

Analiza regresji i szeregów czasowych

Projekt, 2023/24

prof. dr hab. inż. Piotr Cofta

Celem projektu jest analiza i predykcja szeregów czasowych z wykorzystaniem udostępnionych danych.

Problem i dane

Jakość energii elektrycznej w sieci niskiego napięcia (tj. 'domowej', 230V) musi spełniać szereg wymagań dotyczących jej stabilności. Ponieważ system produkcji i dystrybucji energii podlega ciągłym zakłóceniom, cały czas koryguje on parametry dostarczanej energii tak, aby spełniały wymagania. Efektem tego są ciągłe zmiany rzeczywistych parametrów energii.

Ponieważ korekcja zawsze następuje z opóźnieniem, system energetyczny mierzy parametry oraz przewiduje ich zmiany w czasie. Im lepsze jest zrozumienie zachowania się parametrów energii, tym lepsza jest reakcja, a co za tym idzie, tym lepsza jest jakość energii.

Celem projektu jest analiza oraz predykcja zachowania się parametrów w sieci energetycznej niskiego napięcia.

W ciągu ostatniego roku, wartości trzech parametrów sieci energetycznej były mierzone w przybliżeniu co sekundę. Parametry te to

- V1 – napięcie na pierwszej fazie (w systemie trójfazowym)
- F – częstotliwość
- THDV1 – procent harmonicznym w napięciu na pierwszej fazie (tj. odchylenie od oczekiwanej sinusoidy przebiegu napięcia)

Dane z pomiarów są udostępnione w katalogu

https://utpedupl-my.sharepoint.com/:f/g/personal/piotr_cofta_o365_pbs_edu_pl/EIRhg3VcDgJKhrHVU1BTBOWBvmVBggvLh7R3q1yd87coiQ?e=zBaSYR

Katalog ten zawiera trzy podkatalogi: v1, f i thdv1, odpowiadające trzem mierzonym parametrom. W każdym podkatalogu znajduje się 365 plików, po jednym na każdy dzień roku. Data oraz nazwa parametru są częściami nazwy pliku. Plik zawiera skompresowany plik w formacie CSV. Plik CSV zawiera dwie kolumny: moment pomiaru i zmierzoną wartość parametru. Moment pomiaru jest zapisany w milisekundach, według schematu 'Epoch'. Jakkolwiek pomiary powinny być realizowane co sekundę, nie ma gwarancji iż nie wystąpiły opóźnienia, powielanie lub brak danych.

Należy pobrać dane za dwa kolejne tygodnie, dla trzech parametrów. Numer pierwszego tygodnia jest unikalny dla każdego studenta. Razem należy pobrać 42 pliki.

Projekt

Projekt jest realizowany w dwóch etapach.

Etap 1. – wstępna analiza danych

Nie ma gwarancji, iż dane w plikach są właściwie sformatowane, kompletne, ani że są zgodne z opisem. Należy dla obu tygodni dokonać wstępnej analizy danych, określając:

- czy dostępne są wszystkie dane, a jeżeli nie to ilu brakuje; czy konieczne jest uzupełnienie danych, jaką metodą i dlaczego (lub dlaczego nie dokonywać korekty)
- czy dane są właściwie rozmieszczone na osi czasu, czy należy dokonać ich korekty, jaką metodą i dlaczego (lub dlaczego nie należy dokonywać korekty)
- czy dane wymagają jakiegokolwiek czyszczenia, usuwania danych wątpliwych, i dlaczego tak albo nie

Osobno za każdy tydzień:

- jakie są podstawowe parametry danych takie jak wartość średnia, maksymalna, minimalna, itp.
- jak wygląda histogram i czy dane są zgodne z jakimkolwiek ze znanych rozkładów (równomierny, normalny, itp.)
- czy występuje korelacja pomiędzy wartościami parametrów
- czy występuje autokorelacja, czy przebieg jest okresowy, czy można wyodrębnić harmoniczne?

Etap ten powinien zakończyć się raportem, obejmującym

- Opis zastosowanych metod weryfikacji
- Wyniki weryfikacji
- Kod stosowany do weryfikacji (jeżeli weryfikacja wymagała kodu)

Etap 2 – przewidywanie wartości

W tym etapie należy opracować metodę regresji która pozwoli na prognozowanie wartości parametrów w następnym tygodniu na podstawie wartości z poprzedniego tygodnia. Dane z pierwszego tygodnia służą jako zestaw ‘treningowy’, dane z drugiego tygodnia jako zestaw ‘testowy’. Metryka jakości jest pozostawiona do decyzji studenta.

Istnieje znaczna liczba sposobów podejścia do zagadnienia, od bezpośredniego zastosowania metod regresji do zastosowania metod w sposób bardziej wyrafinowany, poprzez np. przekształcenie danych, interpolację, uśredniania, wygładzania, analizę harmonicznych itp. Podobnie możliwa jest analiza każdego parametru osobno, wszystkich razem, jednego w kontekście pozostałych itp.

Nie ma gwarancji, iż istnieje metoda lepsza od prostego uśrednienia. Nie ma również gwarancji, iż wykorzystanie kilku parametrów przyniesie lepsze rozwiązania, lub że prognoza krótkoterminowa będzie lepsza niż długoterminowa.

Etap ten powinien zakończyć się raportem, obejmującym

- Wybrane podejście, wybraną metodę regresji, zakres przewidywania, miarę jakości
- Szczegóły procesu
- Dobór narzędzi, z uzasadnieniem
- Wyniki wraz potrzebnymi wyjaśnieniami
- Wytworzony kod

Ocena

Warunkiem koniecznym zaliczenia jest

- weryfikacja zbiorów danych
- określenie ich charakterystyki
- realizacja i przetestowanie regresji
- dostarczenie raportów w uzgodnionych terminach
- aktywny udział w spotkaniach projektowych zgodnie z programem zajęć
- prezentacja postępu prac na wybranych spotkaniach

Ocena w mniejszym stopniu zależy od uzyskanej jakości przewidywania, a głównie zależy od kompletności i dojrzałości podejścia, udokumentowanej w opracowaniu. Oceniane są takie elementy jak

- uzasadnienie wybranego podejścia / podejść
- uzasadnienie doboru narzędzi i metod pracy
- uzasadnienie doboru metod i ich parametrów
- refleksja nad uzyskanymi wynikami
- jasność prezentacji wyników (opisy, wykresy, tabele)