Farmy Wiatrowe

Marek Wiese

czerwiec 2021

1 Wprowadzenie

Na całym świecie energia wiatrowa jest drugim najpopularniejszym odnawialnym źródłem energii. W związku z szybko rosnącą liczbą monitorowanych turbin wiatrowych, pojawiło się zapotrzebowanie w zakresie algorytmów monitorowania i diagnostyki. Obecnie dane z setek systemów monitorowania przesyłane są do centrów diagnostycznych, gdzie muszą zostać przeanalizowane. Wiążą się z tym wysokie koszty pracy wielu ekspertów, dlatego zaszła potrzeba zautomatyzowania metod diagnostycznych. Artykuł przedstawia zastosowanie metod „Data Mining" do grupowania danych procesowych z turbin wiatrowych.

1.1 Działanie turbiny wiatrowej

Wiatr możemy zdefiniować jako ruch powietrza atmosferycznego. W zależności od prędkości poruszającego się powietrza można określić siłę wiatru i oszacować ilość jego energii. Podstawowe równanie energii wiatru jest postaci:

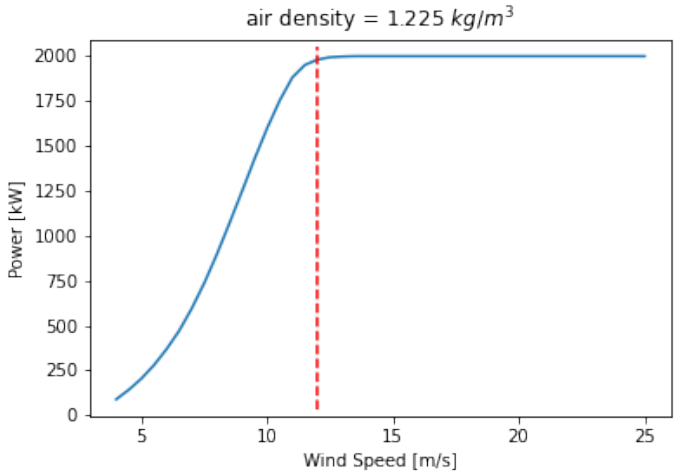
$$P = \frac{1}{2}\rho A v^3,$$

gdzie ρ jest gęstością powietrza, v prędkością wiatru, a A to powierzchnia, jaką zajmują łopaty wirnika.

Turbiny wiatrowe to systemy mechaniczne, które przechwytują energię wiatru i przekształcają ją w energię elektryczną, wykorzystując złożoną technologię dla uzyskania maksymalnej wydajności. Obejmują one różne obszary techniczne, w tym aerodynamikę, mechanikę, dynamikę struktury, meteorologię i inżynierię elektryczną. Wiatraki mogą być instalowane na lądzie lub na morzu.

1.2 Teoretyczna krzywa mocy

Turbina zaczyna pracę, gdy wiatr osiągnie tzw. prędkość „cut-in". Moc turbiny jest również ograniczona, i gdy tylko prędkość wiatru osiągnie tzw. wartość „cut-out", to przestaje ona pracować, aby nie wykonywać niepotrzebnej pracy i zachować jak najlepszą jakość mechaniczną. Krzywe mocy turbiny wiatrowej obliczane są na podstawie prędkości wiatru „cut-in" i „cut-out" oraz teoretycznego wyniku mocy dla różnych prędkości wiatru. Mają one kształt sigmoidy i pokazują jak turbiny będą działać w zależności od wiatru. Poniżej znajduje się wykres przykładowej teoretycznej krzywej mocy wraz z zaznaczoną na czerwono wartością wiatru (rated wind speed), dla którego osiągana jest maksymalna moc:



Rysunek 1: Teoretyczna krzywa mocy (dla analizowanego zbioru EDP)

