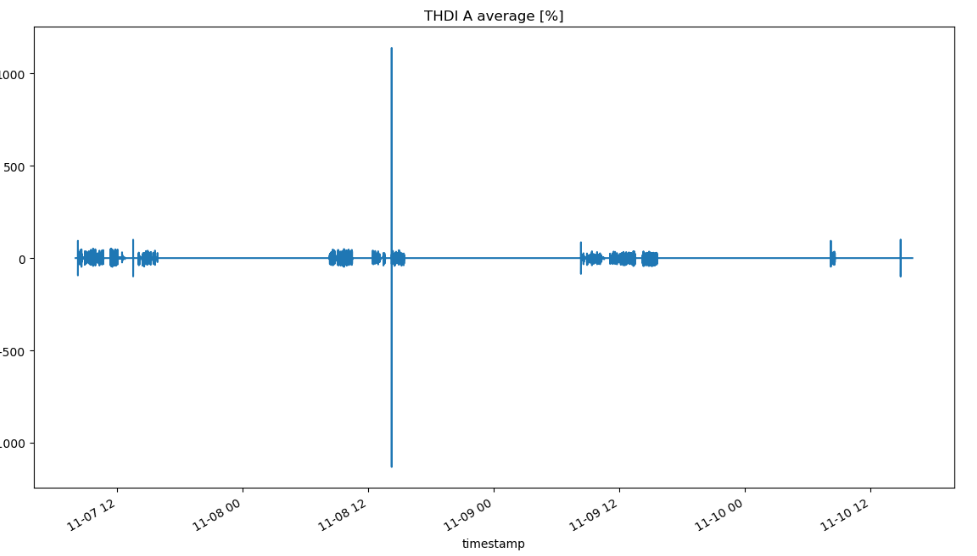
Kako PACF ima vise znacajnih lagova, ali brzo padaju a ACF ima 1 znacajan lag, onda mogu biti isprobana 2 modela.  
Jedan model koji ima samo MA deo, a red tog modela je broj znacajnih lagova na ACF grafu. Dakle q = 1. Ovo resenje je predlozeno na osnovu [ovog videa (08:50).](https://www.youtube.com/watch?v=CAT0Y66nPhs&t=5s&ab_channel=DataScienceShow)  
Drugi u kome je red AR modela odredjen na osnovu znacajnih lagova na PACF grafu. Uzece se u obzir samo lagovi sa korelacijom vecom od 0.5 tako da je p = 1.

### HDI A average [%]'



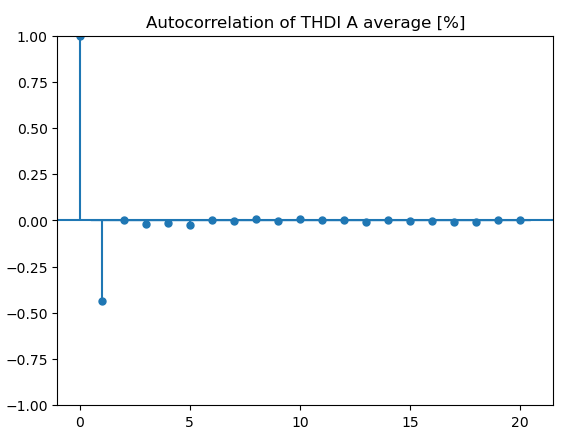


ACF prikazuje preveliki broj lagova kao znacajne. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje.

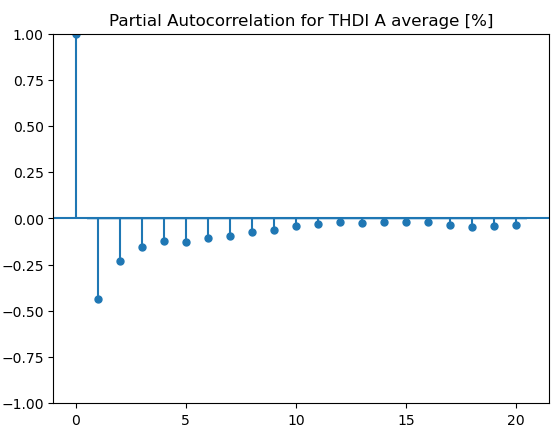


8 Izgled vremenske serije nakon diferenciranja

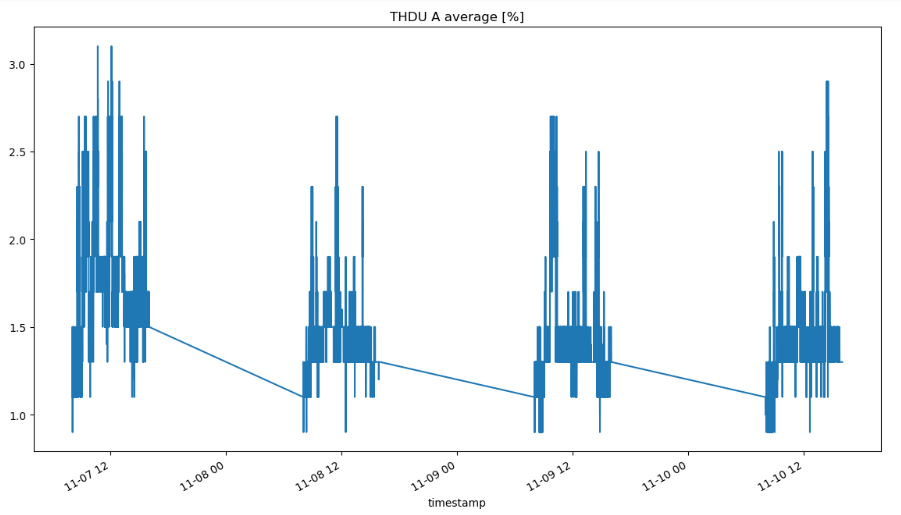
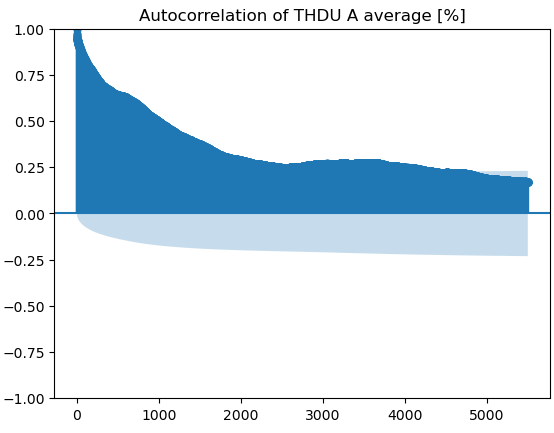
Posto smo diferencirali ovaj feature, parametar d za arima model je 1.

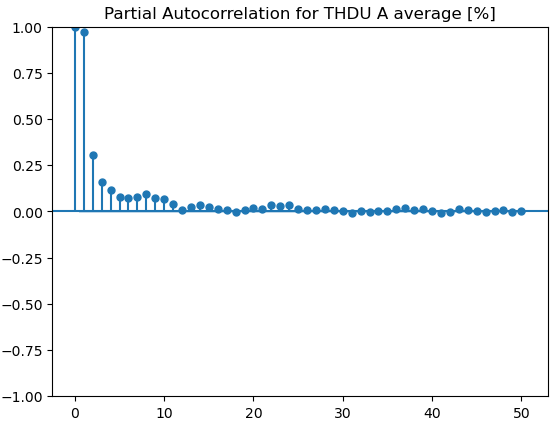


S obzirom da grafik pokazuje 1 znacajan lag, onda ce parametar q biti 1.

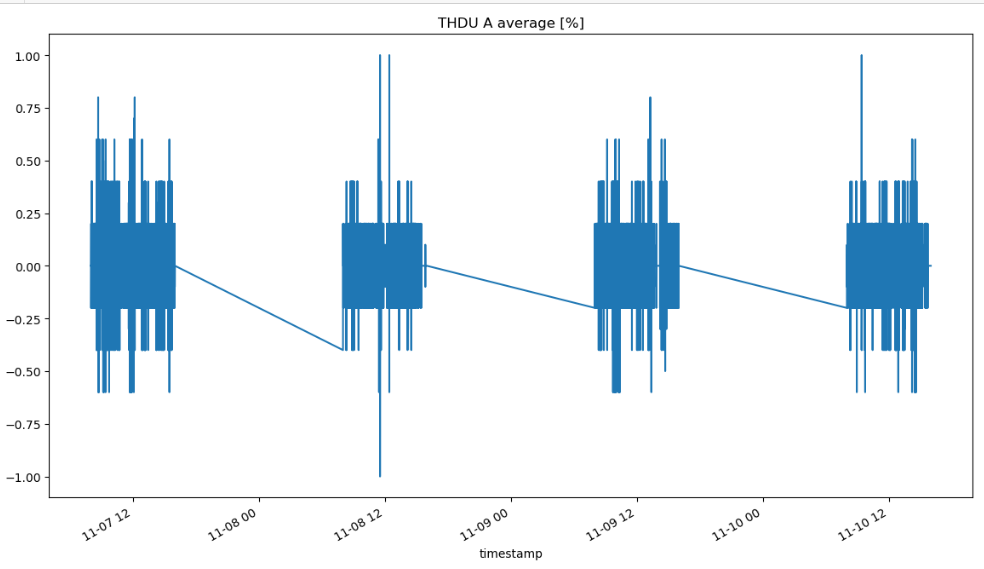
Kako PACF ima vise znacajnih lagova, ali brzo padaju a ACF ima 1 znacajan lag, onda mogu biti isprobana 2 modela.  
Jedan model koji ima samo MA deo, a red tog modela je broj znacajnih lagova na ACF grafu. Ovo resenje je predlozeno na osnovu [ovog videa (8:50).](https://www.youtube.com/watch?v=CAT0Y66nPhs&t=5s&ab_channel=DataScienceShow)  
Drugi u kome je red AR modela odredjen na osnovu znacajnih lagova na PACF grafu. Uzece se u obzir samo lagovi sa korelacijom vecom od 0.25 tako da je p = 1.

