Praca magisterska

Marcin Bartodziej

4 maja 2025

Spis treści

1 Opis eksperymentów **2**

2 Dane: Computers **2**
2.1 Wstępny przegląd danych 2
2.2 Ogólny przegląd metod 4
2.3 Wpływ parametrów na poszczególne metody 5
2.3.1 Metoda Symbolic Aggregate approXimation 5
2.3.2 Metoda oparta na funkcji autokorelacji 5
2.4 Analiza stabilności metod 6

3 Dane: Car **8**
3.1 Wstępny przegląd danych 8
3.2 Ogólny przegląd metod 9
3.3 Wpływ parametrów na poszczególne metody 9
3.3.1 Metoda Symbolic Aggregate approXimation 9
3.3.2 Metoda oparta na funkcji autokorelacji 9
3.4 Analiza stabilności metod 9

1 Opis eksperymentów

W niniejszej pracy rozważono różne sposoby mierzenia odległości dla algorytmu najbliższego sąsiada (1-NN). Pod uwagę zostały wzięte metody:

- Symbolic Aggregate approXimation (SAX),
- metoda oparta na funkcji autokorelacji (ACF),
- metoda oparta na wzajemnej corelacji (CCor),
- Complexity-Invariant Distance (CID),
- odległość Euklidesowska,
- Dynamic Time Warping (DTW),
- odległość bazująca na współczynniku Fouriera.

W ramach części eksperymentalnej przeprowadzono trzy serie analiz mających na celu ocenę skuteczności wybranych metod ekstrakcji cech z szeregów czasowych w kontekście zadania klasyfikacji. Wszystkie eksperymenty przeprowadzono z wykorzystaniem algorytmu najbliższego sąsiada (1-NN) oraz uprzednio znormalizowanych danych. Analizę przeprowadzono na dwóch rzeczywistych zbiorach danych: `computers`, zawierającym obserwacje należące do dwóch klas, oraz `car`, obejmującym cztery klasy.

Pierwszym etapem było zbadanie wpływu doboru parametrów metody Symbolic Aggregate approXimation (SAX) na jej skuteczność. Skoncentrowano się na dwóch kluczowych parametrach: liczbie segmentów w, na które dzielony jest szereg czasowy, oraz liczbie poziomów dyskretyzacji α. W eksperymencie przetestowano kombinacje parametrów $\alpha \in \{2, 4, 8, 12, 16, 20, 24\}$ oraz $w \in \{2, 4, 10, 20, 30, 45, 60\}$. Dla każdej kombinacji przeprowadzono klasyfikację przy użyciu metody 1-NN, a następnie obliczono dokładność predykcji. Uzyskane wyniki zilustrowano w postaci mapy cieplnej, co umożliwiło wizualną ocenę wpływu wartości parametrów na skuteczność metody SAX.

W drugim eksperymencie przeanalizowano metodę opartą na funkcji autokorelacji (ACF), koncentrując się na wpływie parametru `lag.max`, oznaczającego maksymalną liczbę opóźnień uwzględnianych przy obliczaniu ACF. Dla wartości `lag.max` $\in \{1, 2, \ldots, 15\}$ przeprowadzono klasyfikację i obliczono dokładność predykcji. Wyniki przedstawiono na wykresie liniowym, umożliwiającym identyfikację optymalnego ustawienia parametru.

