Projekat iz predmeta: Tehnike i metode analize podataka

Tema: Analiza Environmental data - Data Exploration

Zadatak: Primena raznih metoda koje se tiču istraživanja vremenskih serija nad „environmental parameters” podacima

Student: Damjan Ćupić 1700

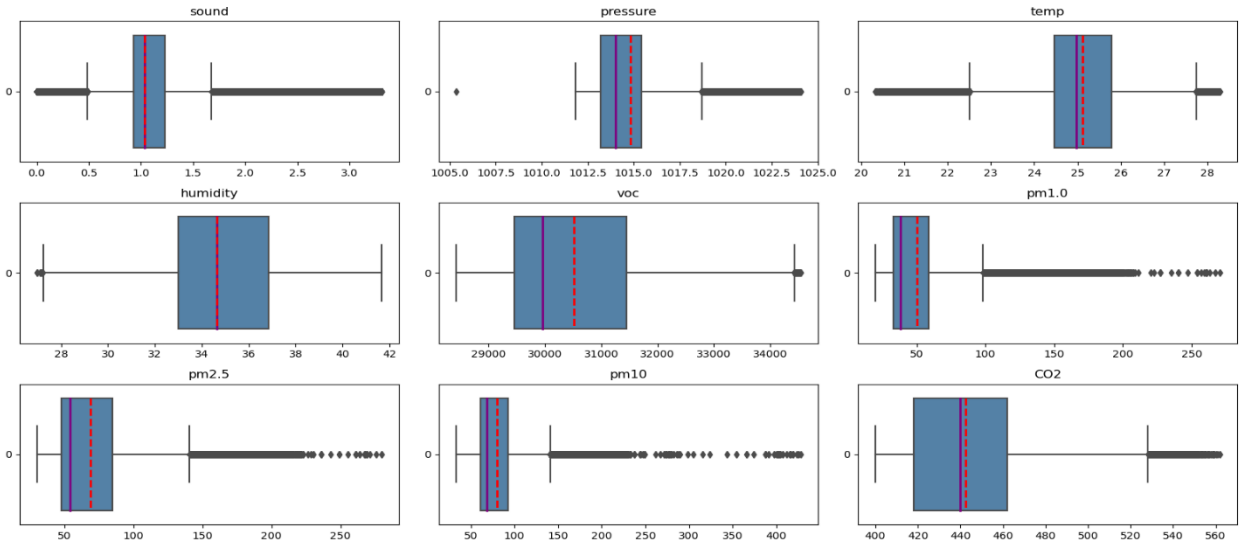
Mentori: Bratislav Trojić – Nissatech, prof. dr Suzana Stojković

Prvi korak:

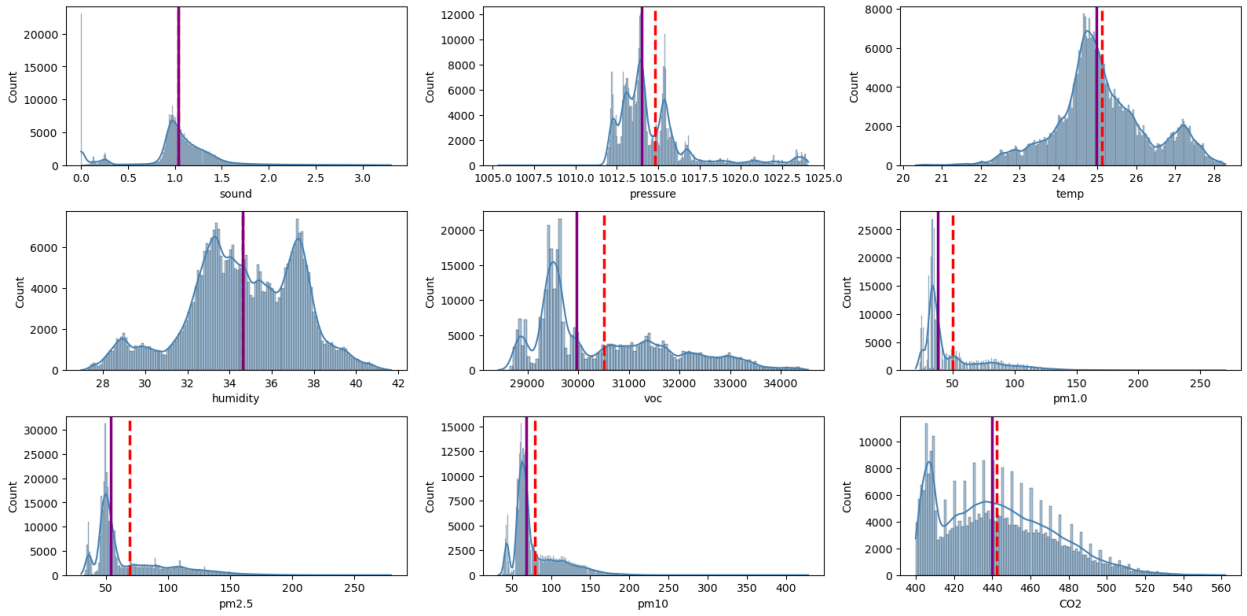
Nakon inicijalnog učitavanja, odrađeno je osnovno istraživanje nad podacima primenom odgovarajućih funkcija. Učitavanjem skupa podataka dobijeno je 345151 redova i 9 atributa, gde je *timestamp* korišćen za indeksiranje. Svi atributi su numeričkog tipa, a osnovni podaci o njihovim vrednostima su dati u sledećoj tabeli:

	sound	pressure	temp	humidity	voc	pm1.0	pm2.5	pm10	CO2
count	345151.000000	345151.000000	345151.000000	345151.000000	345151.000000	345151.000000	345151.000000	345151.000000	344264.000000
mean	1.035109	1014.807431	25.123033	34.636160	30512.354483	49.980410	68.674444	79.883421	442.465660
std	0.493434	2.517912	1.210600	2.652997	1329.625356	27.035828	31.107893	30.668957	28.831611
min	0.000000	1005.339453	20.326285	26.962989	28427.000000	20.000000	30.000000	33.000000	400.000000
25%	0.931100	1013.215952	24.470044	32.984199	29458.000000	33.000000	48.000000	61.000000	418.000000
50%	1.035990	1014.020586	24.972642	34.628268	29964.500000	38.000000	54.000000	68.000000	440.000000
75%	1.229633	1015.415069	25.782958	36.830740	31443.000000	59.000000	85.000000	93.000000	462.000000
max	3.303227	1024.071917	28.293446	41.651724	34535.000000	270.500000	279.500000	428.000000	562.000000

Zatim su nacrtani *boxplot* i *histplot* dijagrami kako bi se stekao uvid u raspodelu vrednosti za svaki atribut i uočili potencijalni *outlier*-i:

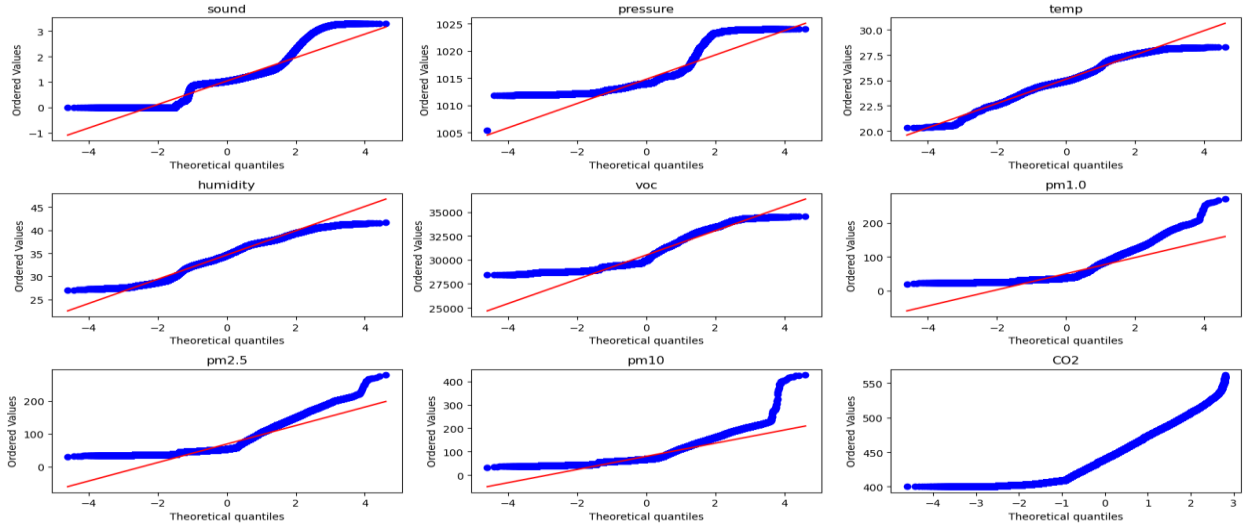


Sa *boxplot* dijagrama se može videti da atributi, sem *humidity* i *voc*, imaju potencijalno veliki broj vrednosti koje odstupaju od ostalih, što je posebno izraženo kod atributa vezanih za *pm* čestice. Kod atributa *sound* i *humidity* postoji poklapanje između srednje vrednosti i medijane, što može ukazati na simetričnu raspodelu podataka.

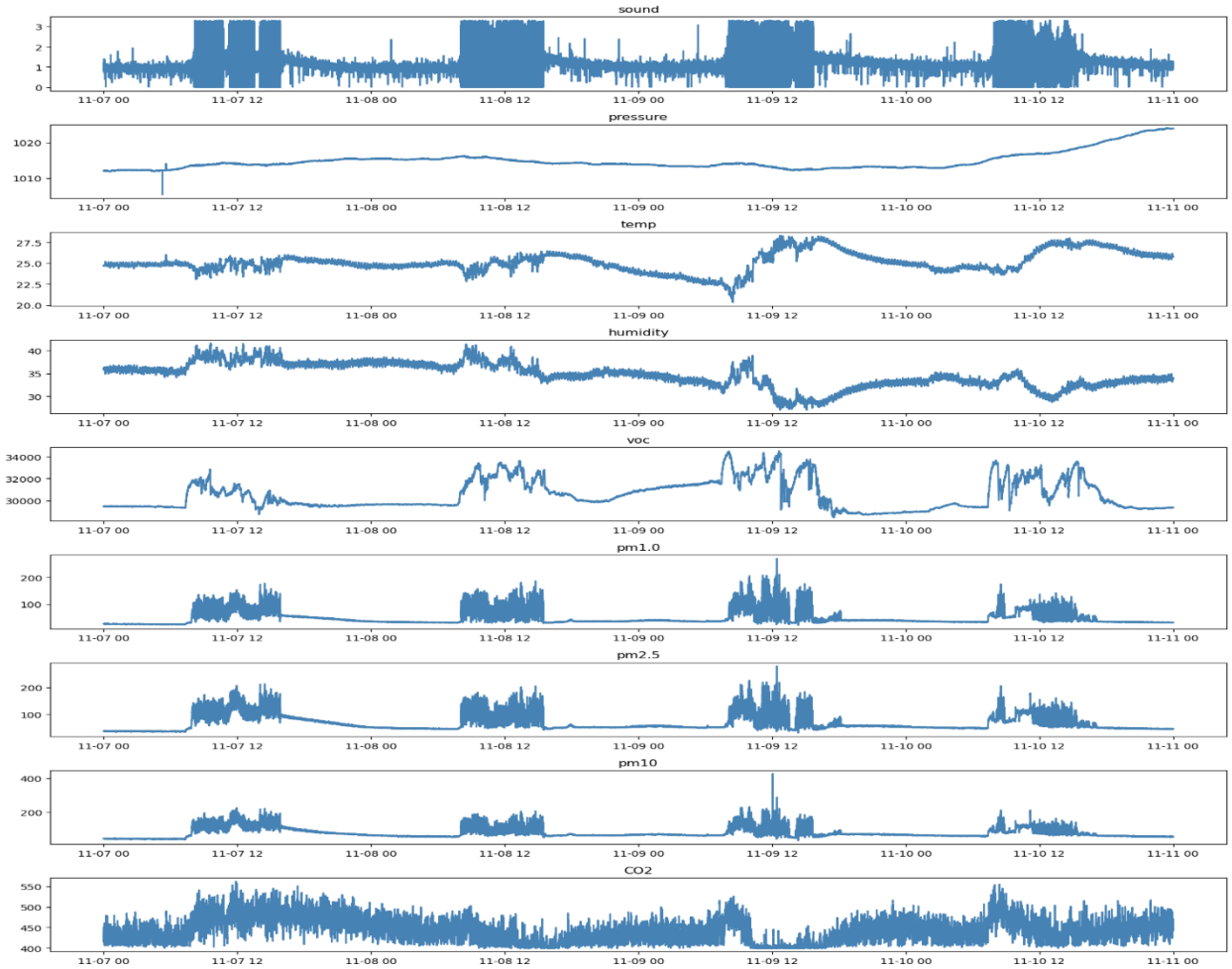


Sa histograma se može videti da kod *pm* čestica zaista postoje vrednosti koje znatno odstupaju od središnje (desna strana dijagrama). Takođe, na dijagramima se primećuje da nijedan atribut nema normalnu raspodelu vrednosti, ali da je raspodela *sound*, *temp* i *humidity* približnija normalnoj u poređenju sa ostalim atributima.

Na slici ispod je prikazan *Q-Q plot*, pomoću kojeg se raspodela vrednosti atributa upoređuje sa normalnom (najbolje poklapanje je za atribut *temp*). Postupak je ponovljen i za uniformnu i eksponencijalnu raspodelu.



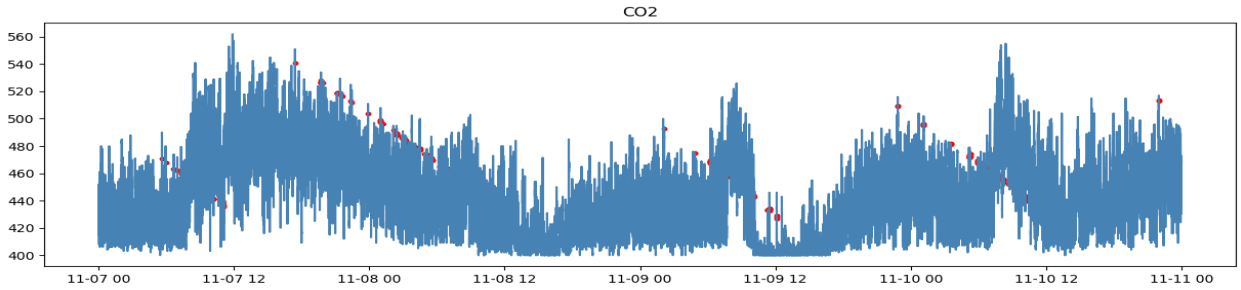
Nakon toga su nacrtani grafici na kojima su predstavljene vrednosti na vremenskoj osi za svaki atribut:



Sa ovih grafika se uočava sezonalnost vrednosti koja odgovara jednom celom danu. Može se zaključiti da se zadati proces obavlja svakog dana približno između 8 i 16 časova. Kod atributa *pressure* uočava se rastući trend koji je primetan poslednjeg dana.