2 Eksploracyjna analiza danych

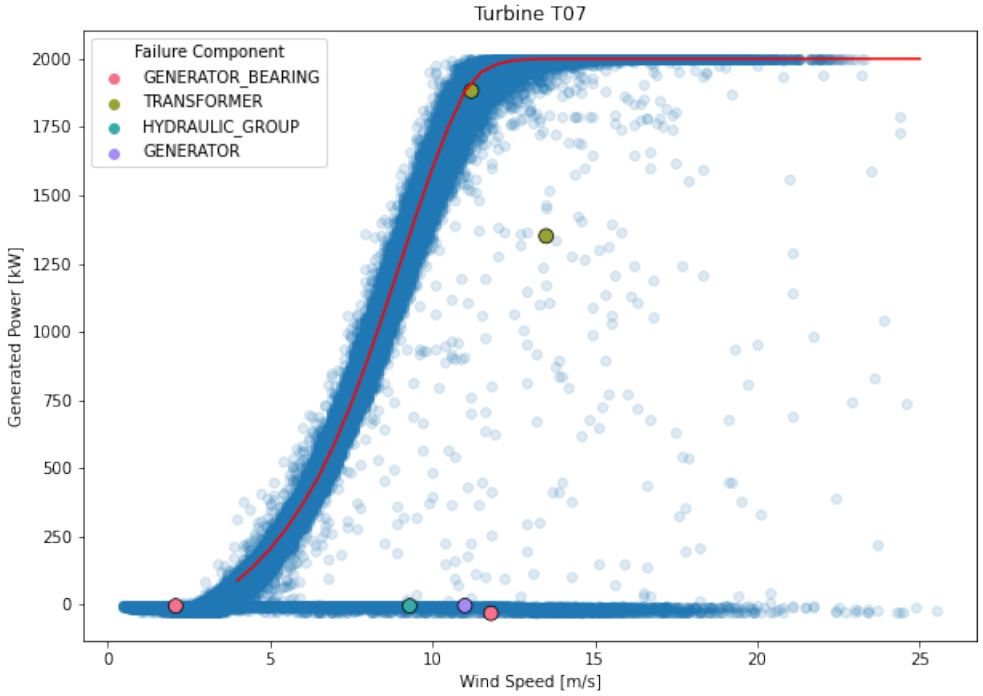
Zbiór danych pochodzi z EDP (Energias de Portugal) i składa się z dwuletnich zapisów systemu SCADA (rekordy co 10 minut) dla pięciu morskich turbin wiatrowych zlokalizowanych w zachodnioafrykańskiej Zatoce Gwinejskiej. Zbiór danych składa się z różnych plików, które zawierają informacje o awariach (failures), jak i informacje techniczne na temat głównych elementów turbiny, takich jak skrzynia przekładniowa (gearbox), generator i wirnik (rotor). Dodatkowe informacje obejmują dane meteorologiczne, a mianowicie prędkość i kierunek wiatru, ciśnienie powietrza, wilgotność, temperaturę oraz sygnały z podzespołów, a mianowicie obroty generatora i temperaturę oleju w grupie hydraulicznej.

W zbiorze danych wykryto kilka brakujących wartości (kolumna `Gen_Bear_Temp_Avg`), było ich na tyle mało (praktycznie 0%), że można by je usunąć, ale aby zachować ciągłość szeregu czasowego, uzupełniono je za pomocą prostej interpolacji liniowej.

Dane obejmują okres od początku 2016 roku do końca 2017 roku. W sumie przez ten czas zanotowano 23 awarie. Najbardziej awaryjnym komponentem turbiny okazał się generator. Poniższa tabela przedstawia ogólną liczbę awarii dla poszczególnych komponentów:

Komponent	Liczba awarii
GENERATOR	7
GENERATOR_BEARING	6
HYDRAULIC_GROUP	5
TRANSFORMER	3
GEARBOX	2
Łącznie	23

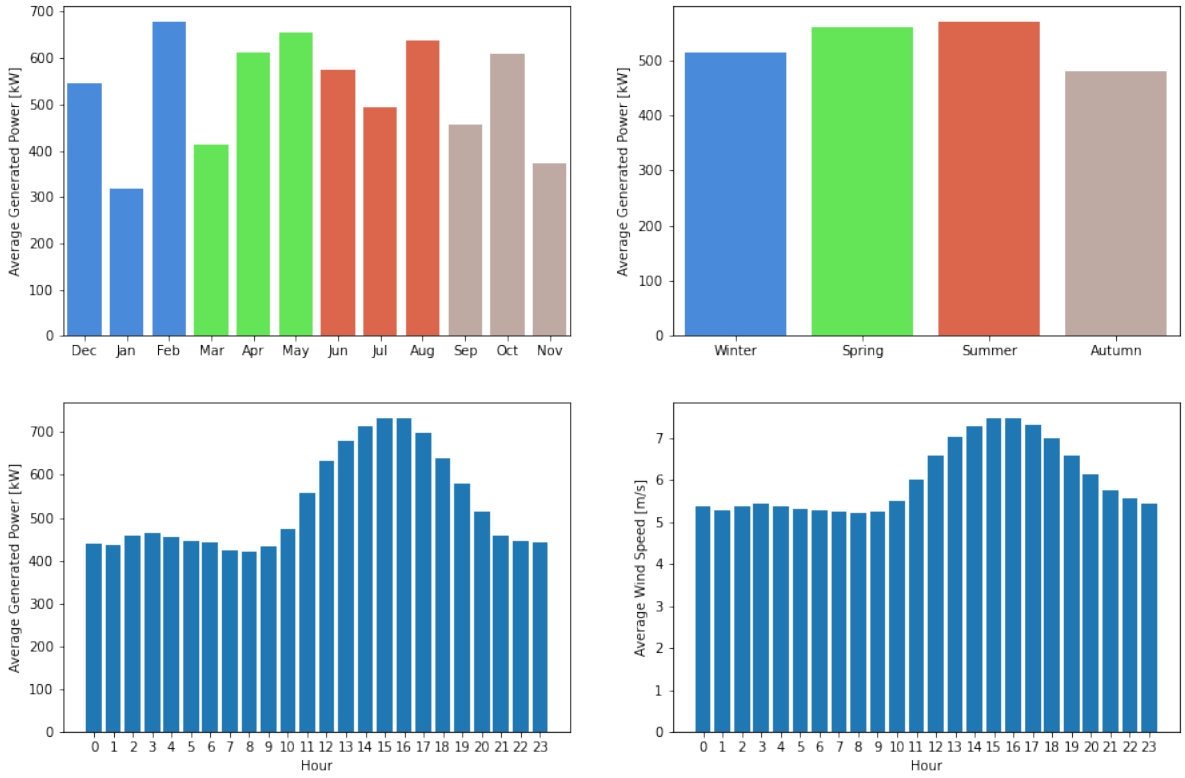
Łatwo jesteśmy w stanie za pomocą dostępnych danych odtworzyć zachowanie poszczególnych wiatraków, tzn. sprawdzić na ile generowany prąd ma się do teoretycznej krzywej mocy. Przykładowy wykres dla turbiny T07, dla której wystąpiło najwięcej awarii:



Rysunek 2: Zachowanie turbiny T07 (wraz z typami awarii)

Pozostałe turbiny zachowują się w podobny sposób, jak przedstawiona powyżej turbina T07. Możemy zauważyć, że występują obserwacje, dla których generowany prąd był ujemny, jak i obserwacje, wskazujące na jakieś anomalie, które znacząco odstają od sigmoidy, dlatego warto te dane oczyścić i skupić się na tzw. regionie operacyjnym, czyli takim, kiedy turbina pracuje i jej moc nie przekracza maksymalnej wartości.

Dodając kilka nowych kolumn związanych z czasem, możemy zagregować dane i zobaczyć kiedy, średnio rzecz biorąc, generowano najwięcej prądu:



Widzimy, że dominują tutaj w większości miesiące ciepłe oraz godziny popołudniowe.

3 Klasteryzacja – teoria

W analizie skupień w odróżnieniu od problemu klasyfikacji nie mamy do dyspozycji próby uczącej. Dana jest tylko prosta próba losowa z rozkładu P_X i na jej podstawie chcemy odtworzyć ukryte wartości Y (przynależności do klasy). Dlatego analizę skupień (cluster analysis) nazywa się czasami klasyfikacją nienadzorowaną (bez nauczyciela).

Postanowiłem zastosować algorytmy k-średnich oraz k-medoidów do modelowania danych oraz analizę składowych głównych (PCA) do wizualizacji otrzymanych klastrów. Do oceny jakości podziału struktury na skupienia wykorzystałem współczynnik oparty na sylwetkach.

3.1 Algorytmy k-średnich i k-medoidów

Algorytm k-średnich klasteryzuje dane, próbując podzielić próbki na k grup o równej wariancji, minimalizując kryterium znane jako inertia lub wewnątrzgrupowa suma kwadratów. Algorytm ten wymaga podania liczby klastrów. Dobrze skaluje się do dużej liczby próbek i był używany w wielu różnych obszarach zastosowań. Dokładny opis jego działania możemy znaleźć w [3].

Jeżeli chodzi o algorytm k-medoidów, to jest to bliźniaczy algorytm do k-średnich. Jest on bardziej odporny na obserwacje odstające. Najczęstszą metodą klasteryzacji jest partycja wokół medoidów (Partitioning Around Medoids – PAM), która polega na:

1. **Inicjalizacja:** Losowo wybierz k z n punktów danych jako medoidy.
2. **Krok przypisania:** Przypisz każdy punkt danych do najbliższego medoidu.

3. **Krok aktualizacji:** Dla każdego medoidu m i każdego punktu danych o związanego z m zamień m i o oraz oblicz całkowity koszt konfiguracji (czyli średnia odmienność o do wszystkich punktów danych związanych z m). Wybierz medoid o o najniższym koszcie konfiguracji.

4. Powtarzaj na przemian kroki 2 i 3, aż nie będzie zmian w przyporządkowaniach.

3.2 Analiza składowych głównych (PCA)

Główną ideą analizy składowych głównych jest zmniejszenie wymiarowości zbioru danych składającego się z dużej liczby powiązanych ze sobą zmiennych, przy jednoczesnym zachowaniu jak największej zmienności obecnej w zbiorze danych. Osiąga się to poprzez transformację do nowego zbioru zmiennych, czyli składowych głównych (PC), które są nieskorelowane, i które są uporządkowane w taki sposób, że kilka pierwszych składowych głównych wyjaśnia największy procent wariancji.