### 'THDU A average [%]'

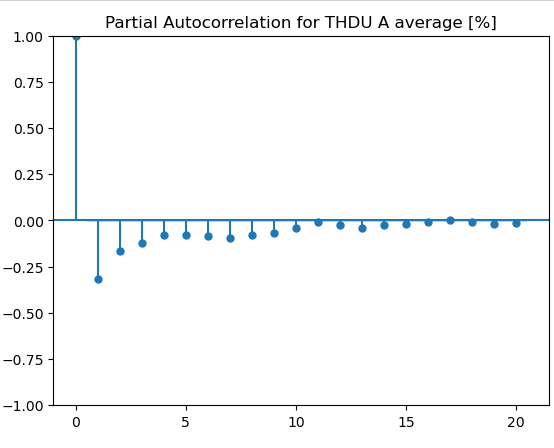
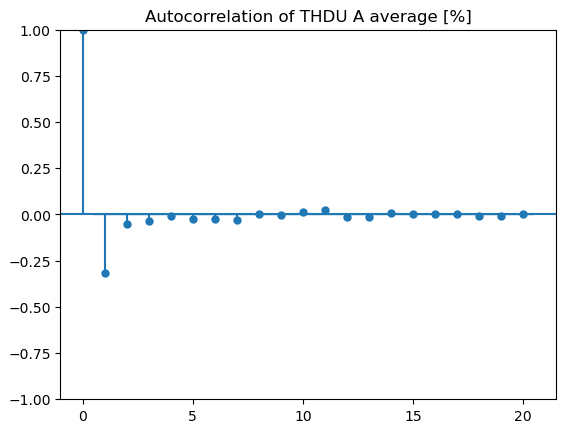


ACF prikazuje preveliki broj lagova kao znacajne. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje.



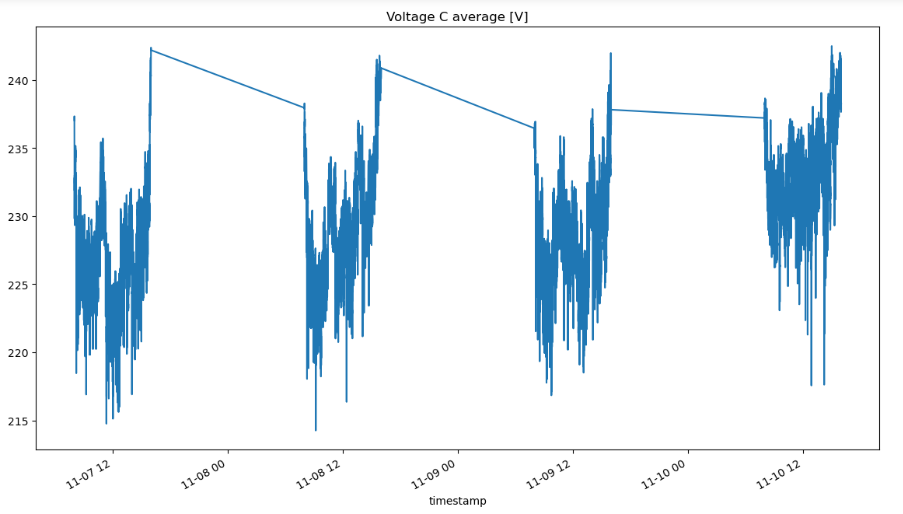
9 Izgled vremenske serije nakon diferenciranja

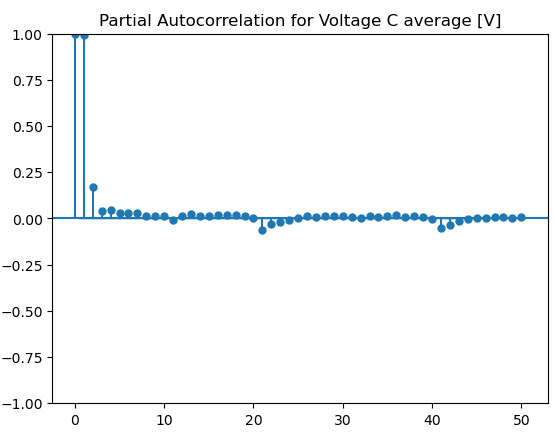
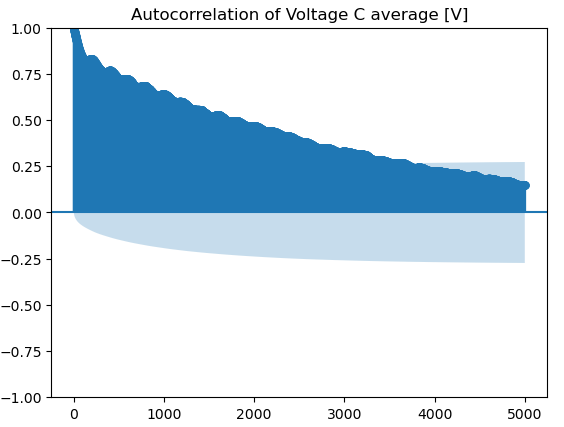
Posto smo diferencirali ovaj feature, parametar d za arima model je 1.



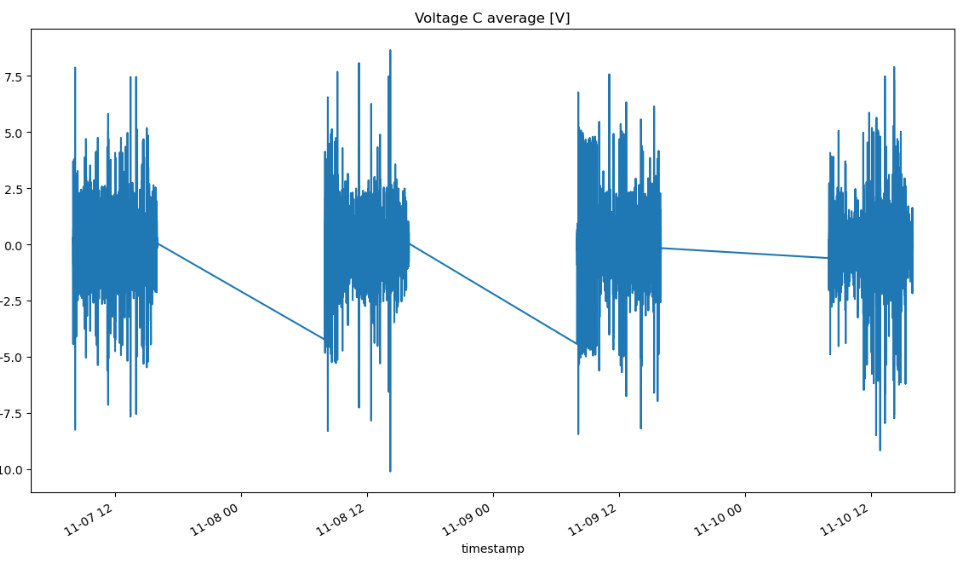
ACF prikazuje jedan znacajan lag sto znaci q=1. Na osnovu ovog grafika p=1.