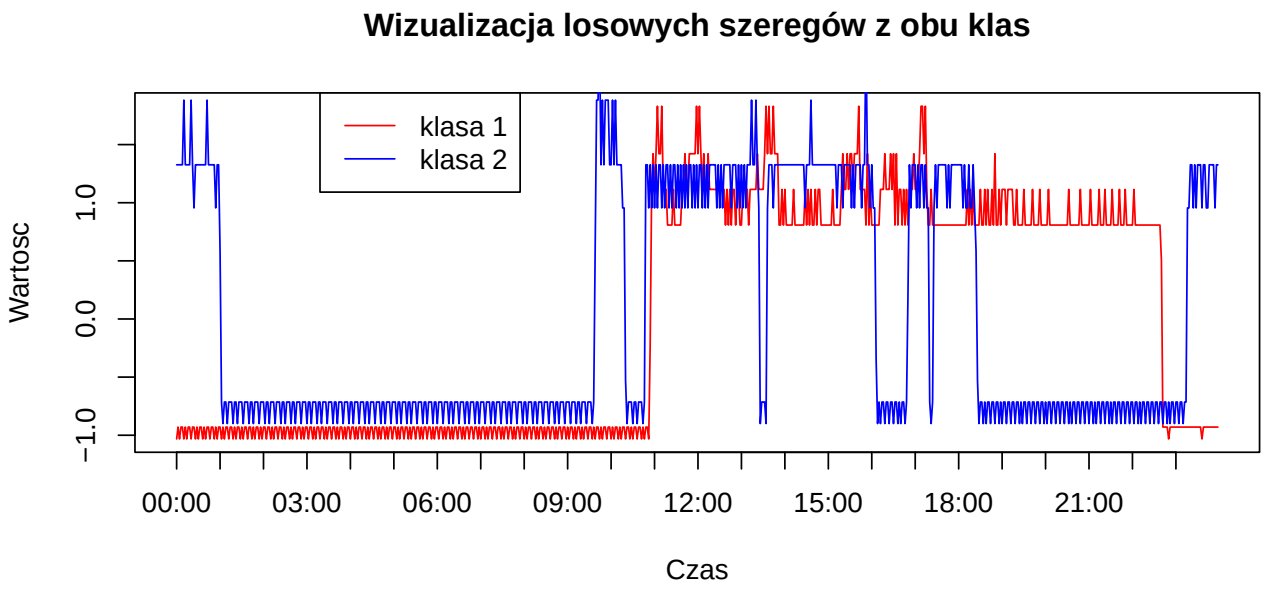
W ostatnim kroku przeprowadzono analizę stabilności metod poprzez wielokrotne losowe podziały danych na zbiór treningowy i testowy. Każdy z eksperymentów został powtórzony 20 razy przy losowym podziale danych. Dla każdego powtórzenia obliczono dokładność klasyfikacji, a zebrane wyniki przedstawiono w formie wykresów pudełkowych (boxplotów). Taka prezentacja pozwala na ocenę nie tylko średniej skuteczności, ale także rozrzutu wyników, co jest istotne przy analizie stabilności i niezawodności badanych metod.