Pseudokod:

- Weź cały zbiór danych składający się z $d+1$ wymiarów i zignoruj etykiety tak, aby nowy zbiór danych stał się d-wymiarowy.
- Oblicz średnią dla każdego wymiaru całego zbioru danych.
- Oblicz macierz kowariancji dla całego zbioru danych.
- Oblicz wektory własne i odpowiadające im wartości własne.
- Posortuj wektory własne według malejących wartości własnych i wybierz k wektorów własnych z największymi wartościami własnymi, aby utworzyć $d \times k$ wymiarową macierz W.
- Wykorzystaj tę macierz W wektorów własnych do przekształcenia próbek na nową podprzestrzeń.

Więcej informacji i zastosowań na temat PCA znajdziemy tutaj [4].

3.3 Sylwetka obserwacji

Niech $C_1, \ldots, C_k$ będzie podziałem na skupienia i niech $\mathbf{x}_i \in C_l$. Definiujemy średnią odmienność $\mathbf{x}_i$ od elementów swego skupienia:

$$a(\mathbf{x}_i) = \frac{\sum_{\mathbf{u} \in C_l} d(\mathbf{x}_i, \mathbf{u})}{|C_l|}$$

oraz średnią odmienność $\mathbf{x}_i$ od skupienia C:

$$d(\mathbf{x}_i, C) = \frac{\sum_{\mathbf{u} \in C} d(\mathbf{x}_i, \mathbf{u})}{|C|},$$

gdzie $|C|$ – liczba elementów w C, d – metryka opisująca odległość (np. euklidesowa).

Odległość $\mathbf{x}_i$ od „najbliższego" skupienia (innego niż C_l):

$$b(\mathbf{x}_i) = \min_{C \neq C_l} d(\mathbf{x}_i, C).$$

Wówczas sylwetka (silhouette) obserwacji $\mathbf{x}_i$ dana jest wzorem:

$$sil(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max(b(\mathbf{x}_i), a(\mathbf{x}_i))} \in [-1, 1].$$

Współczynnik jakości podziału oparty na sylwetkach (Kaufman, Rousseuv):

$$SC = \max_k (\bar{s_k}),$$

gdzie $\bar{s_k}$ – średnia sylwetek k-tego skupienia.

Interpretacja:

- SC = 0.71-1.00: silna struktura podziału,
- SC = 0.51-0.70: umiarkowanie silna struktura podziału,
- SC = 0.26-0.50: słabo zidentyfikowane skupienia,
- SC $\leq$ 0.25: brak identyfikacji struktury.

4 Analiza skupień w danych

W tej sekcji przejdziemy już do praktycznego zastosowania wcześniej wspomnianych algorytmów. Użyjemy ich do zbadania segmentów czasowych, w których wiatraki operowały w podobny sposób, przyjrzymy się bliżej awariom oraz sprawdzimy, czy występuje efekt starzenia się wiatraków.

4.1 Wstępna analiza struktury

Przed dopasowaniem modelu k-średnich w celu zbadania podobieństwa w pracy turbin wiatrowych, dokonano standaryzacji danych, którą można opisać następującym wzorem:

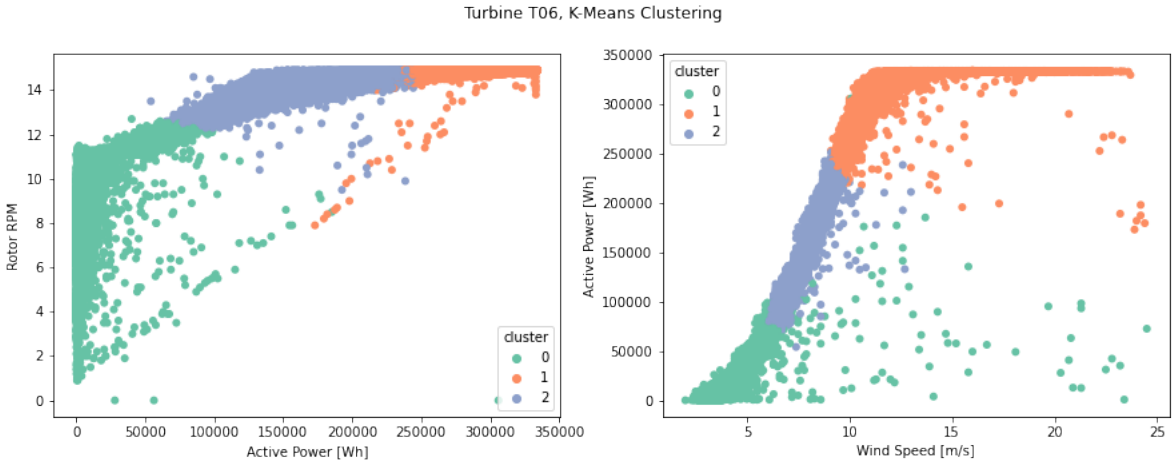
$$Z = \frac{X - \mu}{\sigma},$$

gdzie μ – średnia danego rozkładu, a σ – odchylenie standardowe danego rozkładu. Wówczas $Z \sim N(0, 1)$ jest nazywane wynikiem standardowym (standard score) i reprezentuje liczbę odchyleń standardowych powyżej lub poniżej średniej, do której dana obserwacja wpada. W dalszej części pracy również będziemy działać na danych wystandaryzowanych.

Podobieństwo wiatraków badałem na podstawie trzech zmiennych: prędkości wiatru, generowanego prądu aktywnego i obrotów wirnika. Odpowiedni w tym przypadku wydawał się podział na trzy skupienia, które można łatwo zinterpretować według generowanego prądu – mała moc (turbina dopiero zaczyna swoją pracę, mały wiatr), średnia moc (praca właściwa turbiny, optymalny wiatr) i duża moc (duży wiatr, maksymalna moc). Przykładowe centroidy tego podziału dla turbiny T06 przedstawia poniższa tabela:

cluster	Prędkość wiatru [m/s]	Prąd aktywny [Wh]	Obroty wirnika [RPM]
0	4.78	31977.14	10.95
2	7.78	144687.77	13.79
1	12.30	307568.62	14.84

Możemy również sprawdzić jak nasze wyniki prezentują się na wykresach dwuwymiarowych:



Widzimy, że w grupie średniej mocy (cluster = 2) mamy najmniej obserwacji odstających od teoretycznej krzywej mocy. Współczynniki sylwetki dla wszystkich turbin wynosiły w okolicach 0.56-0.57, co wskazuje na umiarkowanie silną strukturę podziału.

4.2 Analiza awarii

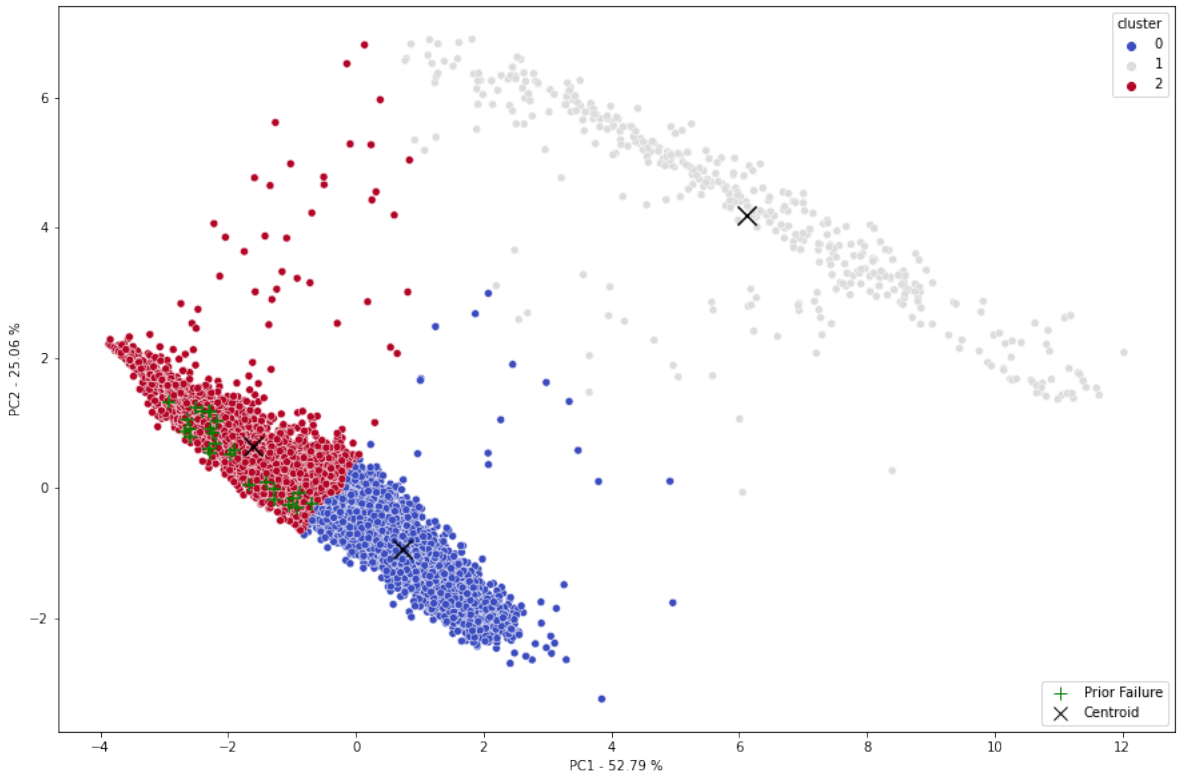
Ważnym elementem diagnostyki turbin wiatrowych jest analiza obserwacji „normalnych" oraz tych, które od awarii dzielił konkretny segment czasowy, na przykład 2 godziny (taki przedział czasowy tutaj rozpatrywałem). Wybrałem model k-średnich w celu podziału danych (z ustalonym segmentem wiatru, jak i temperatury otoczenia!) na skupienia. W wyborze liczby grup pomagał wykres sumy kwadratów dla poszczególnych wartości parametru `n_clusters`, którą oczywiście chcemy minimalizować, ale z rozsądkiem. Nie zawsze jednak była to optymalna liczba i czasem zmniejszałem liczbę skupień dla lepszej interpretacji wyników.

