Analiza danych w naukach o Ziemi – Geoinformatyka II st. – prowadzący Jakub Staszel

Projekt 1 – Standardized Precipitation Index (SPI)

Spis treści

Wstęp

Standardized Precipitation Index (SPI) to miara, która została opracowana w celu kwantyfikacji opadów w danym miejscu przez określony okres. SPI umożliwia określenie suszy i jej nasilenia na podstawie długoterminowych danych o opadach. Jest to wskaźnik stosunkowo prosty w użyciu, ponieważ wymaga jedynie danych o opadach, dzięki czemu można go łatwo zastosować do różnych regionów i okresów czasu.

SPI oblicza się poprzez normalizację opadów dla danego miejsca i okresu (na przykład miesięcznych lub rocznych) do standardowego rozkładu normalnego. W ten sposób SPI pokazuje, ile odchylenia standardowego opady w danym okresie odbiegają od średniej długoterminowej. Na przykład:

- SPI o wartości 0 wskazuje na średnią ilość opadów.
- Dodatnie wartości SPI wskazują na opady większe niż średnie.
- Ujemne wartości SPI wskazują na mniejsze niż średnie opady, co może wskazywać na suszę.

Długość okresu, dla którego obliczany jest SPI, może być różna (np. 1 miesiąc, 3 miesiące, 6 miesięcy itd.), co pozwala na analizę krótkotrwałych jak i długotrwałych trendów opadów i ich wpływu na zasoby wodne, rolnictwo, zarządzanie ryzykiem suszy i inne aspekty związane z wodą.

Obliczanie SPI w teorii

W najbardziej podstawowym wariancie SPI można zapisać następującym wzorem:

$$SPI = \frac{opad - średni\ opad}{odchylenie\ standardowe\ dla\ opadów}$$

Jednak, żeby wzór ten można było zastosować konieczne jest posiadanie danych, które są bliskie rozkładowi normalnemu. Dane opadowe zwykle tak nie wyglądają, przez co stosuje się dla nich raczej rozkład gamma.

Krok 1: Przygotowanie danych opadowych – w zależności od rodzaju SPI obliczamy sumy opadów dla odpowiednich okresów.
Krok 2: Dopasowanie rozkładu prawdopodobieństwa – żeby stosować średnią czy odchylenie standardowe dane muszą posiadać rozkład normalny. W przypadku SPI zwykle pojawia się rozkład gamma.
Krok 3: Transformacja do rozkładu normalnego
Krok 4: Obliczenie SPI
Krok 5: Interpretacja wyników

Wymagania i założenia

1. Akademickie
   1.1. Projekt można wykonać samemu lub w grupach maksymalnie 3-osobowych.
   1.2. Składy zespołów oraz linki do repozytoriów należy wysłać w 1 tygodniu do prowadzącego.
   1.3. Projekt przewidziany jest na 4 lub 5 zajęć, jednak możliwe są modyfikacje – proszę o informację, jeśli potrzebne będzie więcej lub mniej czasu. Chciałbym respektować nakład pracy studenta określony w sylabusie, stąd też zachęcam do kontaktu, jeśli będzie to potrzebne.
   1.4. Weryfikacja postępów w projekcie odbywa się w cyklach tygodniowych, w każdym tygodniu oczekiwany jest postęp w budowaniu analizy i repozytorium (podlegać to będzie weryfikacji).
   1.5. Brak postępów czy opublikowanie całości kodu przed samym oddaniem projektu będzie wpływać negatywnie na ocenę.
   1.6. Oprócz repozytorium na koniec należy oddać sprawozdanie opisujące wszystkie etapy przeprowadzonej analizy. Nie ma konieczności produkowania zbędnego tekstu, liczą się opis każdego elementu analizy oraz opis wyników.
2. Techniczne
   2.1. Organizacja samego repozytorium ma być logiczna, ale występuje tutaj dowolność.
   2.2. Wykorzystywanie Jupyter Notebooks jest wskazane, szczególnie na etapie developmentu, ale ostateczna wersja kodu ma być podzielona na moduły w repozytorium.
   2.3. W pełni należy korzystać z możliwości rozproszonego systemu kontroli wersji (GIT). Proszę oznaczać mnie jako reviewera w swoich PRach do głównej gałęzi, to tutaj będzie prowadzona weryfikacja postępów oraz jakości kodu. Dodatkowo, notatniki, pliki z danymi oraz wynikami nie powinny być przechowywane w repozytorium, to jest miejsce na sam kod – pomocne tutaj będzie wykorzystanie .gitignore.
   2.4. Analizy mają być w pełni reprodukowalne. Oznacza to, że osoba, która wchodzi do repozytorium ma możliwość przeprowadzenia wszystkich analiz podążając za instrukcjami w pliku README.
   2.5. W pliku README mają się znaleźć informacje dot. tego gdzie pobrać dane, jak stworzyć środowisko oraz jak wyzwolić kolejne etapy przetwarzania danych. Do tworzenia i zarządzania środowiskiem polecam korzystać z conda-lock.
   2.6. Każdy kolejny etap ma sprowadzać się do pojedynczych funkcji, których parametry oraz działanie jest opisane w komentarzach.
   2.7. Wszystkie analizowane wykresy mają być zapisywane do plików, podobnie ostatecznie policzone wartości SPI.
3. Tematyczne