2 Dane: Computers

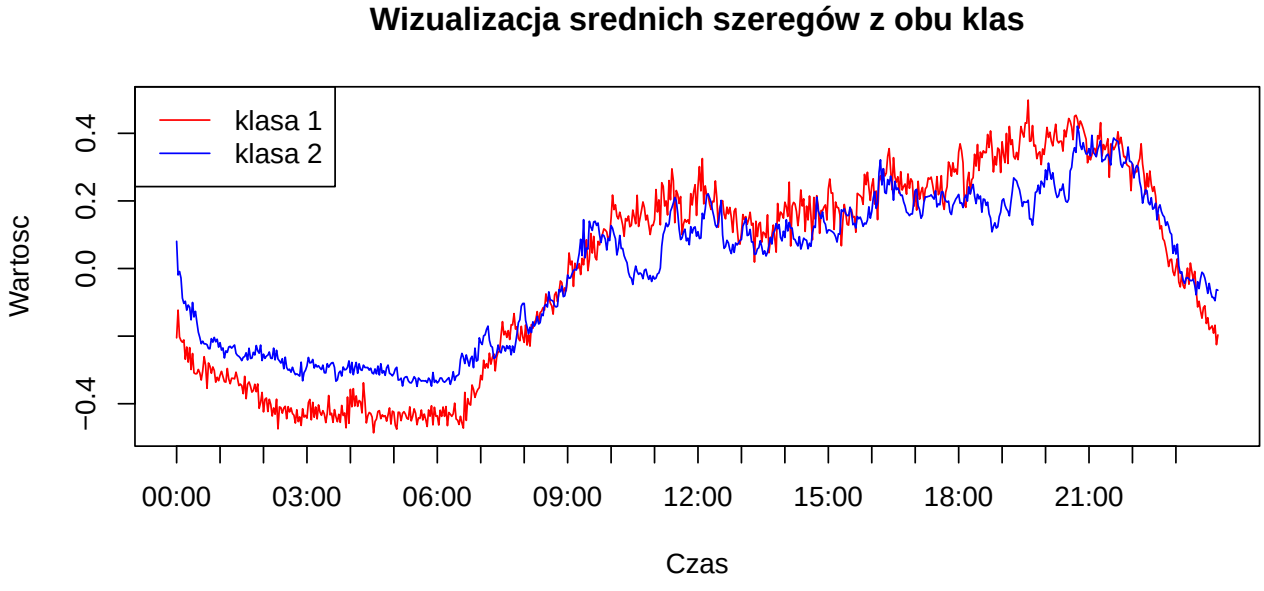
2.1 Wstępny przegląd danych

Dane pochodzą z badania sponsorowanego przez rząd Wielkiej Brytanii, zatytułowanego „Powering the Nation". Celem projektu było zebranie informacji na temat zachowań konsumentów związanych z zużyciem energii elektrycznej w gospodarstwach domowych, aby wspomóc działania na

rzecz ograniczenia śladu węglowego w kraju. Zarówno zbiór treningowy jak i testowy zawierają pomiary zarejestrowane w 250 gospodarstwach domowych, wykonywane co dwie minuty. Każdy szereg czasowy składa się z 720 obserwacji, odpowiadających 24 godzinom pomiarów wykonywanych w dwuminutowych odstępach. Klasy reprezentują dwa typy urządzeń: komputery stacjonarne (Desktop) i laptopy (Laptop). Dane są znormalizowane, nie zawierają brakujących obserwacji oraz klasy są równoliczne (250/250).



Rysunek 1: Wizualizacja losowych szeregów z klasy 1 i 2



Rysunek 2: Wizualizacja średnich szeregów z klasy 1 i 2

Z rysunków 1 i 2 możemy wywnioskować, że szeregi z klasy 2 przyjmują większe wartości w godzinach 00:00 - 07:00 oraz mniejsze wartości od 18:00 do 21:00. Zauważalny jest również spadek przyjmowanych wartości w godzinach około 10:00 - 12:00 dla szeregów klasy 2. Dla obu klas widzimy trend wzrostowy w czasie 07:00 - 10:00, chociaż jest on bardziej zauważalny w przypadku szeregów klasy 1. Ponadto szeregi z obu klas notują spadki wartości w godzinach 21:00 - 24:00.

2.2 Ogólny przegląd metod

Poniższa tabela przedstawia czas wykonania poszczególnych metod oraz ?miar dokładności?

	time (sec)	accuracy	sensitivity	specificity	precision	recall
ACF	11.20	0.59	0.54	0.65	0.60	0.54
DTW	2696.94	0.66	0.72	0.59	0.64	0.72
Euclidean	14.42	0.58	0.51	0.64	0.59	0.51
SAX	9.83	0.50	0.50	0.50	0.50	0.50
CCor	93.02	0.52	0.62	0.42	0.52	0.62
CID	5.69	0.60	0.54	0.65	0.61	0.54
Fourier	6.20	0.58	0.51	0.64	0.59	0.51

Tabela 1: Porównanie wyników klasyfikacji metod na zbiorze computers.

Na podstawie wyników przedstawionych w Tabeli 1 można sformułować następujące obserwacje dotyczące skuteczności i wydajności poszczególnych metod:

1. **Najwyższą dokładność** (accuracy = 0,66) osiągnęła metoda **DTW**, co czyni ją najskuteczniejszą spośród rozważanych metod pod względem ogólnej poprawności klasyfikacji. Również w zakresie czułości (sensitivity) oraz współczynnika recall metoda DTW uzyskała najwyższe wartości (0,72), co wskazuje na jej dużą skuteczność w wykrywaniu przypadków pozytywnych.