Model zbudowany jest na danych wystandaryzowanych i opiera się na 7 zmiennych opisujących stan techniczny poszczególnych części turbiny:

- `Nac_Temp_Avg` – średnia temperatura w gondoli,
- `Rtr_RPM_Avg` – średnie obroty wirnika,
- `Gear_Bear_Temp_Avg` – średnia temperatura w łożysku przekładni na wale wysokoobrotowym,
- `Gen_RPM_Avg` – średnie obroty generatora,
- `Gen_Bear_Temp_Avg` – średnia temperatura w pierwszym łożysku generatora (po stronie bez napędu),
- `Gen_Bear2_Temp_Avg` – średnia temperatura w drugim łożysku generatora (po stronie z napędem),
- `Hyd_Oil_Temp_Avg` – średnia temperatura oleju w grupie hydraulicznej,
- `Gear_Oil_Temp_Avg` – średnia temperatura oleju w przekładni.

Jasne jest, że niektóre z powyższych zmiennych są ze sobą mocno skorelowane, jak np. obroty wirnika i obroty generatora albo średnia temperatura gondoli z temperaturą innych części wewnętrznych. Z pomocą przychodzi tutaj wcześniej wspomniane PCA, które pozwala określić wkład poszczególnych cech na występującą zmienność w zbiorze danych i dodatkowo jesteśmy w stanie zredukować wyniki modelu k-średnich do dwóch wymiarów i przedstawić je na wykresie z nowym układem współrzędnych (PC1 - pierwsza składowa główna, PC2 - druga składowa główna).

Przykład.
Analiza obserwacji na 2 godziny przed awarią dla wieży T09:



Widoczne punkty to obserwacje, dla których prędkość wiatru była co najmniej równa 9 m/s, a temperatura otoczenia mieściła się w przedziale $[24, 41)$. Na pierwszy rzut oka widzimy wyraźny podział na dwie grupy – cluster=1 i cała reszta. W grupie pierwszej średnie obroty wirnika są bliskie zera, co oznacza brak pracy turbiny. Natomiast w drugiej grupie, gdzie wiatrak generuje prąd, mamy dodatkowy podział na dwie podgrupy: cluster=0 i cluster=2. Tutaj w klastrze drugim wystąpiła awaria łożyska generatora i widzimy, że obserwacje na 2 godziny przed samą awarią również wpadają do tego skupienia,

zatem możemy powiedzieć, że ta grupa opisuje stan bliski awarii (znacznie podwyższone temperatury łożysk generatora). W tym wypadku dobrym pomysłem było by wprowadzenie systemów alarmujących.

Podobne analizy wykonałem dla każdej turbiny z osobna, dla różnych podziałów ze względu na prędkość wiatru i temperaturę otoczenia (tam, gdzie występowały jakieś awarie) i na tej podstawie możemy wysunąć następujące wnioski:

- **Turbina T01:** podział na 5 skupisk, 2 z nich opisują praktycznie wyłączony wiatrak (w jednej wystąpiła awaria przekładni – tutaj zanotowano znaczny spadek średniej temperatury oleju w przekładni), w pozostałych grupach, gdzie wiatrak pracuje, można wyróżnić klaster z podwyższoną temperaturą łożyska generatora – tam wystąpiła również awaria transformatora.
- **Turbina T06:** podział na 3 skupiska, w jednym z nich wystąpiła awaria generatora – przyczyna: podwyższona temperatura w łożyskach generatora.
- **Turbina T07:** w tym przypadku wystąpiło dużo różnych awarii i ciężko wysunąć konkretne wnioski, bo obserwacje z odmiennych grup nachodzą na siebie.
- **Turbina T11:** dwa różne podziały w zależności od wiatru i temperatury otoczenia – dla małego wiatru i wysokiej temperatury mamy podział na 4 skupiska, gdzie w jednym wystąpiła awaria grupy hydraulicznej (hamulców); dla wyższego wiatru w przedziale $[5.5, 6)$ i niższej temperatury otoczenia mamy podział na 2 skupiska, gdzie również w jednym wystąpiła awaria hamulców grupy hydraulicznej.
- Ogólnie najwięcej awarii miało miejsce, gdy wiatr przekraczał 9 m/s, a temperatura otoczenia wynosiła ponad 24 stopnie.
- Największy wkład na podział mają temperatury poszczególnych części gondoli (w szczególności przekładni i generatora).