Drugi korak:

Na početku ovog koraka su proverene nedostajuće vrednosti i uočeno je da atributu CO2 nedostaje 887 vrednosti. Nedostajuće vrednosti su popunjene primenom linearne interpolacije:



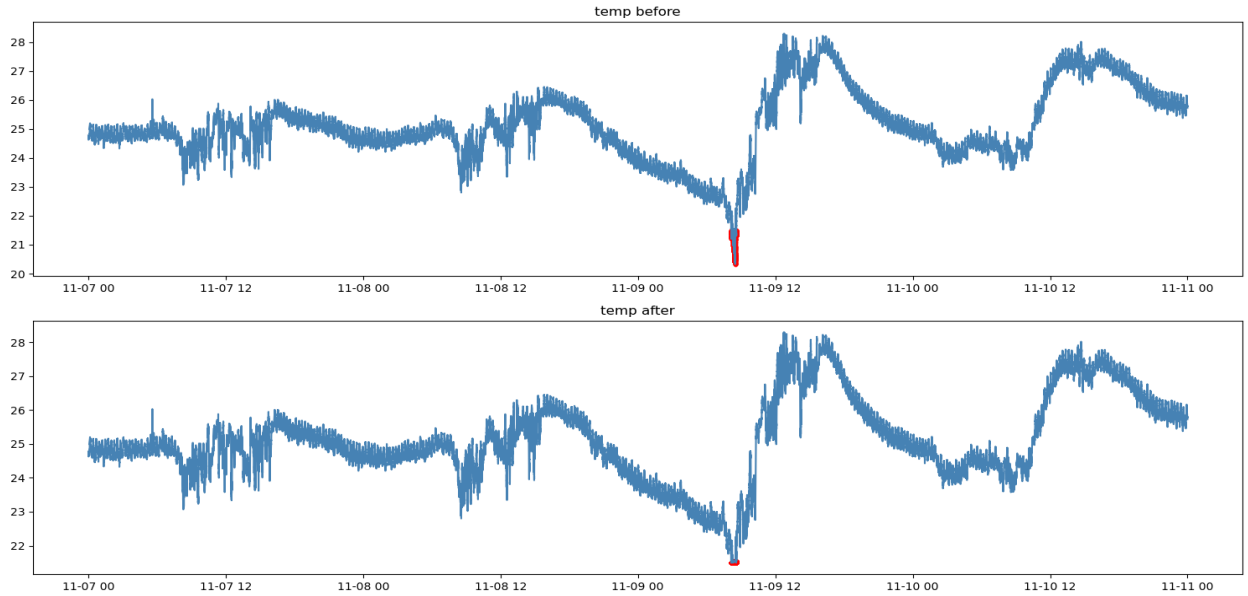
Nakon toga provereno je da li u skupu podataka postoje redundantni redovi. Detektovano je da postoji 24 reda sa istom vrednošću indeksa, koji su uklonjeni usrednjavanjem odgovarajućih vrednosti. Zatim je odrađeno odmeravanje vremenskog signala na jednu sekundu.

Za svaki od atributa provereno je da li se nad njim može primeniti „3-sigma pravilo“ („empirijsko pravilo“) za detekciju odstupajućih vrednosti. Ta tehnika je primenjena nad atributima *temp*, *humidity* i *voc*, s obzirom na to da su oni jedini koji su (približno) zadovoljili uslove da im se 68% vrednosti nalazi u opsegu $\mu \pm \sigma$, 95% u opsegu $\mu \pm 2*\sigma$ i 99.7% u opsegu $\mu \pm 3*\sigma$. Nad ostalim atributima je za detekciju odstupajućih vrednosti primenjena IQR metoda, pri čemu je IQR = Q3 - Q1, gde su Q1 i Q3 25. i 75. percentil, respektivno.

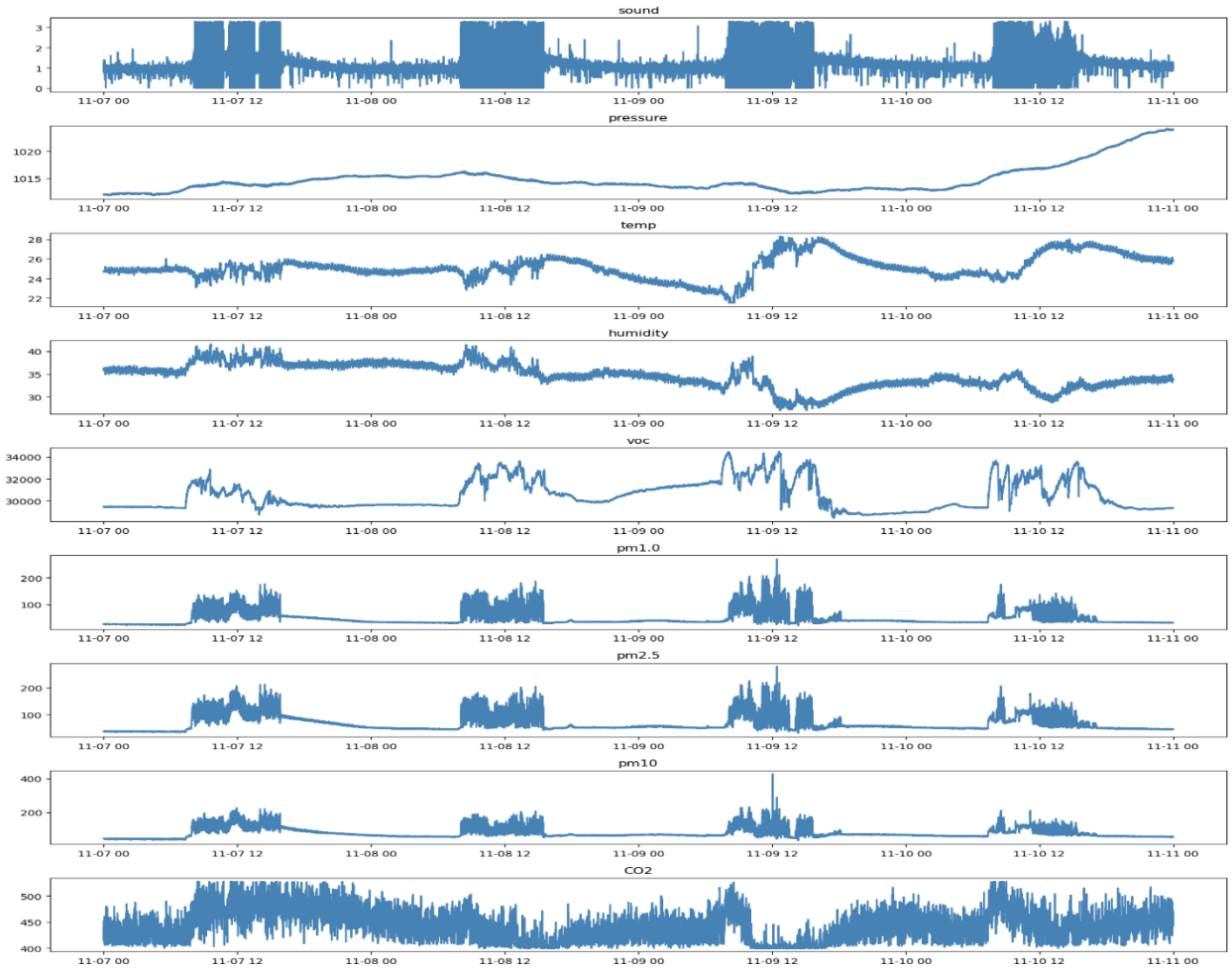
Kod nekih parametara nisu detektovani *outlier*-i (*humidity*), dok je kod nekih bio prisutan veoma veliki broj (npr. 65015 kod *sound*). Kod atributa kod kojih su odstupajuće vrednosti detektovane IQR tehnikom trebalo bi odbaciti značajan broj instanci, tako da je u nastavku rađeno sa originalnim vrednostima kako se ne bi izgubile bitne informacije (jedino je kod atributa *CO2* odbačeno 1269 vrednosti). Pored toga, na vremenskom dijagramu atributa *pressure*, mogu se uočiti dve vrednosti koje su značajno manje (07.11. u 05:35:21), odnosno veće (07.11. u 05:15:54) od okolnih, pri čemu je prva detekovana i IQR metodom, a druga vizuelnom analizom vremenskog dijagrama (najverovatnije je reč o greškama pri merenju).