2. **Metody ACF, CID i Fourier** osiągnęły porównywalne wyniki dokładności (w zakresie od 0,58 do 0,60), przy czym charakteryzowały się znacznie krótszym czasem działania niż DTW. Przykładowo, metoda CID wykonała się w zaledwie **5,69 sekundy**, co czyni ją atrakcyjną alternatywą pod względem kompromisu między jakością a wydajnością.

3. **Metoda SAX** uzyskała najniższe wyniki we wszystkich miarach klasyfikacyjnych (accuracy = 0,50, sensitivity = 0,50, specificity = 0,50), co sugeruje, że przy zastosowanych parametrach nie nadaje się do klasyfikacji danych z tego zbioru.

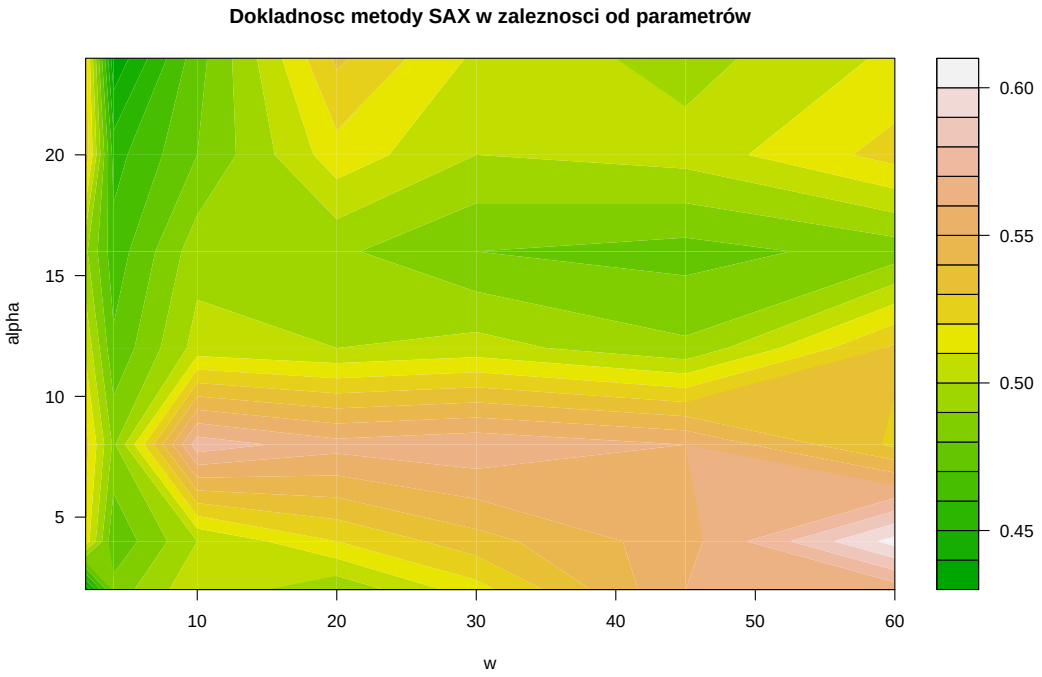
4. **Czas działania** metod istotnie się różnił. Najdłużej wykonywała się metoda DTW - około 45 minut, co może być czynnikiem ograniczającym jej praktyczne zastosowanie ze względu na złożoność obliczeniową.

5. **Metody Euclidean i Fourier** osiągnęły umiarkowaną dokładność (ok. 0,58), przy krótkim czasie działania (od 6 do 14 sekund), co czyni je dobrymi kandydatami do dalszego rozważenia w kontekście klasyfikatorów bazowych.