3.1. W projekcie mają zostać przeanalizowane 3 warianty SPI: SPI-1, SPI-3 oraz SPI-12.
3.2. Analizie poddane mają zostać dane ze stacji zlokalizowanych na wybranym obszarze.
3.3. Po uruchomieniu całej analizy w repozytorium mają pojawić się pliki z ostatecznymi wartościami SPI oraz wykresy.
3.4. W poniższym opisie projektu przedstawiono przydatne funkcje czy sposoby analizy wyników, ale są to jedynie przykłady – zachęcam do stworzenia własnych elementów czy dyskusji na temat innego podejścia.

Cele projektu

Ocena i analiza zmienności warunków hydrologicznych oraz ryzyka suszy w wybranym regionie poprzez obliczenie i analizę Standardized Precipitation Index (SPI) na różnych skalach czasowych.

Kolejne etapy projektu i minimalne wymagania

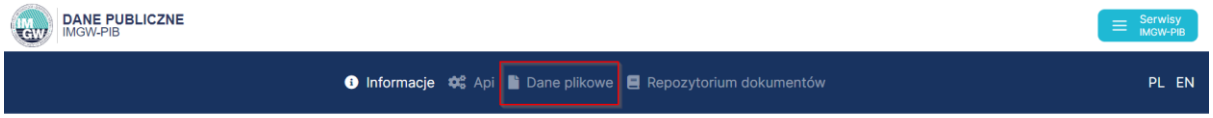
To jest sugestia dot. kolejności wykonywania działań, nie jest obowiązkowa.

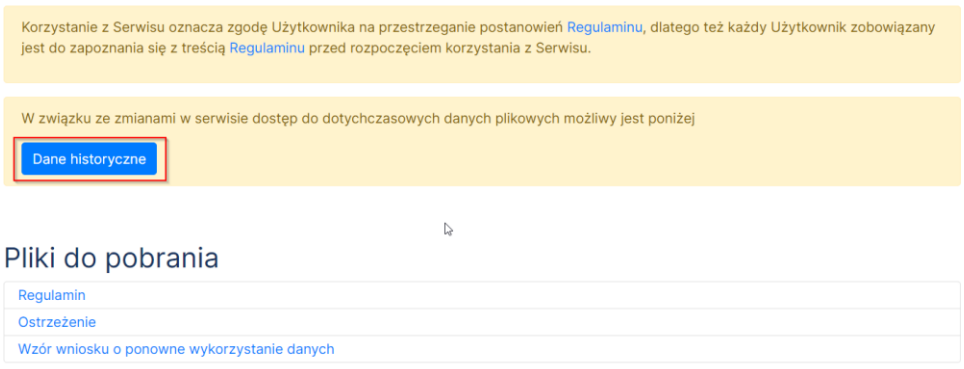
Pobranie danych oraz zapoznanie się z ich formatem

Proszę wykorzystać dane publiczne udostępniane przez IMGW – https://danepubliczne.imgw.pl/. Dane można pobrać ręcznie lub zbudować automatyczny proces.

Szczegóły praktyczne:

- Jedynym ewentualnie ręcznym krokiem ma być pobranie danych do wybranego katalogu. Rozpakowywanie (unzipping) i przeszukiwanie plików ma odbywać się z poziomu kodu.
- **Danymi wsadowymi są dane dobowe**.
- Konieczne są dane nie tylko dotyczące opadów, ale także lokalizacje poszczególnych stacji.
- SPI jest indeksem, który wymaga szerokiego zakresu czasowego – powinniśmy posiadać co najmniej 30 lat danych. Nie znam pełnej struktury danych, więc mogą się pojawić przypadki, gdzie tych danych będzie za mało / nigdzie nie będzie takiej sytuacji, więc proszę o otwartą dyskusję.
- Proszę zwrócić uwagę, że dane można znaleźć w 2 katalogach, w jednym z nich można znaleźć współrzędne dla stacji, ale dane tam są dopiero od 2008 roku. W drugim dostępne są także wcześniejsze dane. Dane z obu źródeł mogą mieć różny format.





Analiza lokalizacji stacji pomiarowych

Przed przystąpieniem do analizy samych danych, należy wybrać obszar, dla którego prowadzona będzie analiza – tego dotyczy ten punkt.