Vrednosti koje se od središnje razlikuju za više od $3*\sigma$, odnosno koje se nalaze van opsega [Q1 - 1.5*IQR, Q3 + 1.5*IQR] (*outlier*-i) zapravo nisu odbačene (kako se ne bi prekidala vremenska serija), već su ažurirane primenom tehnike linearne interpolacije. Nakon toga je kod nekih atributa i dalje ostao određeni broj odstupajućih vrednosti.

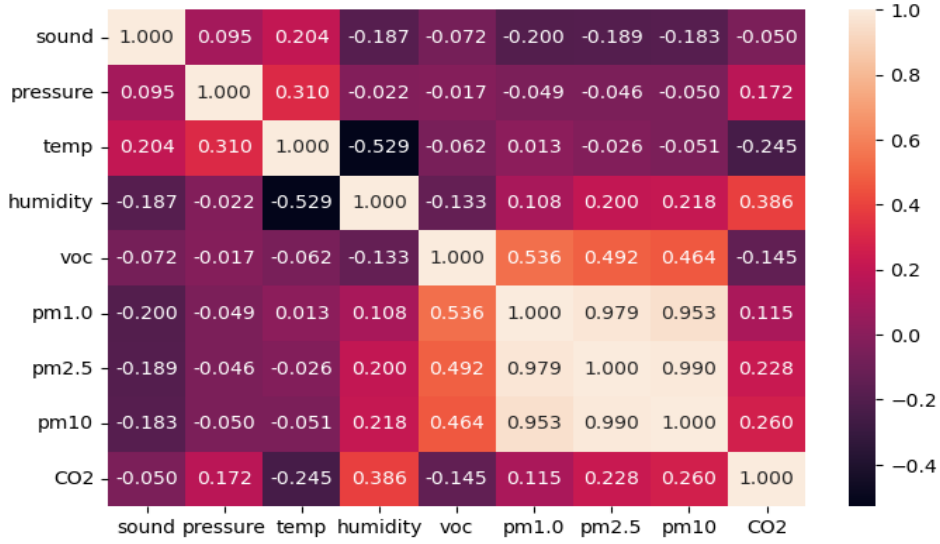
Na narednoj slici je na primeru signala *temp* prikazano kako izgledaju odstupajuće vrednosti pre i posle izmene. Detektovano je 834 vrednosti koje su odstupale od signala, a nakon ažuriranja ostalo ih je 106.



Izgled svih atributa nakon otklanjanja *outlier*-a:



Na slici je prikazana *heatmap-a* kojom je prikazana korelacija između atributa. Može se primetiti da postoji visoka korelacija između *pm1.0*, *pm2.5* i *pm10*, srednje negativna korelacija između *temp* i *humidity*, kao i srednje pozitivna korelacija između *voc* i svih *pm* čestica.