2.3 Wpływ parametrów na poszczególne metody

2.3.1 Metoda Symbolic Aggregate approXimation

W tej sekcji zbadano wpływ parametrów `w` i `alpha` na metodę SAX. Wzięto pod uwagę $\alpha \in \{2, 4, 8, 12, 16, 20, 24\}$ oraz $w \in \{2, 4, 10, 20, 30, 45, 60\}$.



Rysunek 3: Mapa ciepła przedstawiająca dokładność metody SAX w zależności od parametrów

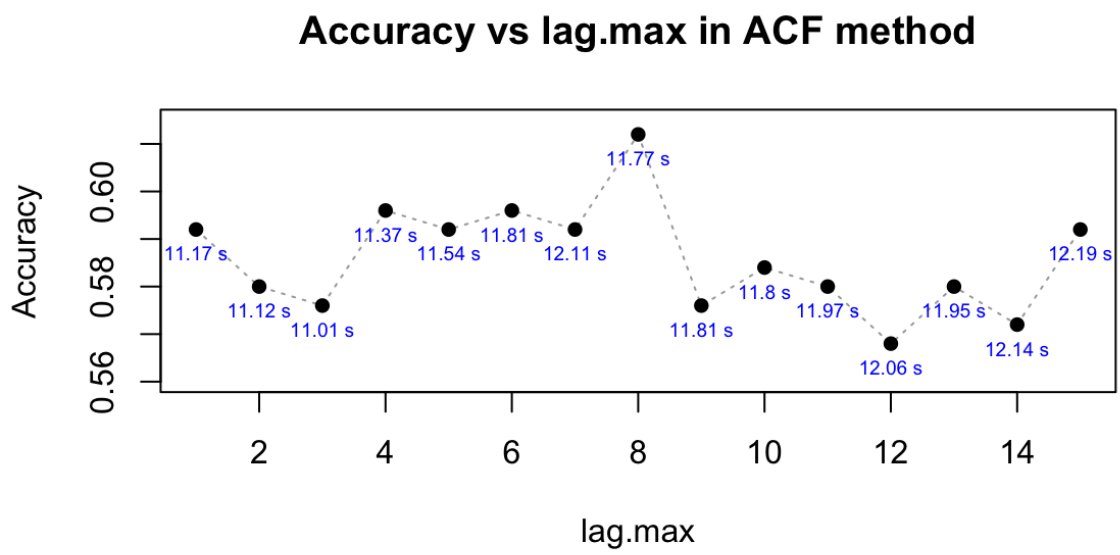
Na podstawie mapy ciepła przedstawionej na rysunku 3 można zauważyć, że dokładność metody SAX silnie zależy od odpowiedniego doboru parametrów `alpha` i `w`. Najlepsze wyniki osiągane są przy umiarkowanych wartościach `alpha` (około 10–12) oraz mniejszych wartościach `w` (ok. 10–20). W tym obszarze mapa wskazuje na najwyższy poziom dokładności - około 0,6.

Wraz ze wzrostem wartości `w` przy niskich `alpha`, dokładność istotnie spada. Z kolei zbyt wysokie wartości `alpha` również nie wpływają korzystnie na skuteczność metody, co może wynikać z nadmiernego rozdrobnienia danych na symbole, co utrudnia modelowi identyfikację wzorców.

Podsumowując, dla metody SAX optymalny zakres parametrów to umiarkowane `alpha` oraz relatywnie niskie `w`. Dobór tych parametrów ma kluczowe znaczenie dla osiągnięcia wysokiej dokładności klasyfikacji.

2.3.2 Metoda oparta na funkcji autokorelacji

W tej sekcji zbadano wpływ paramertru `lag.max` na metodę ACF. Rozpatrzono `lag.max` $\in \{1, 2, \ldots, 15\}$.