### 'Voltage C average [V]



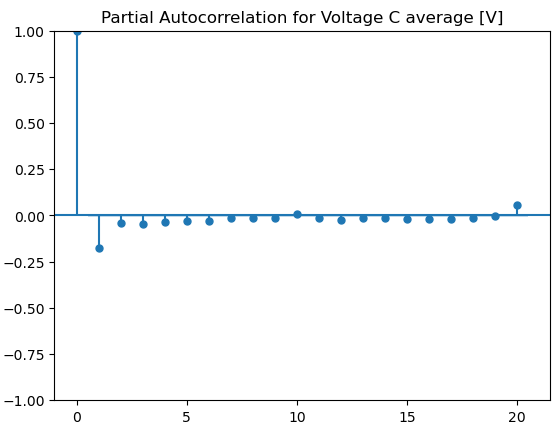
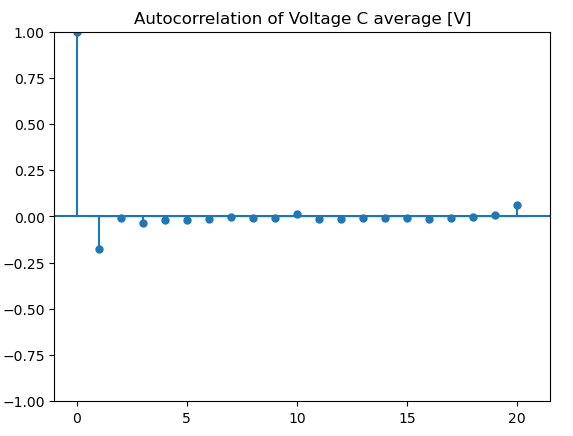


ACF prikazuje preveliki broj lagova kao znacajne. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje.



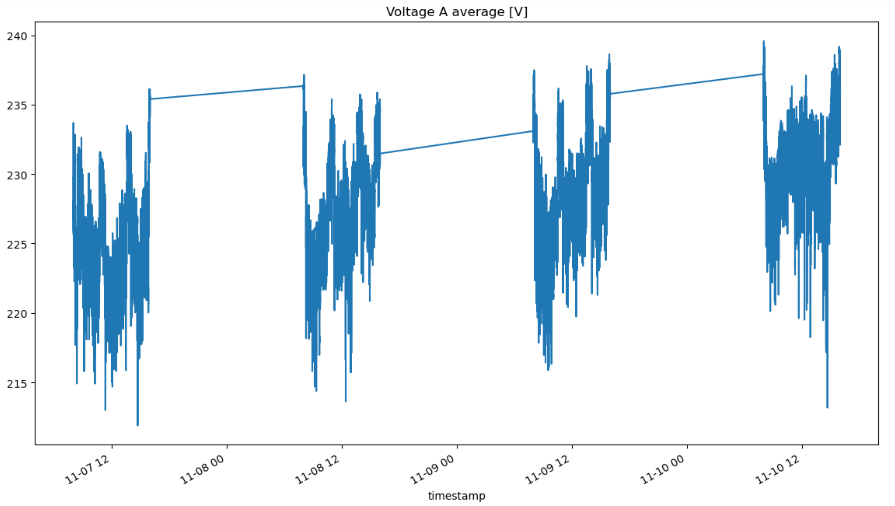
10 Izgled vremenske serije nakon diferenciranja

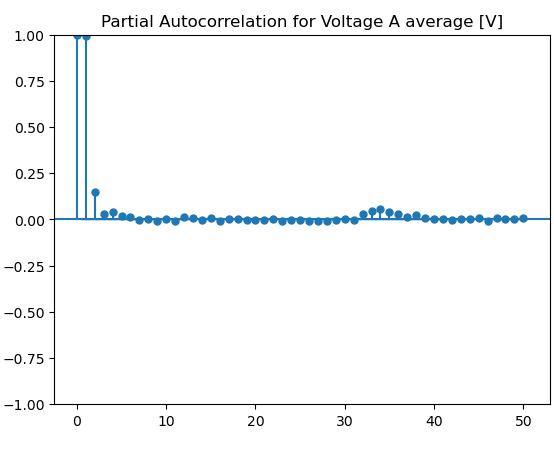
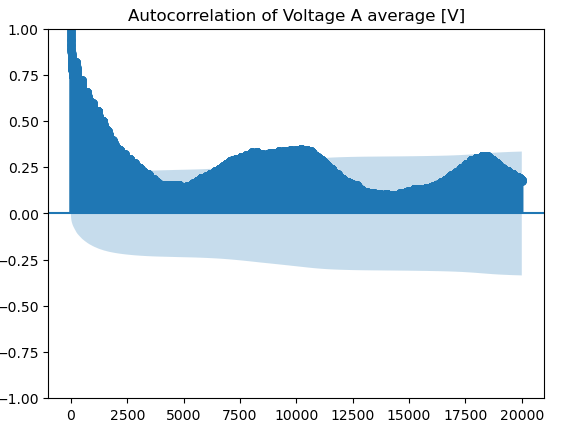
Zbog diferenciranja d = 1.



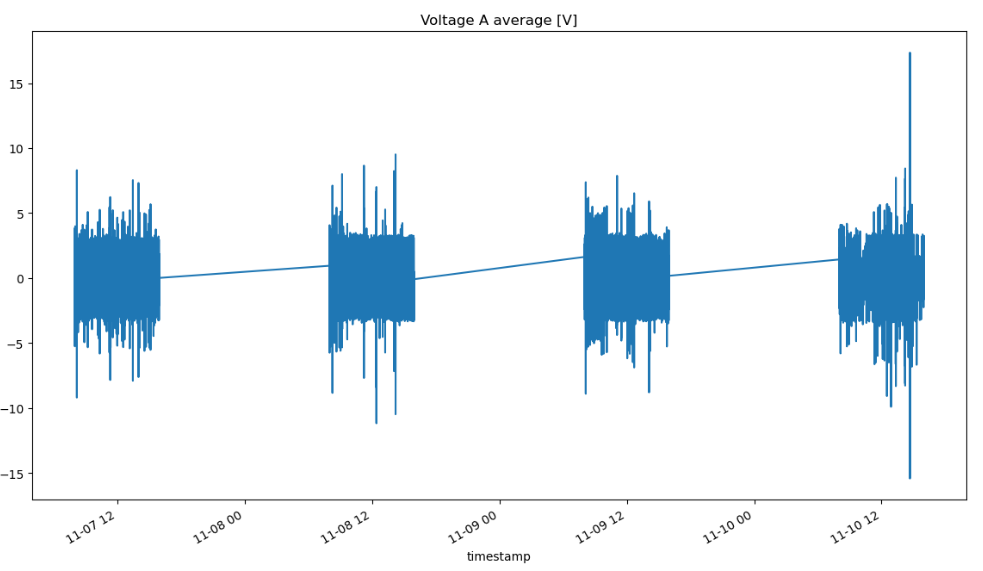
Na osnovu ACF grafika q = 1. Na osnovu PACF grafika p = 1.

### 'Voltage A average [V]'



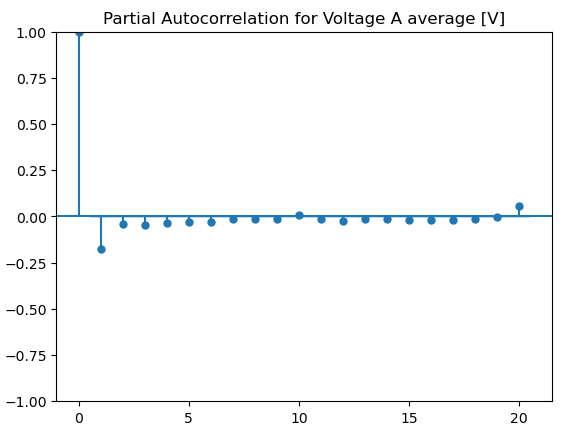
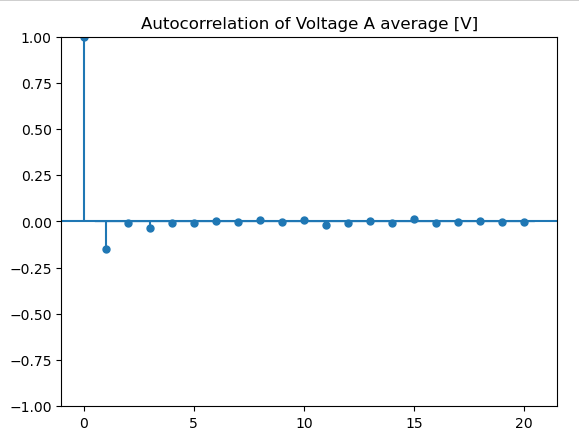


ACF prikazuje preveliki broj lagova kao znacajne. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje.