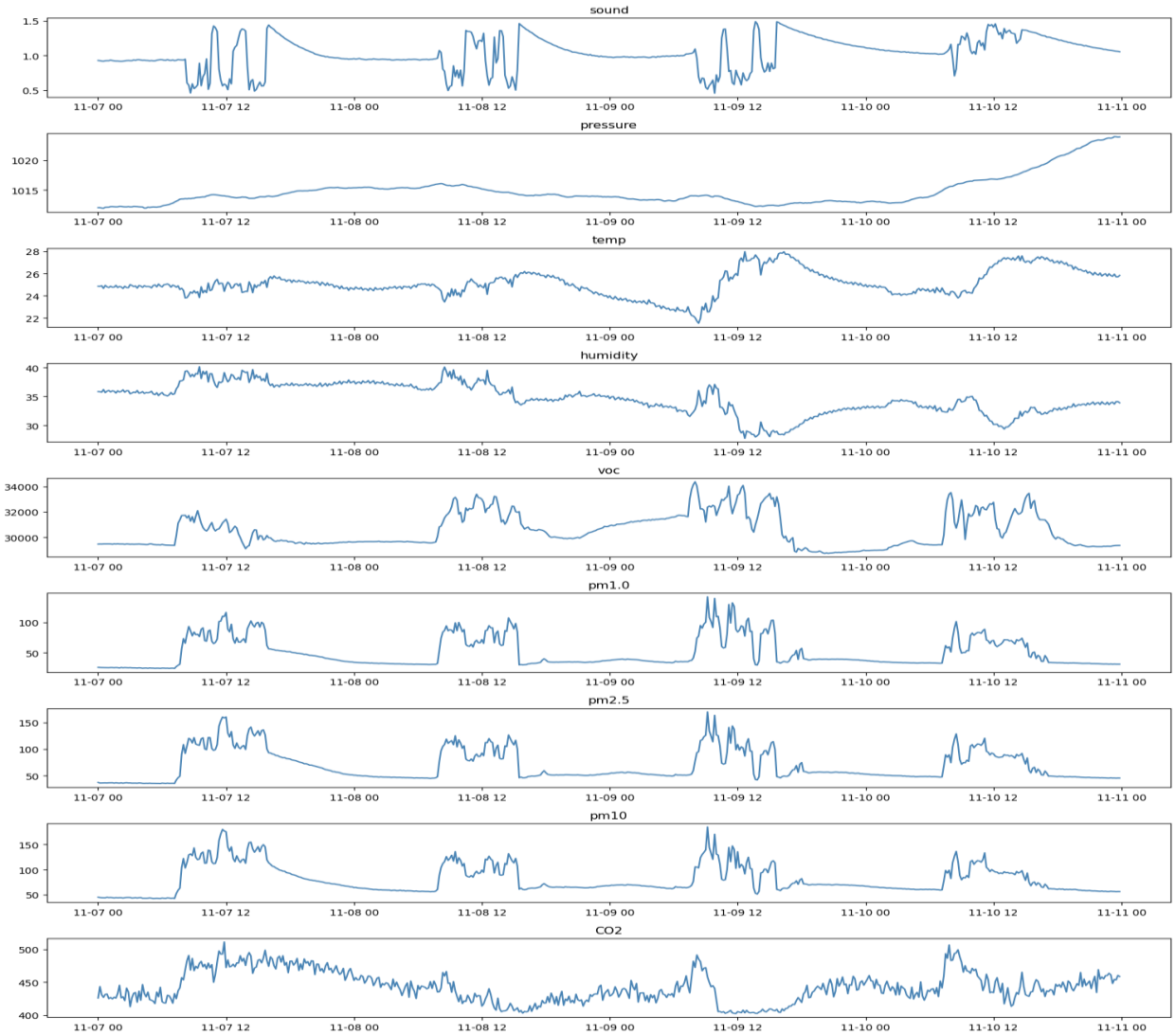


Treći i četvrti korak:

U nastavku je rađeno sa podacima koji su odmeravani na 10 minuta, zato što je izvršavanje algoritama nad podacima uzorkovanim na jednu sekundu trajalo predugo, ili nije moglo da se izvrši usled nedostatka resursa na računaru. Prvih pet podataka iz novog skupa prikazano je u sledećoj tabeli:

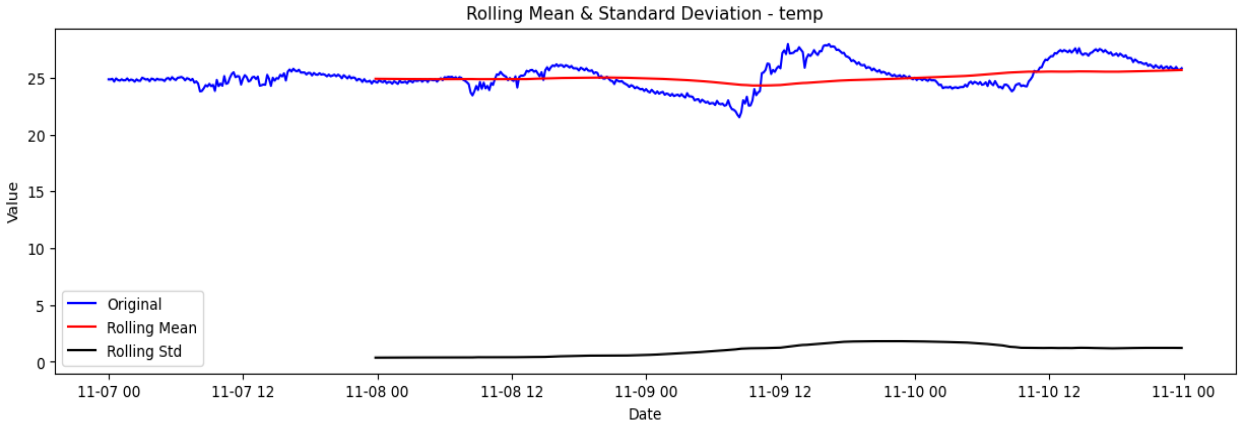
timestamp	sound	pressure	temp	humidity	voc	pm1.0	pm2.5	pm10	CO2
2022-11-07 00:00:00	0.928750	1012.065458	24.850302	35.857482	29446.195326	25.854758	37.353088	44.924875	425.767112
2022-11-07 00:10:00	0.925973	1012.068418	24.843975	35.837883	29454.563333	25.651667	36.292500	44.330000	443.254167
2022-11-07 00:20:00	0.920606	1011.998360	24.897359	35.751384	29449.128333	25.535833	36.405833	43.887500	430.708333
2022-11-07 00:30:00	0.918829	1011.966773	24.662044	36.203402	29459.695000	25.512500	36.616667	43.861667	426.738333
2022-11-07 00:40:00	0.924403	1012.119962	24.929130	35.676675	29461.610833	25.426667	36.548333	43.742500	431.318333

Vremenski dijagram ovih signala:

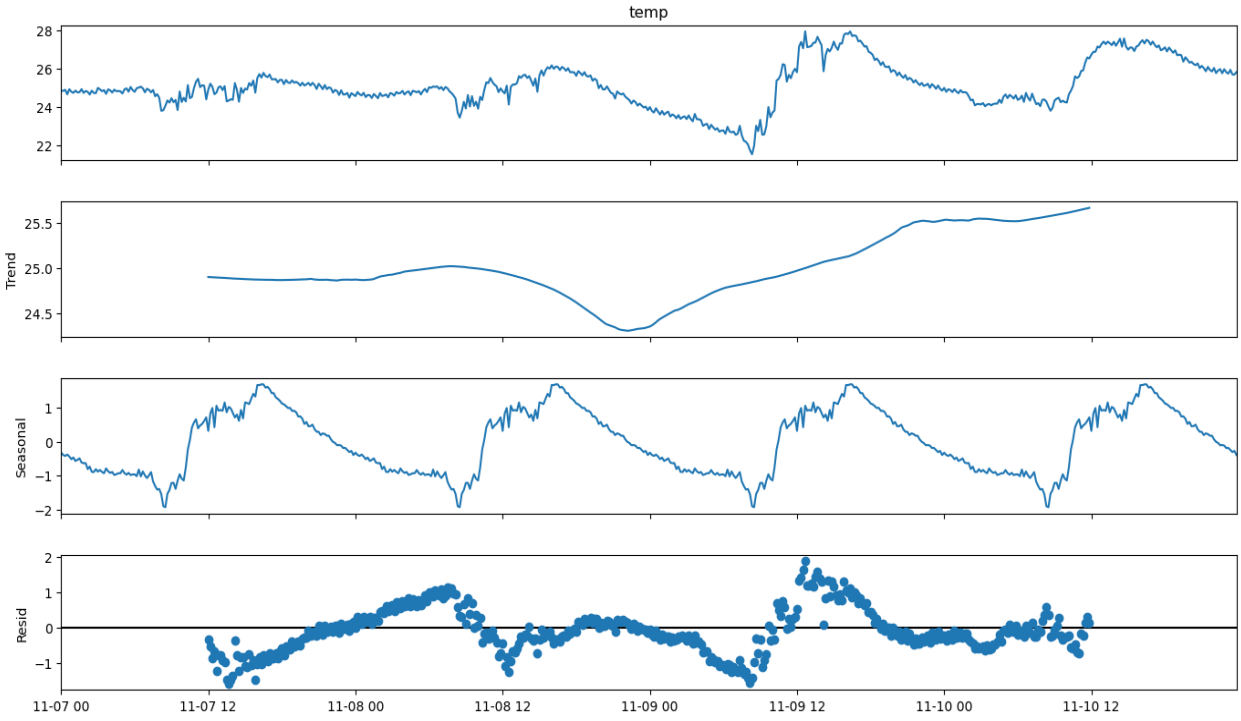


Nakon toga je nad svakim atributom primenjen isti postupak, koji se sastoji posmatranja komponenti dobijenih dekompozicijom signala, podele skupa na trening i test deo, provere stacionarnosti i korelisanosti signala, i na kraju primene ARIMA (AutoRegressive Integrated Moving Average) modela. Za svaki segment analize kreirane su odgovarajuće funkcije, a ovaj proces je ilustrovan na primeru atributa *temp*.