- Jako podstawowe warianty proszę przyjąć **obszary województw**, wszelkie inne geometrie proszę przedyskutować z prowadzącym.
- Gęstość przestrzenna stacji jest mocno zróżnicowana, przez co wybranie obszaru jest ważne. Proszę o stworzenie mapy z lokalizacjami stacji.
- Stacje mogą zmieniać nazwy w czasie, identyfikatory pewnie też, więc należy to przeanalizować i przyjąć, że to jest ta sama stacja jeśli ma te same współrzędne.

Przydatne na tym etapie: biblioteki pandas i geopandas.

Wynik tego etapu: lista stacji znajdujących się na obszarze analizy, mapa stacji naniesiona na obszar, ewentualne szczegóły i modyfikacje w danych, które mają wpływ na uzyskane wyniki.

Mapowanie stacji z odpowiadającymi im danymi

Interesujące z punktu widzenia analizy są tylko dane dla stacji, które znajdują się na analizowanym terenie. Co więcej, wszystkie obliczenia prowadzone będą na poziomie jednej stacji, więc na tym etapie można się zastanowić czy nie warto zapisać plików w nowym formacie – dla jednej stacji, ale z dłuższym zakresem czasu. Dane wsadowe posiadają także więcej atrybutów, czy będą nam one potrzebne?

- W danych pomiarowych mogą występować stacje, których brakowało w wykazie stacji. Należy je zidentyfikować i ewentualnie uzupełnić plik z lokalizacjami[1].
- Dla każdej stacji na obszarze należy przedstawić zakres dat dla których dostępne są dane.

Przydatne na tym etapie: manipulacja danymi wbudowana w biblioteką pandas.

Wynik tego etapu: pliki w nowym formacie (opcjonalnie), informacje o zakresie dat w których posiadamy dane dla poszczególnych stacji, informacje o stacjach, których brakowało w głównym wykazie.

Przygotowanie danych – preprocessing; Eksploracja danych - exploratory data analysis (EDA)

Na tym etapie należy przyglądnąć się danym, zdefiniować braki, błędne wartości czy duplikaty, zastanowić się w jaki sposób można je uzupełnić.

- Opad to zmienna ciągła, ale zgodnie z opisem, niektóre wartości nie są związane z samymi wynikami, czy istnieją tam jakieś flagi (wartości oznaczające coś innego niż sam pomiar)?
- Jaki jest format dla braku danych? Czy te dane można jakoś uzupełnić?

Czyszczenie danych

- Szukanie brakujących wartości: użycie df.isnull() lub df.isna() do identyfikacji brakujących danych. Sprawdzenie, jak dużo danych brakuje w poszczególnych kolumnach, może to pomóc podjąć decyzję, czy wartości te należy uzupełnić, czy też usunąć dane wiersze/kolumny.
- Usuwanie duplikatów: funkcja df.drop_duplicates() pomoże usunąć zduplikowane wiersze, należy sprawdzić jak funkcja definiuje czym jest duplikat.

[1] Jeśli stacji tych będzie dużo, proszę o informację, wspólnie ustalimy kolejne działania w tym temacie.

Uzupełnianie danych

- Uzupełnianie braków: df.fillna(value) pozwala na uzupełnienie brakujących danych stałą wartością, średnią (df.fillna(df.mean())), medianą itp.
- Interpolacja: metoda df.interpolate() pozwala na bardziej zaawansowane uzupełnianie braków, na przykład przez interpolację liniową czy metodę najbliższych sąsiadów.

Transformacja danych

- Zmiana typów danych: df.astype(type) umożliwia zmianę typu danych w kolumnach, co jest przydatne np. przy konwersji typów numerycznych czy konwersji dat.
- Raczej w przypadku tych danych to ciekawostka: normalizacja i standaryzacja: biblioteka sklearn.preprocessing oferuje funkcje takie jak StandardScaler czy MinMaxScaler do skalowania danych numerycznych.

Podstawowa analiza statystyczna

- Podsumowanie statystyczne: użycie df.describe() w Pandas do szybkiego przeglądu statystyk podsumowujących takich jak średnia, mediana, min/max, kwartyle dla danych numerycznych.
- Liczebność i unikalność: funkcje df.count() i df.nunique() pomagają zrozumieć, ile jest niepustych wartości i ile unikalnych wartości zawiera każda kolumna.

Wizualizacja danych

- Histogramy i dystrybucje: użycie df.hist() lub bibliotek jak Matplotlib i Seaborn do generowania histogramów, które pomagają zrozumieć rozkład danych.
- Wykresy pudełkowe (boxplots): służą do identyfikacji wartości odstających i zrozumienia rozkładu danych.
- wykresy punktowe (scatter plots): pozwalają zobaczyć zależności między dwoma zmiennymi i identyfikować wzorce lub korelacje.
- Raczej ciekawostka (nie mamy tylu atrybutów): wykresy ciepła korelacji (correlation heatmaps): użyteczne do wizualizacji siły i kierunku związku między zmiennymi.