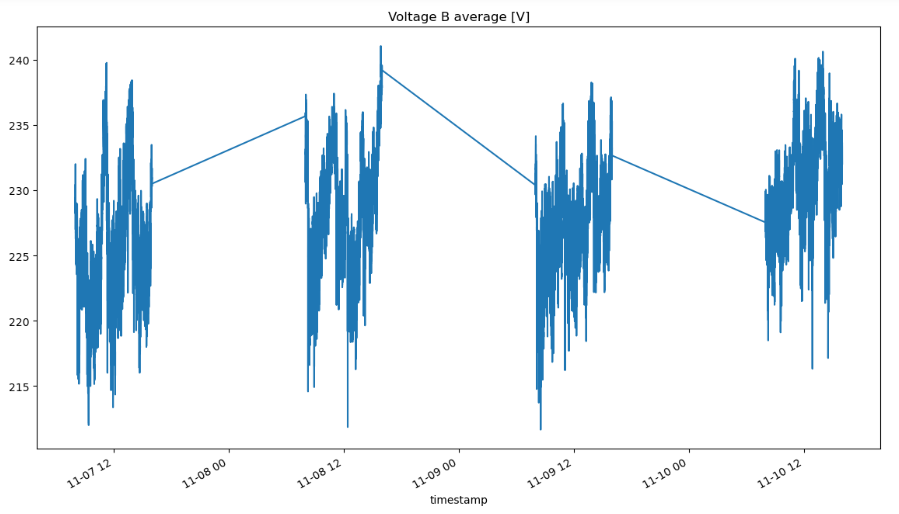
11 Izgled vremenske serije nakon diferenciranja

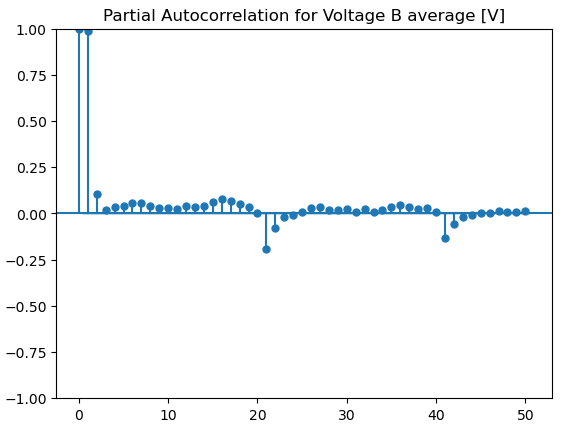
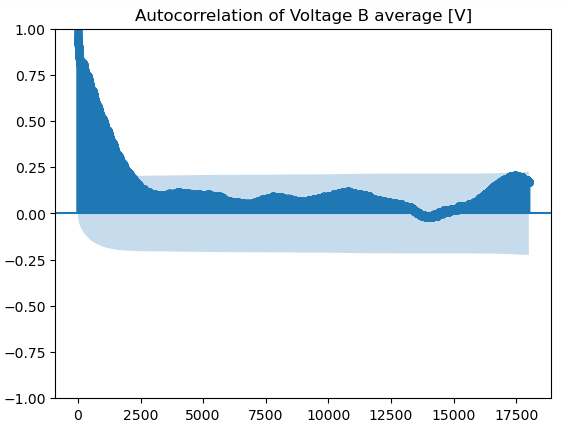
Zbog diferenciranja d = 1.



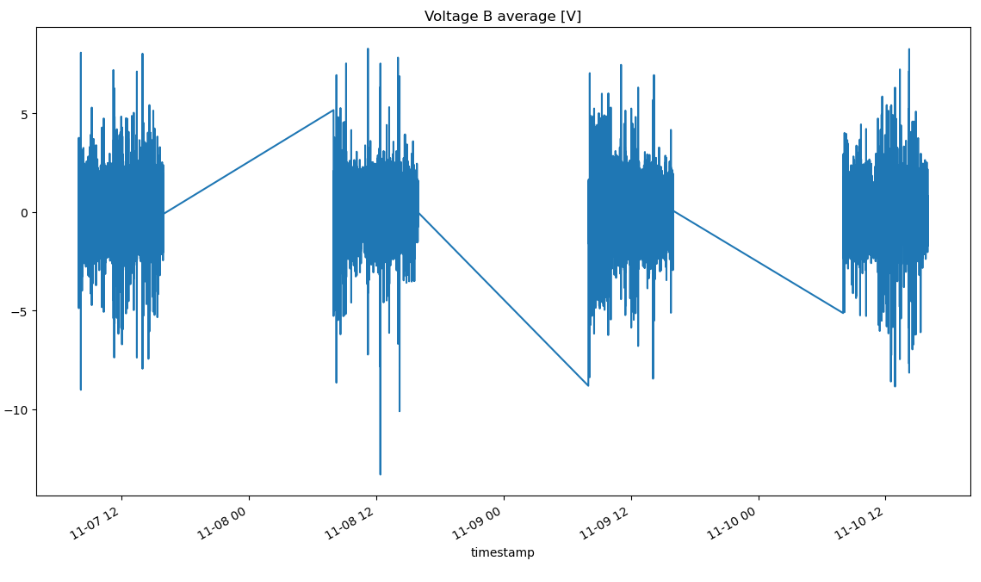
Na osnovu ACF grafika q = 1. Na osnovu PACF grafika p = 1.

### 'Voltage B average [V]'



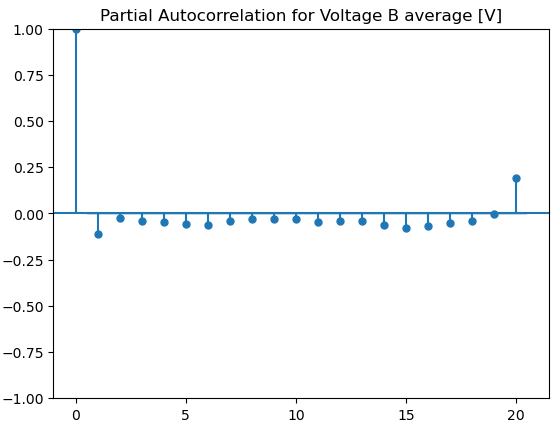
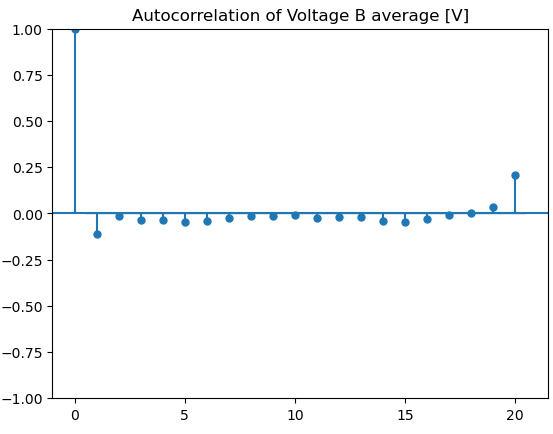


ACF prikazuje preveliki broj lagova kao znacajne. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje.



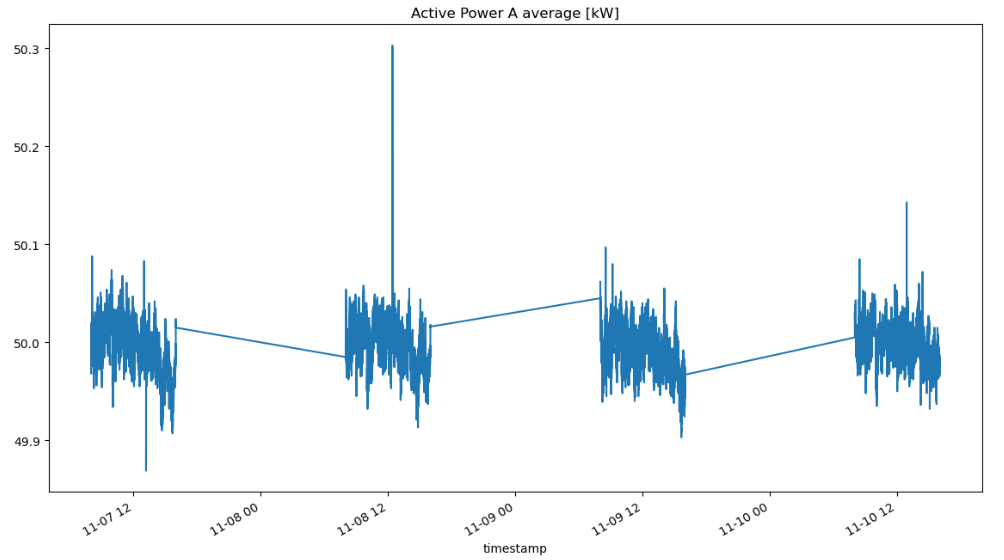
12 Izgled vremenske serije nakon diferenciranja

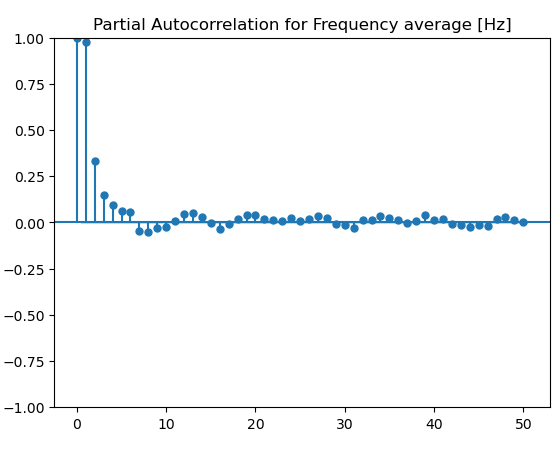
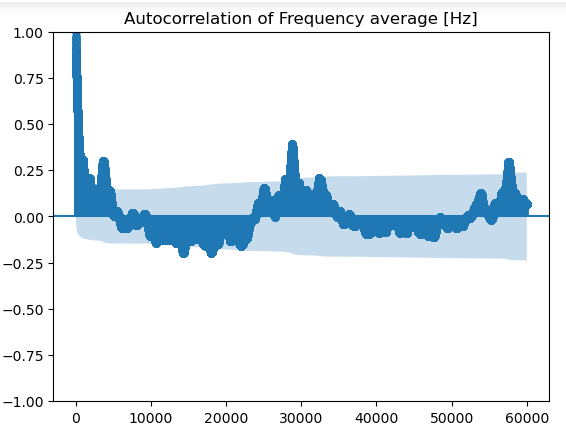
Zbog diferenciranja d = 1.