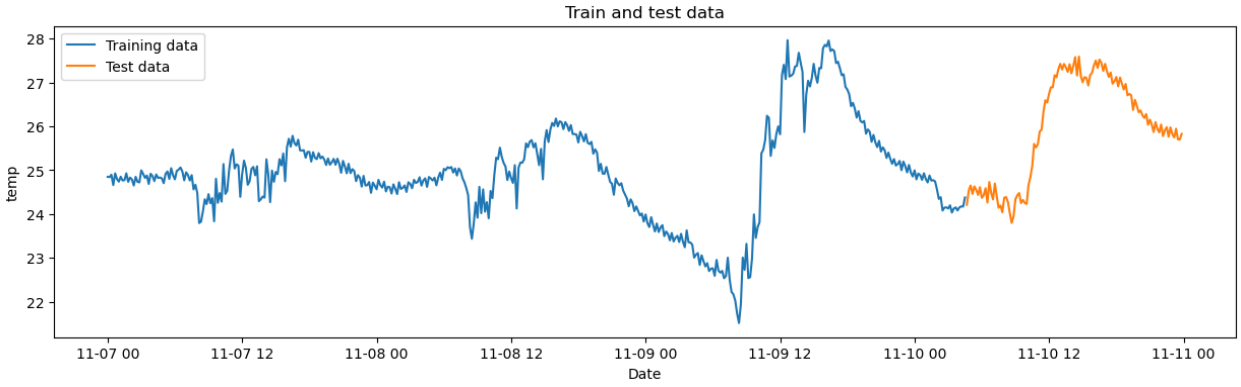
Na sledećoj slici je prikazan grafik na kome se vidi kakve su vrednosti za *mean* i *std* kada se koristi „klizajući prozor“ od jednog dana. Primećuje se da ove vrednosti nisu u potpunosti konstantne, već imaju male varijacije.



Nakon toga je odrađena dekompozicija signala na tri komponente: trend, sezonsku i rezidualnu. Može se primetiti da je prisutna sezonalnost, kao i da je postojao trend opadanja, nakon koga sledi trend blagog rasta.

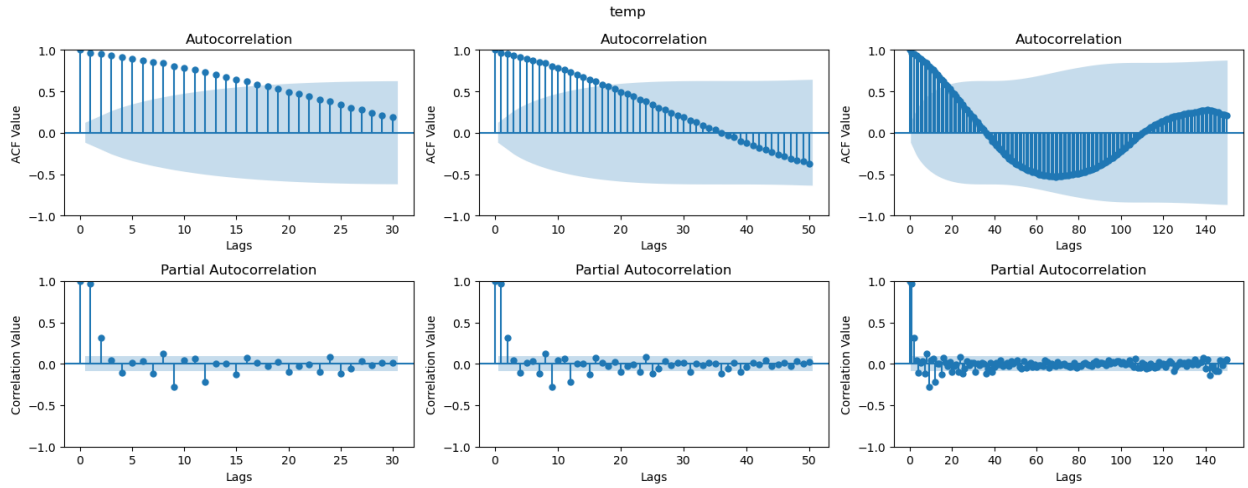


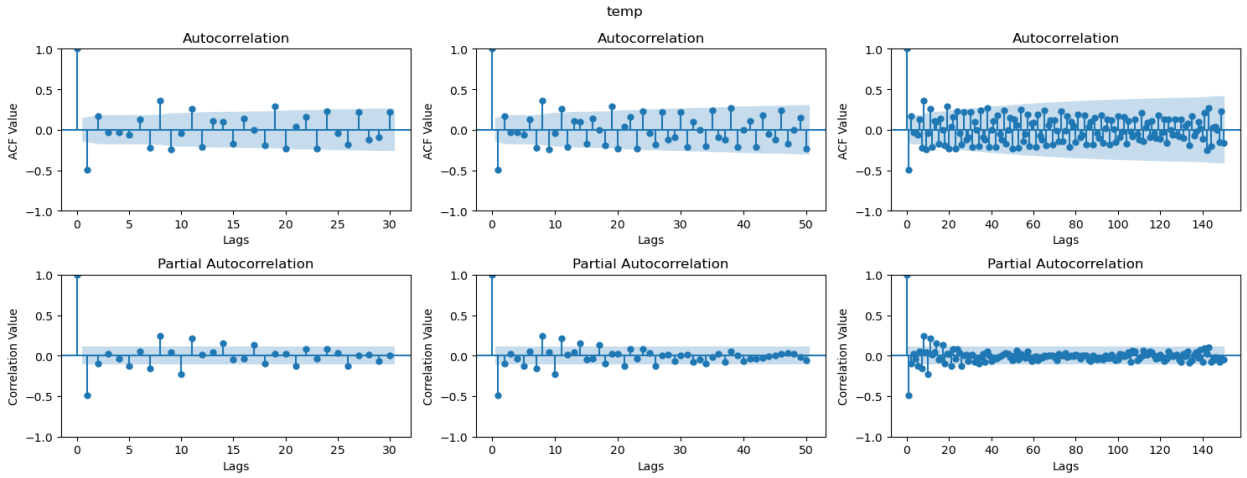
Podaci su podeljeni tako da 80% pripada treningu skupu, a preostalih 20% test skupu, tako da su približno tri dana korišćena za treniranje, a jedan dan za testiranje modela. Na narednoj slici je prikazano kako izgleda ova podela za *temp*:



Zatim je proverena stacionarnost primenom statističkih testova ADF (Augmented Dickey-Fuller) i KPSS (Kwiatkowski-Phillips-Schmidt-Shin). Kod ADF testa inicijalna hipoteza je da signal nije stacionaran, a kao izlaz testa dobija se *p* vrednost. Ukoliko je ova vrednost manja od zadatog praga (najčešće 0.05), inicijalna hipoteza se odbacuje, odnosno može se zaključiti da je signal stacionaran. Kod KPSS testa inicijalna hipoteza je obrnuta – da signal jeste stacionaran. U tom slučaju se zaključuje da je test nestacionaran ukoliko je *p* vrednost manja od praga. Za primer atributa *temp* oba testa su ukazala na to da je signal stacionaran, međutim, s obzirom na to da postoji sezonska komponenta, odrađeno je sezonsko diferenciranje. Nakon toga su testovi pokazali da je signal nestacionaran, pa je urađeno i nesezonsko diferenciranje, nakon čega je signal bio stacionaran.