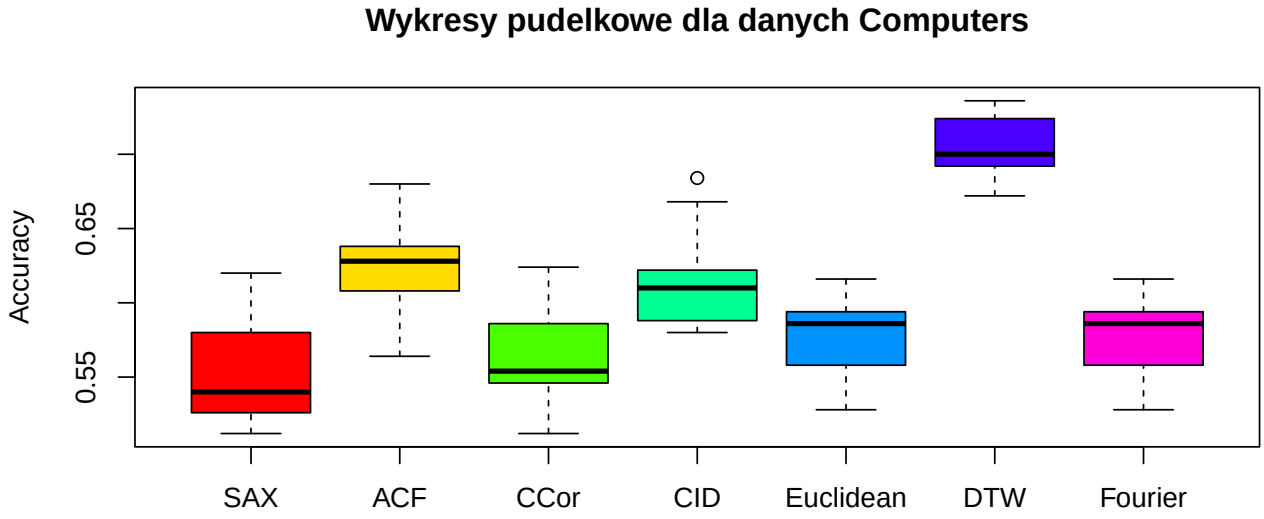


Rysunek 4: Wykres dokładności metody ACF w zależności od parametru lag.max

Jak możemy zauważyć na rysunku 4 dokładność metody SAX na tym zbiorze danych wacha się pomiędzy 0,57 a 0,61. Najwyższą dokładność mamy dla `lag.max` = 8. Dość dobre wyniki (w okolicach 0,59) notujemy też dla `lag.max` = {1, 4, 5, 6, 7, 15}. Ponadto warto zwrócić uwagę, że czasy wykonania dla różnych wartości parametru są zbliżone do siebie.

2.4 Analiza stabilności metod

W tej sekcji przeprowadzono analizę stabilności metod poprzez 20-krotne losowe podziały danych na zbiór treningowy i testowy. Zostały zachowane oryginalne proporcje zbiorów treningowego i testowego (50/50). Dla każdego powtórzenia obliczono dokładność klasyfikacji. Wyniki przedstawiono w formie wykresów pudełkowych i tabeli ze średnią dokładnością oraz odchyleniem standardowym wyników. W metodzie SAX ustawiono parametry `w`=10 i `alpha`=8, natomiast w metodzie ACF zastosowano `lag.max`=8.



Rysunek 5: Wykresy pudełkowe dokładności dla rozpatrywanych metod, przy 20-krotnym powtórzeniu losowego podziału danych

	Średnia	Odchylenie standardowe
SAX	0.55	0.032
ACF	0.63	0.027
CCor	0.56	0.028
CID	0.61	0.028
Euclidean	0.58	0.027
DTW	0.71	0.020
Fourier	0.58	0.026

Tabela 2: Średnia i odchylenie standardowe dokładności dla badanych metod

Na podstawie wykresu pudełkowego 5 oraz wyników przedstawionych w Tabeli 2 można zauważyć wyraźne różnice w skuteczności i stabilności rozpatrywanych metod klasyfikacji.