4.3 Analiza starzenia turbin

Ostatnim zadaniem było sprawdzenie, czy w danych można zauważyć efekt starzenia się wiatraków, to znaczy czy występuje spadek jakości wraz z upływem czasu. Będziemy badać zależność prądu aktywnego oraz jego zmianę w stosunku do poprzedniego interwału czasowego.

Zajmiemy się danymi oczyszczonymi, dla których prędkość wiatru mieści się w przedziale $[4, 9)$ i będziemy rozpatrywać segmenty wiatru co 0.5 m/s. Ponadto ustalimy również temperaturę otoczenia (5 segmentów), do wyboru tutaj są dwie metody: k-średnich lub kwantyle. Wybieramy tylko te obserwacje, dla których prąd aktywny jest dodatni i ograniczony z góry przez 200000 Wh.

Tak otrzymane ramki danych dalej dzielimy na mniejsze kostki, każda o ustalonej wielkości, np. $N = 70$. Na tych małych kostkach danych uruchamiamy algorytm k-medoidów (ustalamy liczbę skupień równą 3: mała moc, średnia moc i duża moc), a otrzymane centroidy nanosimy na wykres razem z punktami centralnymi. Na rysunku 3 znajdują się przykładowe wyniki dla turbiny T06.

Wynikowe ramki danych składają się z kilkudziesięciu centroidów, indeks jest wyznacznikiem czasu – im większa liczba, tym nowsza sygnatura czasu (newest timestamp). Obserwacje dzielimy na dwie połowy: pierwsza połowa to obserwacje najstarsze, a druga najnowsze. Dla każdej z tych grup liczymy średnią z odległości euklidesowej centroidów od punktu centralnego danego skupienia (czerwone trójkąty).

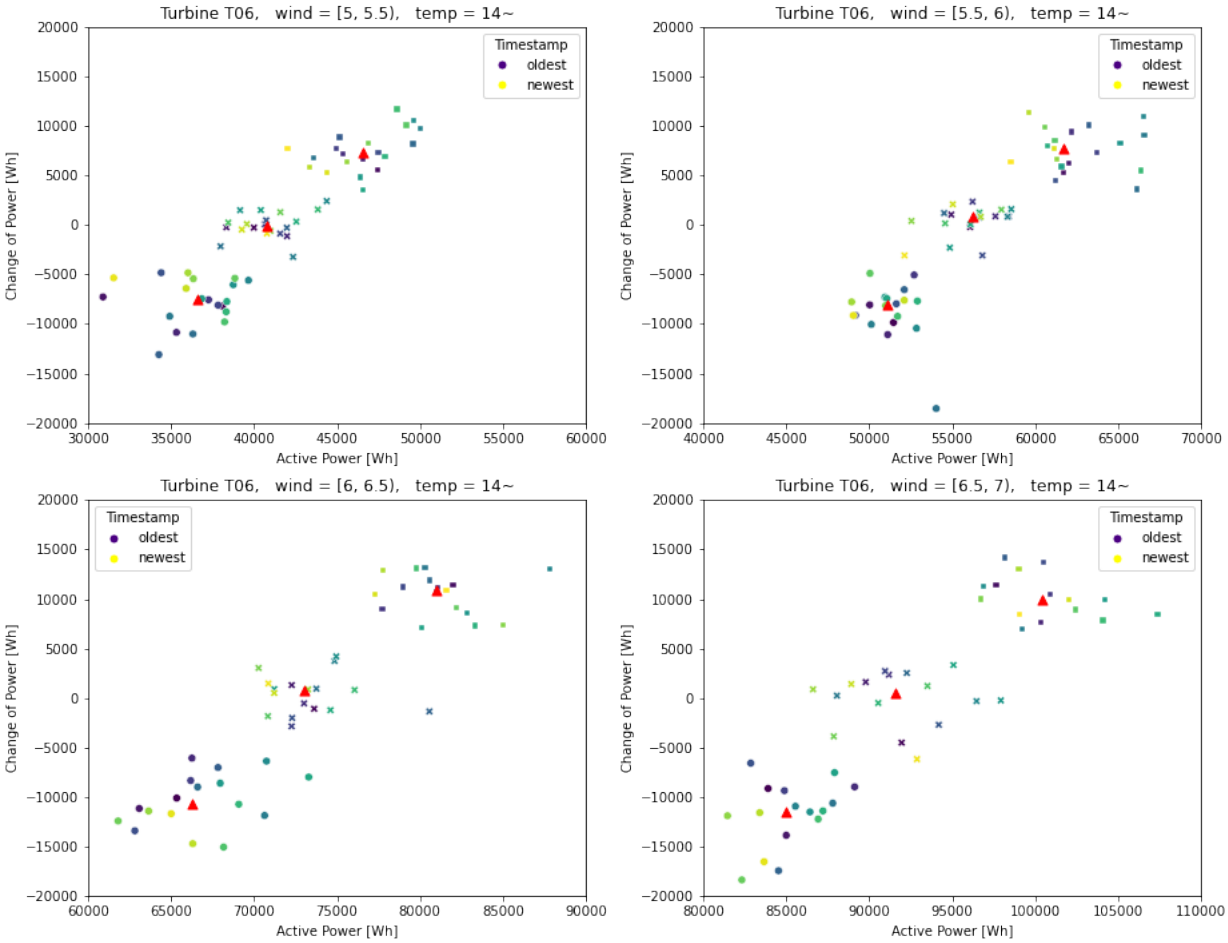
Ostateczny współczynnik D_k odległości dla k-tego skupienia:

$$D_k = \bar{N}_k - \bar{O}_k, \quad k = 1, 2, 3,$$

gdzie $\bar{N}_k$ – średnia odległość obserwacji najnowszych w k-tym skupieniu, $\bar{O}_k$ – średnia odległość obserwacji najstarszych w k-tym skupieniu.

Zauważmy, że jeżeli $D_k > 0$, to obserwacje nowsze są bardziej odchylone od środka niż obserwacje starsze, co świadczy o pewnym działaniu efektu starzenia. Ostatecznie współczynnikiem oceniającym może być średnia z wyliczonych różnic odległości:

$$D = \frac{D_1 + D_2 + D_3}{3}.$$



Rysunek 3: Analiza starzenia turbiny T06

Przykład.
Dla turbiny T06 wyniki przedstawiają się następująco:

```
D_0 = -619.33
D_1 = 302.71
D_2 = 1009.48
---------------
D = 230.95
```

```
D_0 = 854.45
D_1 = -360.65
D_2 = 995.68
---------------
D = 120.7
```

```
D_0 = -1332.26
D_1 = 509.36
D_2 = -997.35
---------------
D = -375.8
```

```
D_0 = -486.38
D_1 = 1412.46
D_2 = -462.29
---------------
D = 275.29
```

W trzech przypadkach współczynnik D jest większy od 0, więc możemy powiedzieć, że występuje jakiś efekt starzenia dla ustalonych konkretnie warunków atmosferycznych (por. tytuły na rysunku 3). Warto też spojrzeć, czy obserwacje układają się od najnowszych do najstarszych (od żółtego do fioletowego) – czasem ciężko stwierdzić, czy efekt starzenia rzeczywiście występuje i jasno go widać.

5 Podsumowanie

Niniejsza praca miała na celu pokazanie, że metody modelowania, a w szczególności klasteryzacja, są pomocne w diagnostyce sygnałów z turbin wiatrowych.

Dzięki analizie skupień jesteśmy w stanie zbadać momenty, w których turbiny operują w podobny sposób, możemy wyróżnić różne regiony pracy turbiny. Ponadto algorytmy te mogą posłużyć nam do klasyfikacji awarii, co może być przydatne, na przykład do budowania systemów alarmujących, bazujących na parametrach technicznych poszczególnych komponentów turbiny. Możemy również zbadać, czy upływający czas ma wpływ na ilość generowanego prądu.

Literatura

[1] D. Menezes, M. Mendes, J.A. Almeida, T. Farinha, *Wind Farm and Resource Datasets: A Comprehensive Survey and Overview*, Energies 2020, 13(18), 4702

[2] M. Gibiec, T. Barszcz, M. Bielecka, *SELECTION OF CLUSTERING METHODS FOR WIND TURBINES OPERATIONAL DATA*, DIAGNOSTYKA, 4(56)/2010

[3] D. Arthur, S. Vassilvitskii, *k-means++: The Advantages of Careful Seeding*, SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms

[4] R. Bro, Age K. Smilde, *Principal component analysis*, Anal. Methods, 2014, 6, 2812

Oświadczam, że niniejszą pracę wykonałem samodzielnie, co poświadczam własnoręcznym podpisem.

Marek Wiese