Još jedan od pokazatelja stacionarnosti mogu biti grafici na kojima su prikazane autokorelacija (ACF, Autocorrelation Function) i parcijalna autokorelacija (PACF, Partial Autocorrelation Function). Na narednim slikama je prikazano kako oni izgledaju pre i posle primenjenog diferenciranja. Na prvom ACF grafiku se može primetiti da korelacija između vrednosti signala sporo opada (što može biti znak nestacionarnosti), dok na drugom to više nije slučaj.



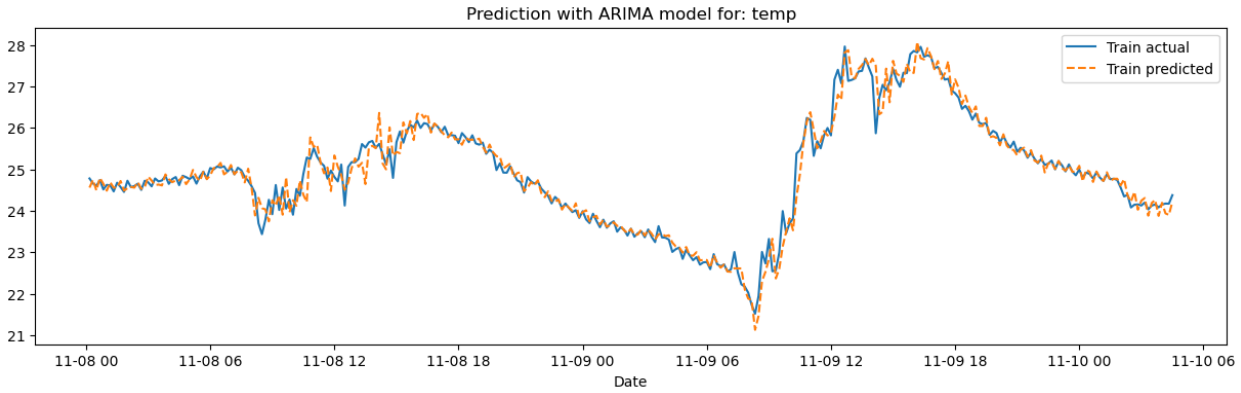


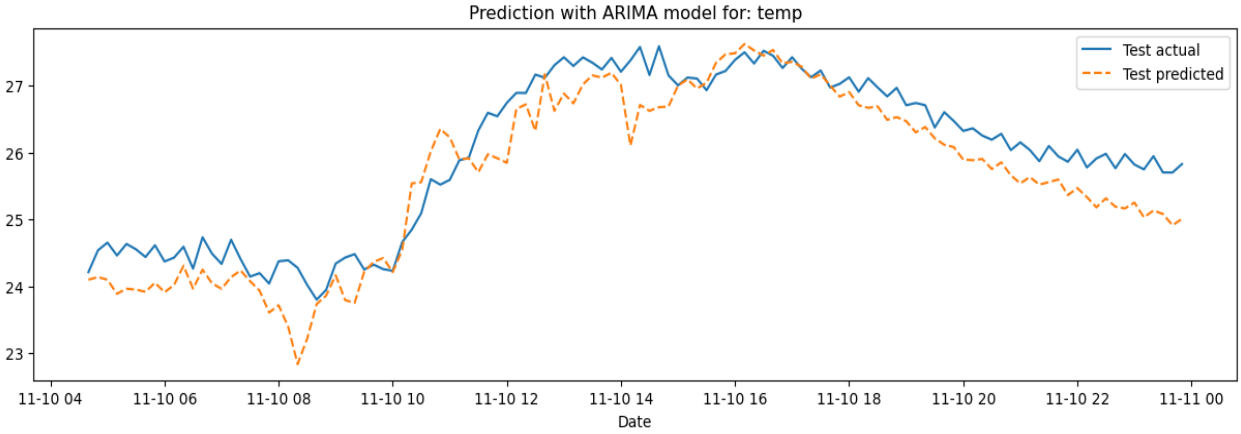
Sledeći korak podrazumevao je odabir adekvatnih hiperparametara i modelovanje signala ARIMA metodom. ARIMA model se sastoji od tri komponente: AR (AutoRegression), I (Integrated) i MA (Moving Average). Hiperparametri koje treba odrediti su (p, d, q), gde p predstavlja broj prethodnih vrednosti koje će biti uključene u model (AR komponenta), d predstavlja broj diferenciranja ulaznog signala (I komponenta), dok q predstavlja veličinu *moving average window* (MA komponenta). Parametri p i q mogu se odrediti na osnovu PACF, odnosno ACF dijagrama, a pored toga se za odabir najboljeg modela mogu koristiti mere AIC (Akaike Information Criterion) i BIC (Bayesian Information Criterion).

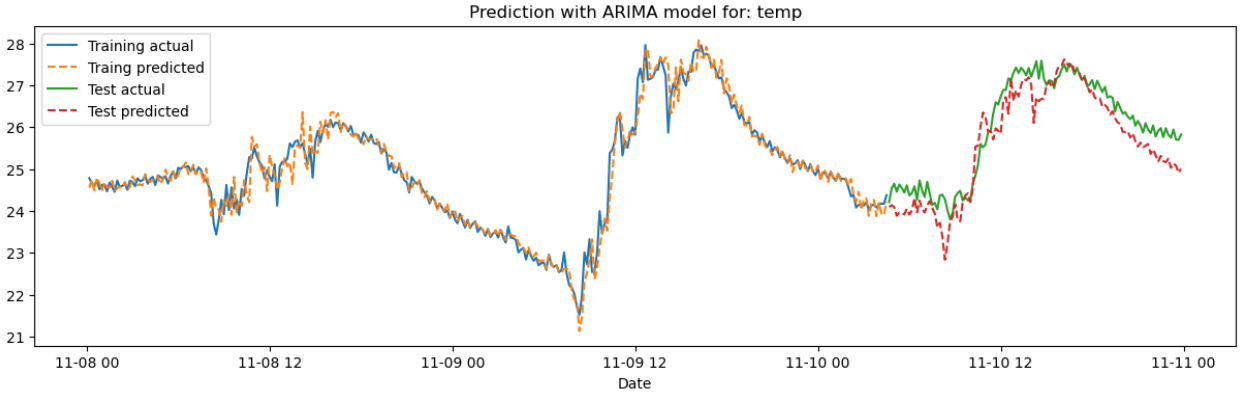
U ovom slučaju je, s obzirom na postojanje sezonalnosti, bilo pogodno primeniti *seasonal* ARIMA model (SARIMA) umesto standarnog. Ovaj model zahteva specificiranje više hiperparametara: (p, d, q)(P, D, Q, m), gde su (p, d, q) nesezonski parametri, a (P, D, Q, m) sezonski. Parametar m predstavlja sezonalnost (u ovom slučaju odgovara jednom danu).