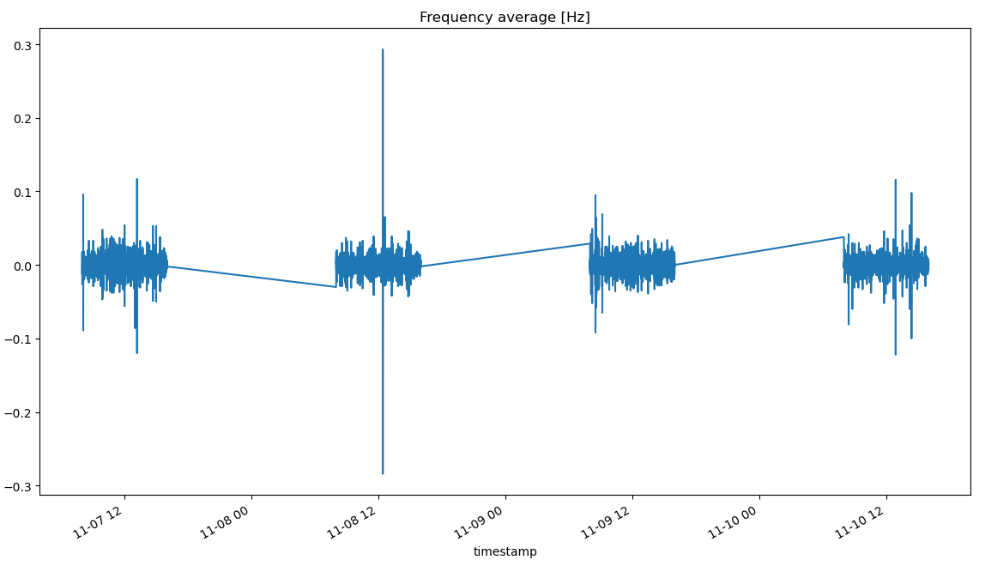
Na osnovu ACF grafika q = 1. Na osnovu PACF grafika p = 1.

### 'Frequency average [Hz]'



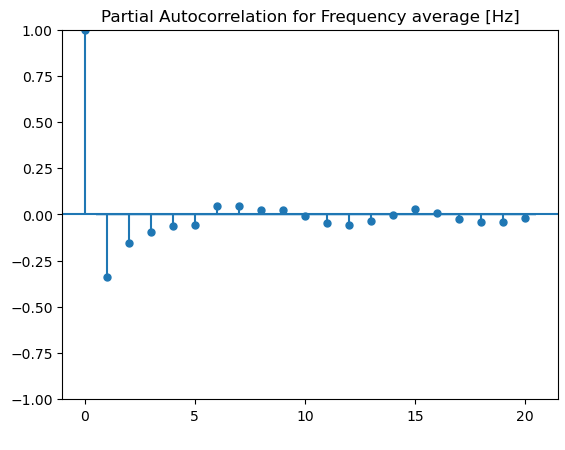
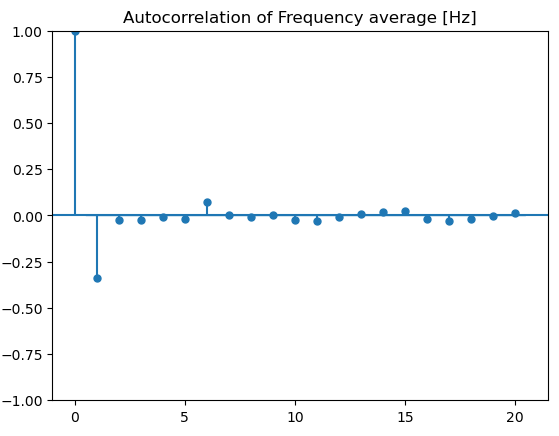


ACF prikazuje preveliki broj lagova kao znacajne, sto ukazuje na nestacionarnu vremensku celiju. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje.



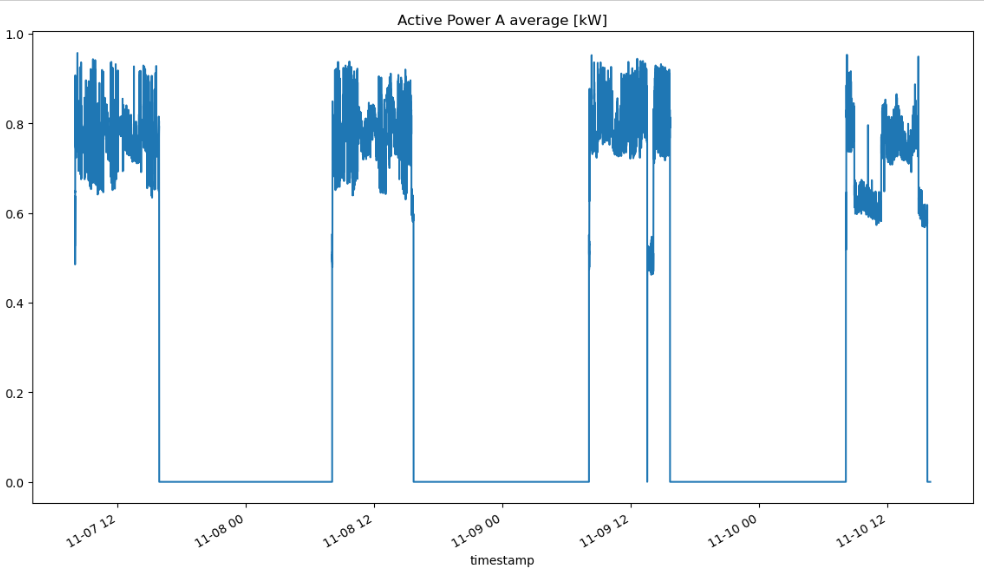
13 Izgled vremenske serije nakon diferenciranja

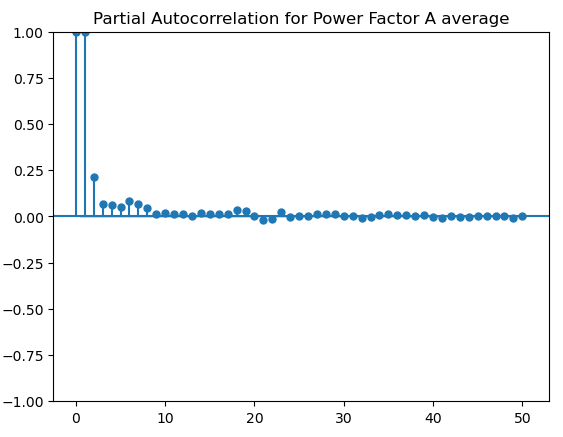
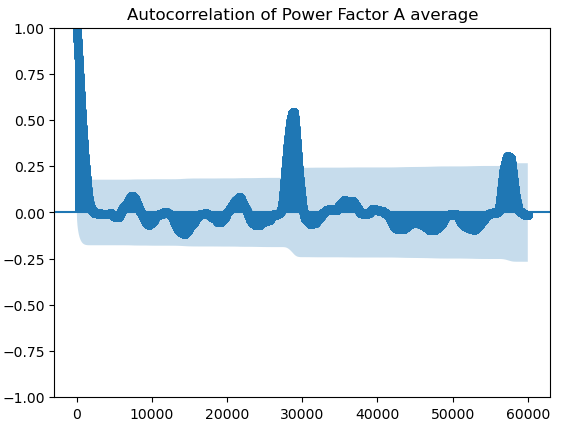
Zbog diferenciranja d = 1.