Przydatne na tym etapie: dokumentuj swoje kroki – przechowuj notatki dotyczące tego, jakie transformacje zostały wykonane na danych, aby zapewnić przejrzystość i możliwość reprodukcji analizy.

Wynik tego etapu: dane o nowej jakości, wszystkie uwagi i ważne kwestie dot. wprowadzonych zmian i poznanych danych.

Obliczenie SPI

SPI obliczamy na poziomie 1 stacji i 1 miesiąca. Należy obliczyć SPI w 3 wariantach:

- SPI-1: suma opadów z 1 danego miesiąca,
- SPI-3: suma opadów z danego miesiąca i 2 miesięcy przed,
- SPI-12: suma opadów z danego miesiąca i 11 miesięcy przed.

Kolejne kroki (odpowiadające tym z rozdziału Obliczanie SPI w teorii):

1. Obliczenie sum opadów dla odpowiedniego okresu (zgodnie z poprzednim wypunktowaniem),
2. Obliczenie parametrów rozkładu gamma – scipy.stats.gamma.fit(),

W przypadku tych danych i rozkładu gamma naturalne jest przyjęcie wartości parametru floc (lokalizacji) jako 0. Zapewnia to, że modelowany rozkład zaczyna się od 0, co jest logiczne dla danych o opadach. Model nigdy także nie powinien przyjmować wartości niższych niż 0.

3. Obliczenie prawdopodobieństwa (CDF) dla danych opadów za pomocą rozkładu gamma – scipy. stats.gamma.cdf() i transformacja CDF do wartości Z (standardowego rozkładu normalnego) – stats.norm.ppf().

W tym przypadku metoda fit() powinna zwrócić 3 wartości: kształt (tutaj alfa), lokalizację i skalę (tutaj beta). Funkcja cdf() wykorzystuje następnie stworzony model gamma do obliczenia prawdopodobieństwa.

Jako input funkcja ppf() przyjmuje wartości prawdopodobieństwa z funkcji wyżej. Jej output to otrzymane wartości SPI.

Przydatne na tym etapie: zalecane jest przekształcenie danych do typu array z biblioteki numpy przed zastosowaniem funkcji statystycznych.

Wynik tego etapu: obliczone wartości różnych wariantów SPI dla stacji na wybranym obszarze – zalecane jest wyeksportowanie tych danych do plików.

Analiza, wizualizacja i interpretacja wyników[2]

Można przyjąć następujące przedziały przy interpretacji wartości:

- SPI 2.0 lub więcej: ekstremalnie mokro
- SPI od 1.5 do 1.99: bardzo mokro
- SPI od 1.0 do 1.49: umiarkowanie mokro
- SPI od -0.99 do 0.99: średnie warunki
- SPI od -1.49 do -1.0: umiarkowana susza
- SPI od -1.99 do -1.5: silna susza
- SPI -2.0 lub mniej: ekstremalna susza

Na poziomie pojedynczych stacji proponowane elementy to:

1. Statystyczne podsumowanie wyników (statystyki opisowe),
2. Wizualizacja zmienności w czasie (jakieś trendy?),
3. Porównanie ze sobą wariantów SPI (czy skala czasowa (wariant SPI) ma wpływ na obserwowane zmiany?).

Na poziomie wszystkich danych dla analizowanego obszaru – różne stacje mogą mieć różny zakres czasowy, a więc zwracane mogą być listy różnej długości dla SPI – należy to ujednolicić:

1. Statystyczne podsumowanie wyników (statystyki opisowe),
2. Wizualizacja zmienności w czasie dla wszystkich stacji na jednym wykresie i/lub dla statystyk (np. dla średniej),
3. Mapa przedstawiająca wartości SPI w różnych lokalizacjach (czy widoczne są jakieś miejsca bardziej narażone na suszę?).

Dotyczy raczej samych danych o opadach:

[2] Ten element to prawdopodobnie będzie większość objętości sprawozdania, ale jak należy stworzyć kod, który będzie zastosowany na wszystkich danych (tj. w kwestii wykresów czy obliczonych statystyk), to tak nie trzeba wszystkiego opisywać w sprawozdaniu – proszę wybrać najbardziej ciekawe wizualizacje czy statystyki.

1. Mapa przedstawiająca zakres czasowy dla różnych stacji – np. może to być suma dni pomiędzy pierwszym i ostatnim pomiarem (czy posiadamy wystarczającą ilość danych, żeby dokonywać analiz?).