Za konkretan slučaj atributa *temp*, analizom stacionarnosti je određeno da parametri d i D dobiju vrednost 1, parametar m je postavljen na 144 (24*6, jer su podaci odmeravani na 10 minuta), dok su kandidati za parametre p, P, q i Q određeni na osnovu PACF i ACF dijagrama. Na PACF dijagramu se može primetiti da postoji statistički značajna korelacija za lag=1, a nakon toga je jedna od poslednjih značajnih vrednosti za lag=12, pa su te dve vrednosti razmotrene za parametar p, a parametar P je postavljen na 1 jer za lag=144 postoji vrednost koja prelazi prag značajnosti. Na ACF dijagramu se primećuje da postoji značajna korelacija za lag=1, a poslednja značajna vrednost je za lag=19, pa su te dve vrednosti razmotrene za parametar q, dok je parametar Q postavljen na 0. Kombinacije sa potencijalnim vrednostima za ove parametre su upoređene korišćenjem AIC mere, tako da je kao najbolji model odabran onaj sa najmanjom AIC vrednošću: SARIMA(12, 1, 19)(1, 1, 0, 144).

Zatim je vizuelizovano koliko dobro je modelovan trening signal, kao i koliko dobro su predviđene nove vrednosti. Na sledećim graficima se može primetiti da postoji dobro poklapanje između prediktovanih i tačnih vrednosti:

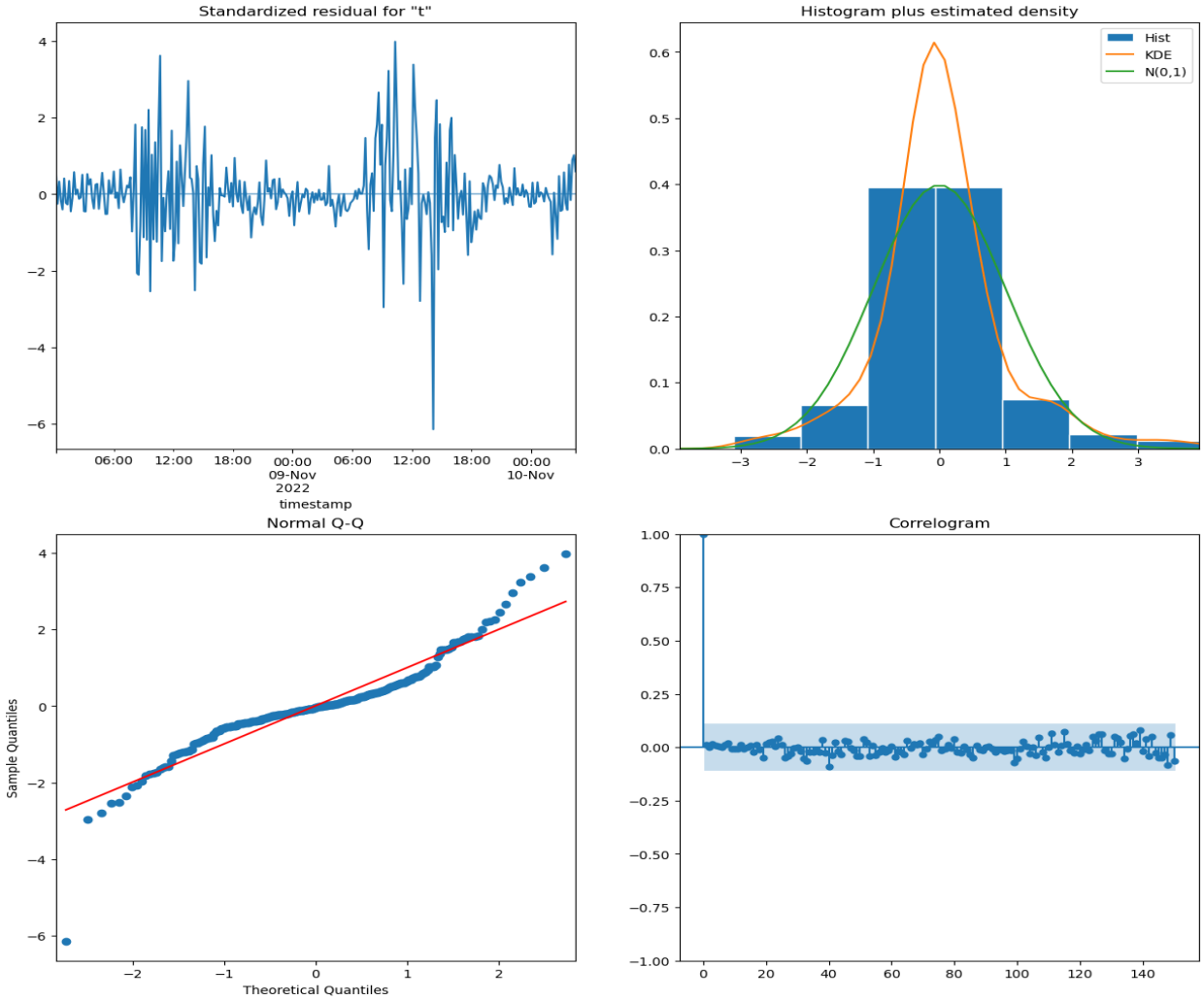






Nakon toga je izvršena evaluacija modela pozivom ugrađene *plot_diagnostics* funkcije nad njim. Ova funkcija generiše četiri dijagrama pomoću kojih se mogu analizirati reziduali (razlika između tačnih i prediktovanih vrednosti): na prvom se nalaze standardizovani reziduali (koji treba da izgledaju kao šum, bez uočljivog šablona), na drugom se nalazi histogram ovih reziduala (koji treba da se što bolje poklapa sa histogramom normalne raspodele), na trećem se prikazuje Q-Q plot (kojim se raspodela upoređuje sa normalnom), a na četvrtom se prikazuje autokorelacija reziduala (za dobar model se očekuje da sve tačke budu unutar obojenog intervala, a ukoliko postoji značajna korelacija između reziduala, to bi značilo da postoje informacije u podacima koje nisu obuhvaćene modelom).

Za atribut *temp* se može reći da su skoro svi uslovi zadovoljeni (reziduali nisu korelisani, a raspodela nije u potpunosti normalna, ali ne odstupa značajno).



Reziduali se mogu evaluirati i na osnovu rezultata poziva ugrađene funkcije *summary* nad modelom. U dobijenoj tabli *Prob(Q)* se odnosi na korelaciju reziduala (ako je manja od 0.05 reziduali su korelisani), dok *Prob(JB)* odnosi na informaciju o tome da li reziduali imaju normalnu raspodelu (ako je manja od 0.05 reziduali nemaju normalu raspodelu). U slučaju atributa *temp* ove mere pokazuju da reziduali nisu korelisani, ali takođe i da nemaju raspodelu koja se poklapa sa normalnom.

U fajlu „arima_model_residuals.csv“ sačuvani su dobijeni reziduali primenom ARIMA metode za svaki od atributa. Ovaj fajl ima 10 kolona (kao i skup podataka), pri čemu prva predstavlja *timestamp*, a ostale odgovaraju posmatranim parametrima.