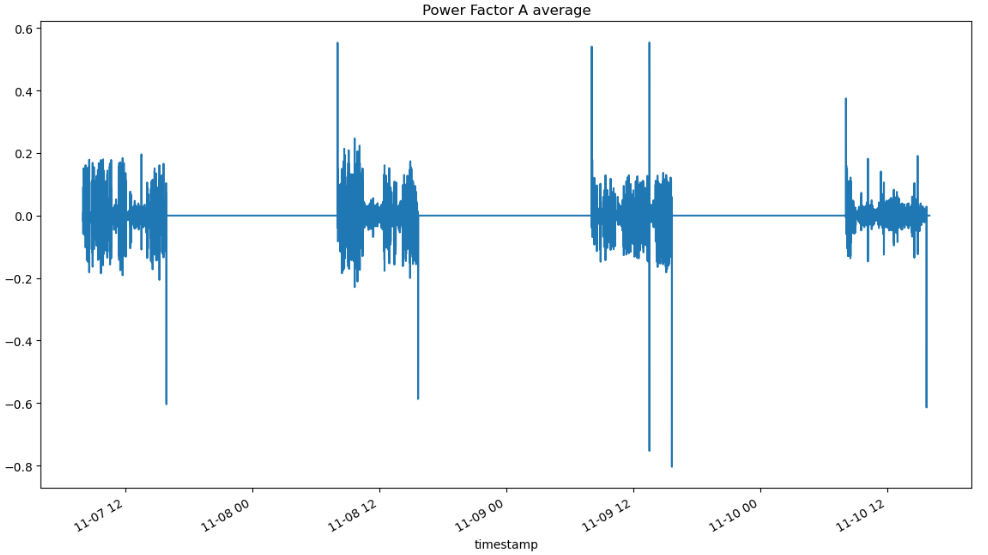
Na osnovu ACF grafika q = 1. Na osnovu PACF grafika p = 1.

### 'Power Factor A average'

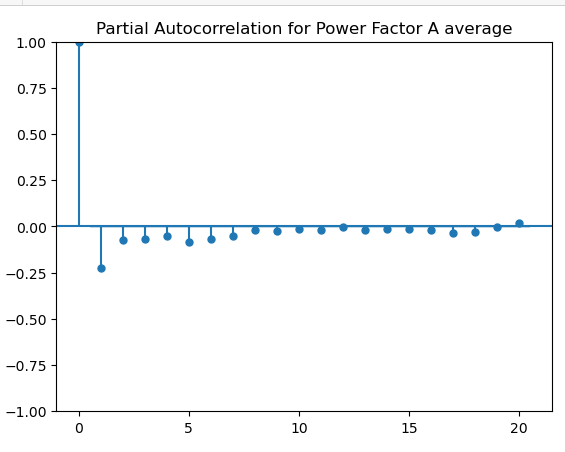
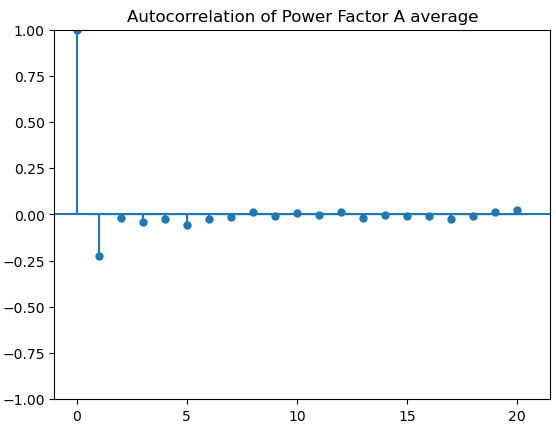




ACF prikazuje preveliki broj lagova kao znacajne, sto ukazuje na nestacionarnu vremensku celiju. Zbog toga ce se na ovu vremensku seriju primeniti diferenciranje.



14 Izgled vremenske serije nakon diferenciranja

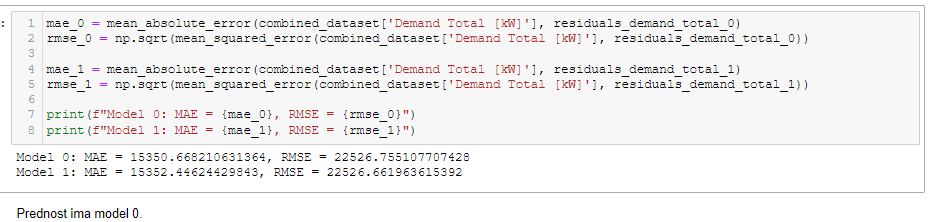


Na osnovu ACF grafika q = 1. Na osnovu PACF grafika p = 1.

## ARIMA model

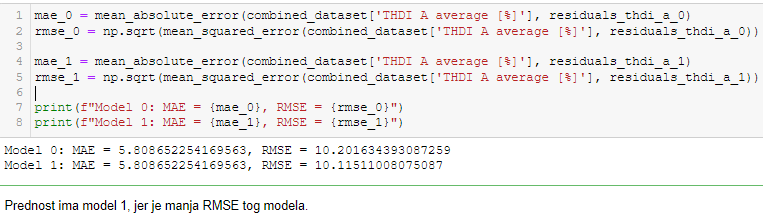
Nad svim vremenskim serijama su primenjeni arima modeli sa parametrima koji su odredjeni na osnovu acf i pacf grafa. Za vremenske serije atributa “Demand Total” i “THDI A” su primenjena dva modela, a onda je korišćenjem mera evaluacije Mean Abosute Error i Root Mean Sqquared Error odabran bolji model od ta dva.

Nad 'Demand Total [kW]' primenjeni su modeli sa parametrima: (0, 2, 1) i (1, 2, 1).



Na osnovu ovih rezultata prednost ima model (0, 2, 1) zbog manjeg Mean Absolute Error parametra.

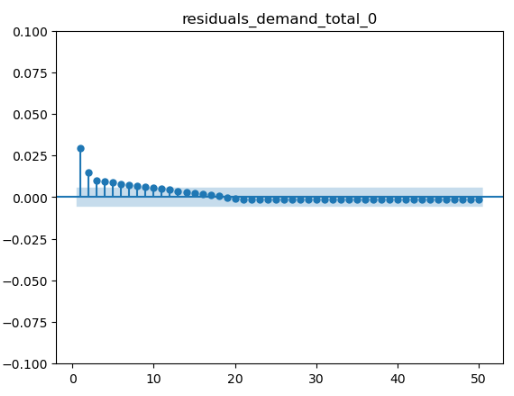
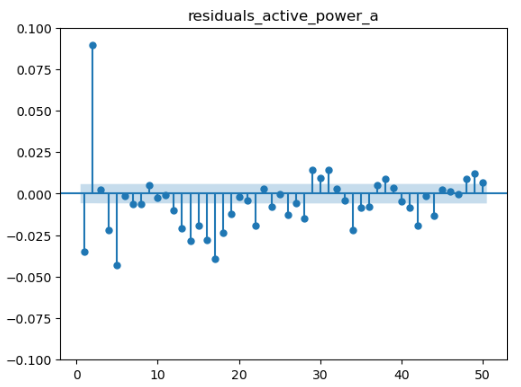
Nad vremenskom serijom 'THDI A average [%]' primenjena su dva modela: (0, 1, 1) i (1, 1, 1).

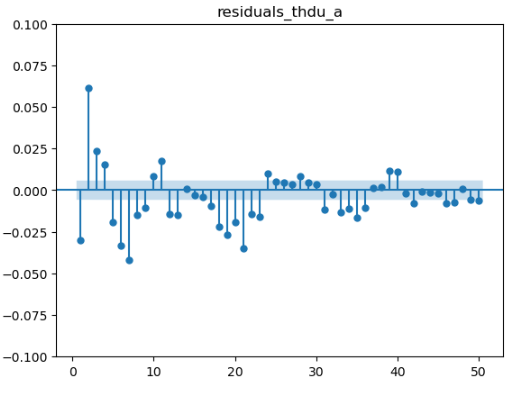
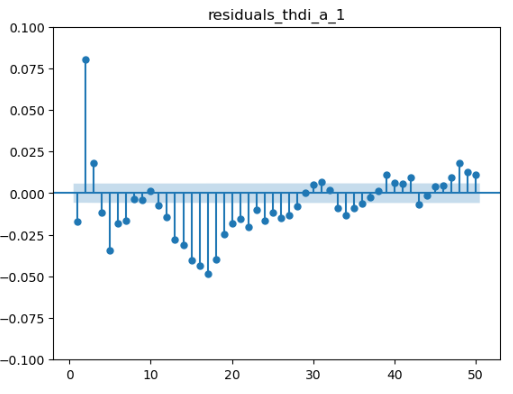


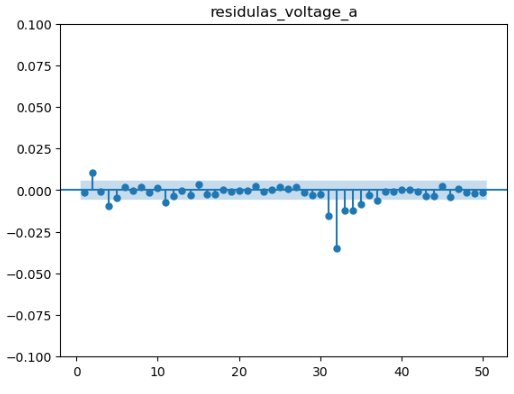
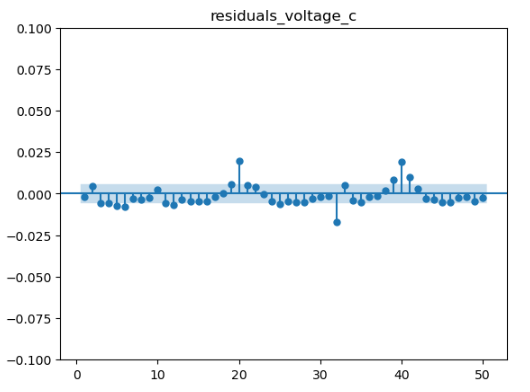
# Evaluacija

