

Tema A1 Analiza Environmental data: Data Exploration

Tema obuhvata primenu raznih metoda koje se tiču istraživanja vremenskih serija nad environmental parameters data. Cilj je dublje istražiti podatke koji nam daju značajan uvid u statistički i analitički pristup nad datim setom podataka, koja otvara puteve za dalje adekvatne predikcione metode nad vremenskim serijama.

Proces:

1. **Prvi korak** je osnovno istraživanje podataka. Pre svega je bitno razumeti kontekst i priču iza podataka da bi output bilo koje analize mogao da bude validan. Zato vršimo početno istraživanje podataka, koje služi kao određivanje osnovnih podataka o samim podacima, kao što je tip podataka, vrednosti, jedinice, a onda i analitički deo, ovde spada opsežna statistička analiza podataka u vidu računanja osnovnih parametara o podacima, kao što su mean, median, standard variation, kao i posmatranje boxplot-ova, histogram-a, distribucije podataka (Q-Q plot).
2. **Drugi korak** je preprocesiranje podataka. Nakon istraživanja podataka, već imamo bolji uvid u kontekst, pa lakše možemo odrediti i metode preprocesiranja. U osnovne metode spada pronalaženje i ispunjavanje missing values (NaN) ukoliko ih ima, otklanjanje duplikata, resampling (poželjno na frekvenciji od 1 Hz), otklanjanje outliers statističkim metodama. Da li je moguće ostvariti empiric rule (pravilo “3 sigma”) za outlier detection? Da li je moguće ostvariti IQR metodu za outlier detection? Kakva je korelisanost između podataka?
3. **Treći korak** je napredna analiza vremenskih serija. Cilj je proučiti, istražiti i primeniti ove metode nad environmental dataset.
 - Proveriti stacionarnost signala koji je bitan zbog predikcionih metoda kod vremenskih serija, otklanjanje stacionarnosti metodama diferenciranje, korenovanje, logaritmovanje i proveru
 - Računanje (auto)kovarijanse i (auto)korelacije kao glavnog pokazatelja vremenskih serija u odnosu na druge veličine, a pored toga i parcijalne autokorelacije. Proveriti zašto su signali vremenske serije.
 - Modelovanje vremenskih serija AR(IMA) metodom, određivanje adekvatnih hiperparametara.
 - Neke predikcione metode zahtevaju da signal ne bude autokorelisan. Otklanjanje autokorelacija u signalu primenom AR(IMA) metode i računanje reziduala.
 - Krajnji rezultat su reziduali za svaki environmental parametar.
4. **Četvrti korak** je evaluacija i validacija algoritama. Evaluirati metodu koja je korišćena tako što proverimo da li su reziduali sada korelisani i kakva je distribucija signala.

Krajnji rezultat se očekuje kao .csv fajl sa rezidualima za svaki od parametara dobijenih u environmental data, gde je prva kolona kao i pre timestamp. Očekuje se da su reziduali dobijeni uz pomoć AR(IMA) metode.

Pored ovog word dokumenta, u zip fajlu se nalazi environmental dataset par radova kao propratna literatura na temu konkretnih metoda za analizu koji će biti od pomoći za bolje razumevanje algoritma.

Ceo proces je poželjno dokumentovati na par strana (**dužina dokumenta nije bitna**).