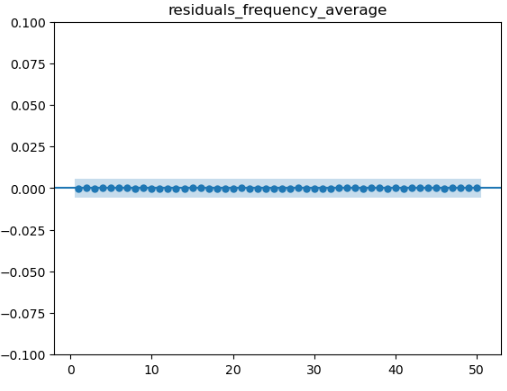
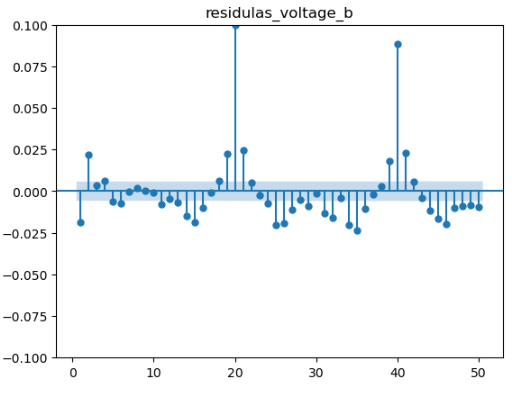
## Korelacija reziduala(autokorelacija)

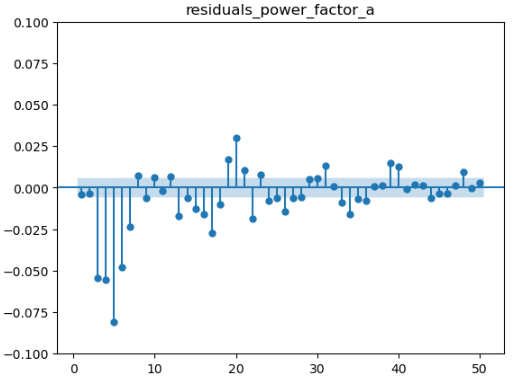
Kao sto se moze videti na ACF grafu korelisanost izmedju lagova je vrlo niska. Ovaj graf je prilagodjen da pokazuje korelisanost za granice od -0.1 do 0.1 da bi grafici bolje izgledali. Iako je korelacija procentualno veoma niska, granica za significant threshold je pomerena previse nisko nakon diferenciranja svake vremenske serije, pa lagovi prelaze significant threshold.







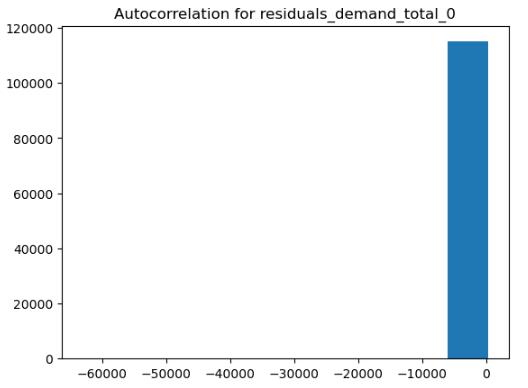
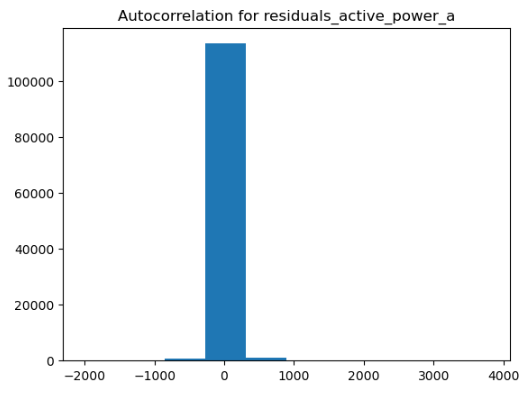




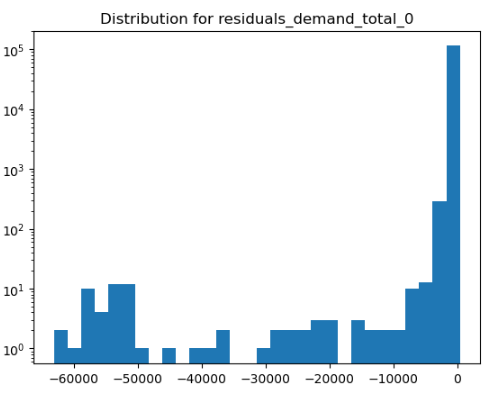
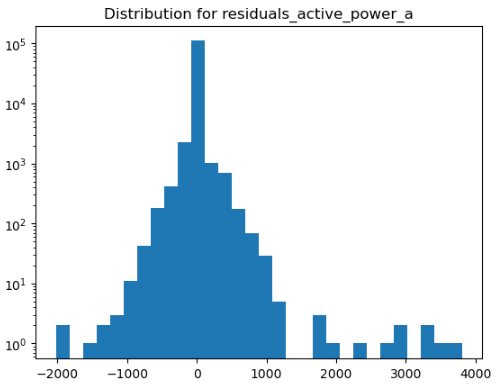
## Distribucija reziduala

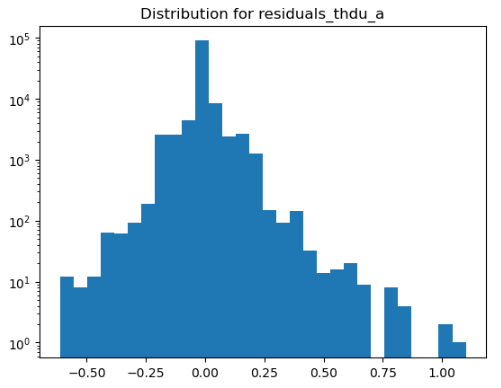
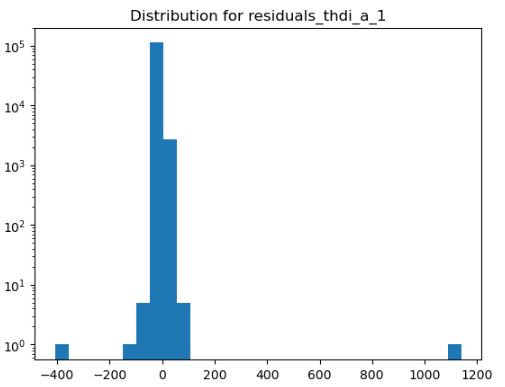
### Histogrami reziduala

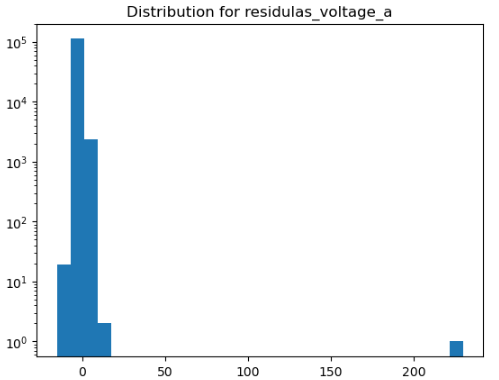
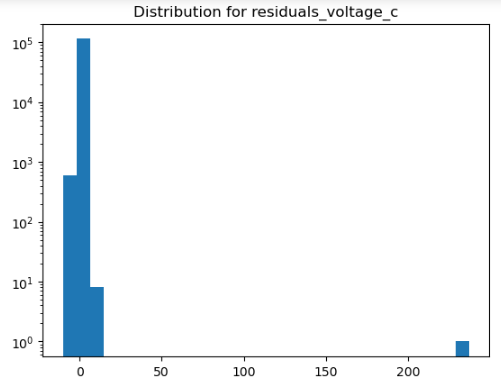
Nakon pravljenja histograma reziduala na linearnoj skali rezultati su loše izgledali. Za skoro svaki atribut na histogramu reziduala je bio prikazan samo uspravni stub. Primeri:

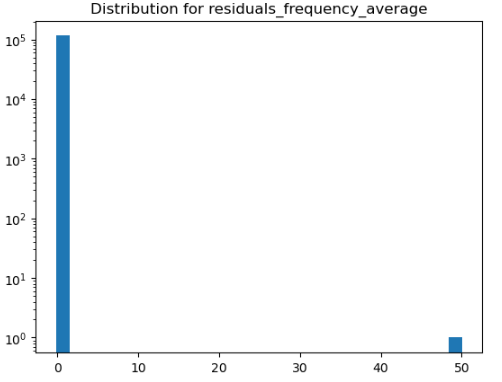
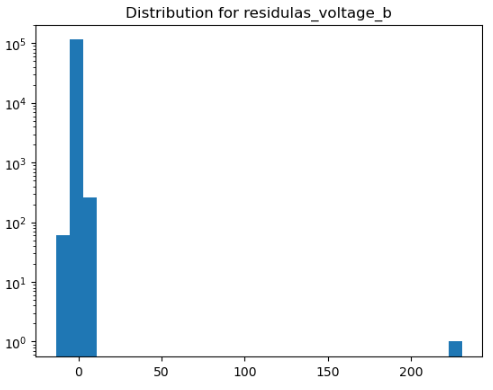


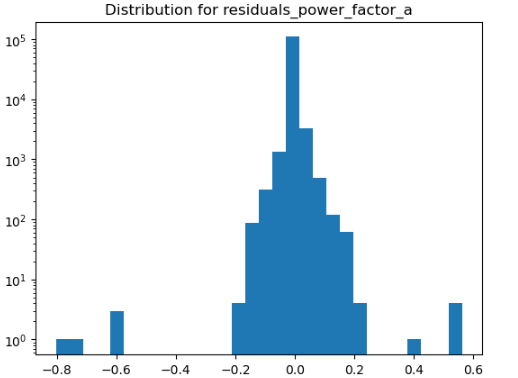
Zbog je za histograme reziduala umesto linearne skale korišćena logaritamska skala.





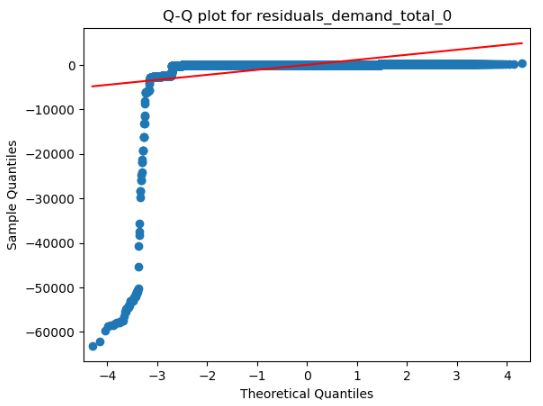
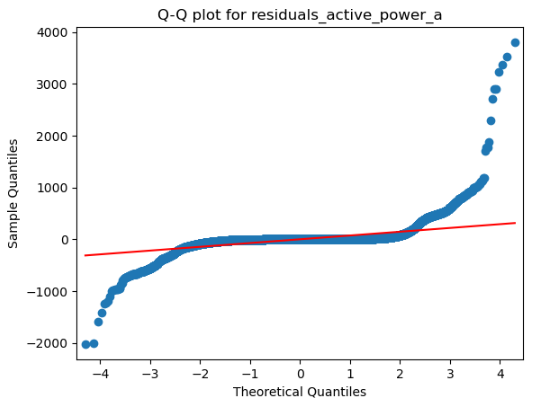


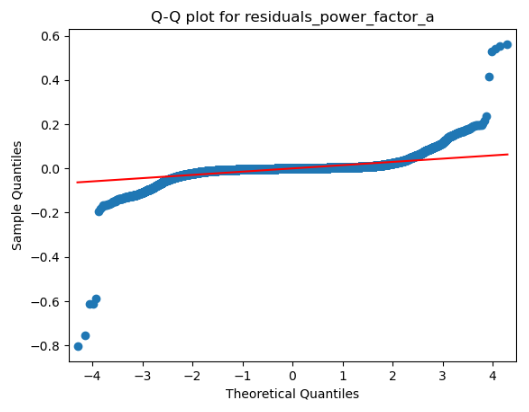
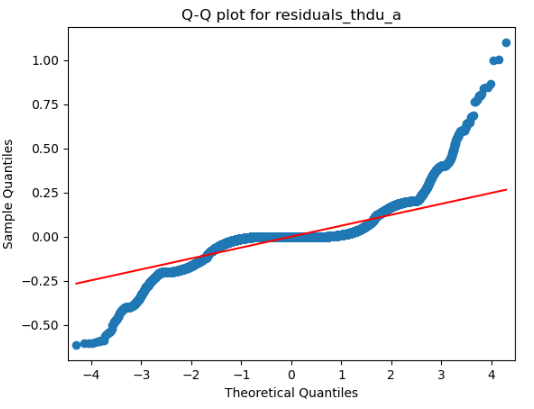




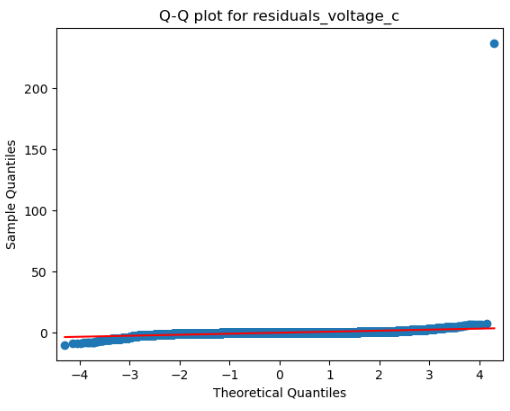
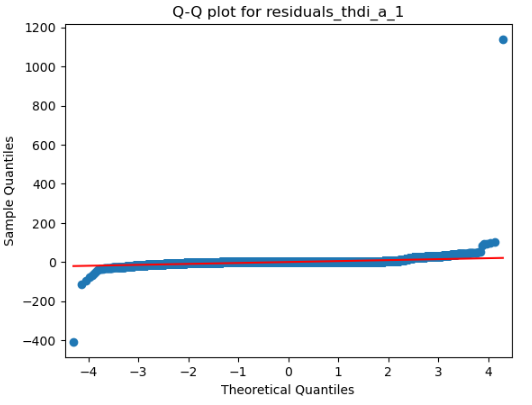
Na osnovu histograma reziduala vremenskh serija teško je zaključiti vizuelno da li vremenske serije prate normalnu distribuciju. Bolji uvid se dobija posmatranjem Q-Q plotova, koji su dati u nastavku.

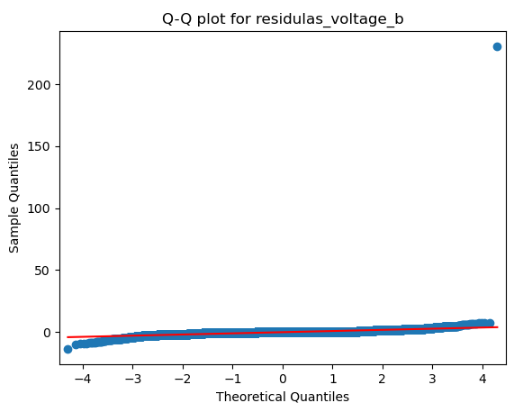
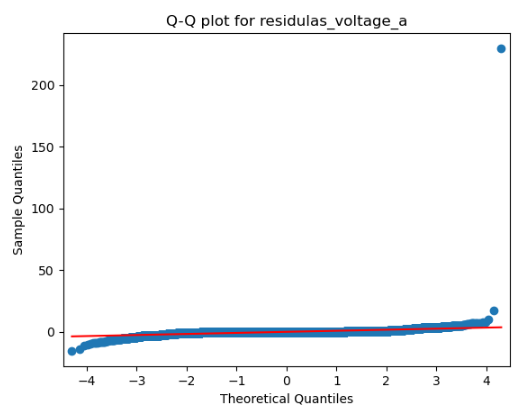
### Q-Q plot

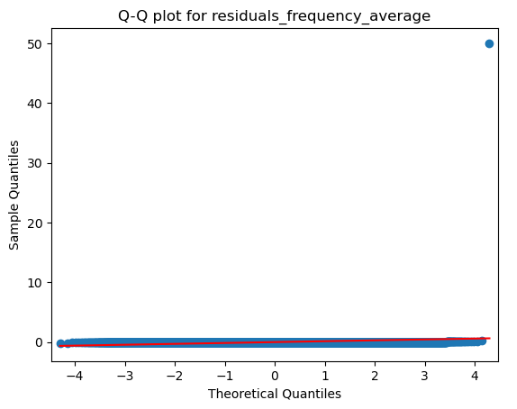




Na prethodne 4 slike su prikazani Q-Q plotovi reziduala za vremenske serije kod kojih distribucija reziduala ne prati normalnu raspodelu. To su reziduali za sledeće vremenske serije: 'Active Power A average [kW]', 'Demand Total [kW]', 'THDU A average [%]', 'Power Factor A average'.







Na prethodnih 5 slika su prikazani reziduali vremenskih serija koji prate normalnu raspodelu, a to su: 'THDI A average [%]', 'Voltage C average [V]', 'Voltage A average [V]', 'Voltage B average [V]', 'Frequency average [Hz]'.