Najwyższą średnią dokładność klasyfikacji uzyskała metoda **DTW** (0,71), która jednocześnie charakteryzuje się najniższym odchyleniem standardowym (0,02). Oznacza to, że DTW nie tylko osiąga najlepsze wyniki, ale również wykazuje bardzo wysoką stabilność względem losowego podziału danych.

Na drugim miejscu pod względem dokładności znajduje się metoda **ACF** ze średnią dokładnością 0,63. Cechuje się stosunkowo niską zmiennością wyników z odchyleniem standardowym 0,027. Podobne właściwości wykazuje metoda **CID**. Tutaj mamy średnią 0,61 i odchylenie 0,028, co czyni ją również godną uwagi.

Najniższą średnią dokładność na poziomie 0,55 uzyskała metoda **SAX**. Dodatkowo wykazuje ona najwyższe odchylenie standardowe na poziomie 0,032, co świadczy o jej niskiej stabilności. Również metoda **CCor**, mimo niskiej dokładności (średnio 0,56), cechuje się dość dużą zmiennością wyników na poziomie 0,028.

Pozostałe metody, takie jak **Euclidean** i **Fourier**, osiągają umiarkowaną średnią dokładność 0,58 i niskie odchylenia standardowe - odpowiednio 0,027 i 0,026. Wskazuje to na ich stabilność, choć niekoniecznie wysoką skuteczność.

Podsumowując, metoda **DTW** zdecydowanie wyróżnia się pod względem zarówno skuteczności, jak i niezawodności. Na przeciwnym biegunie znajduje się metoda **SAX**, która wypada najsłabiej w obu aspektach.

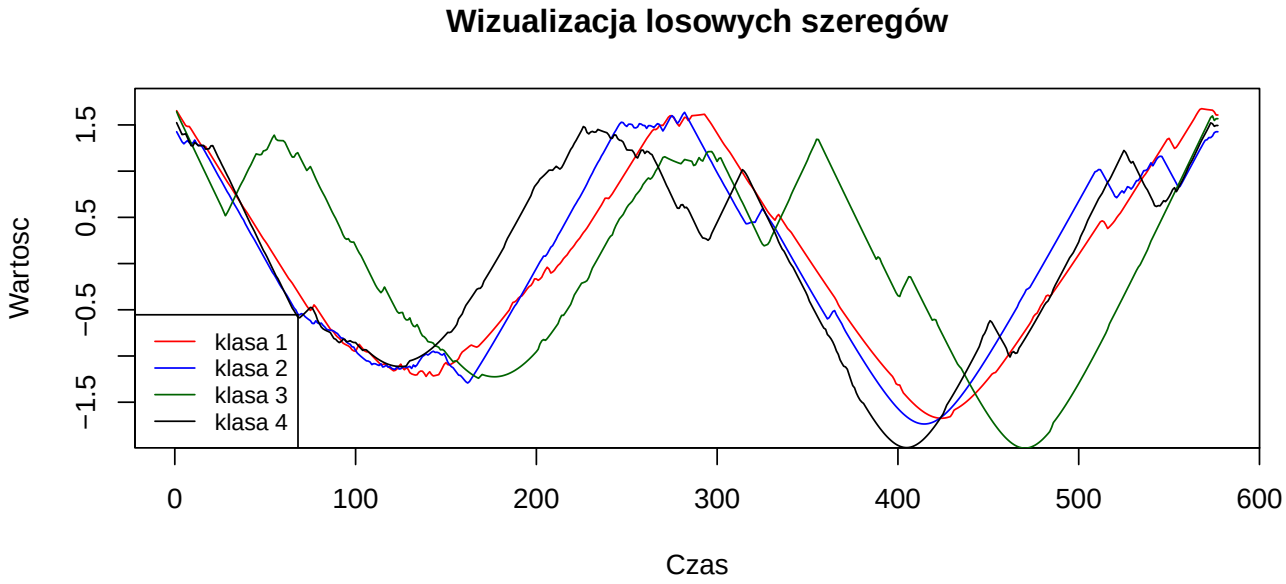
3 Dane: Car

3.1 Wstępny przegląd danych

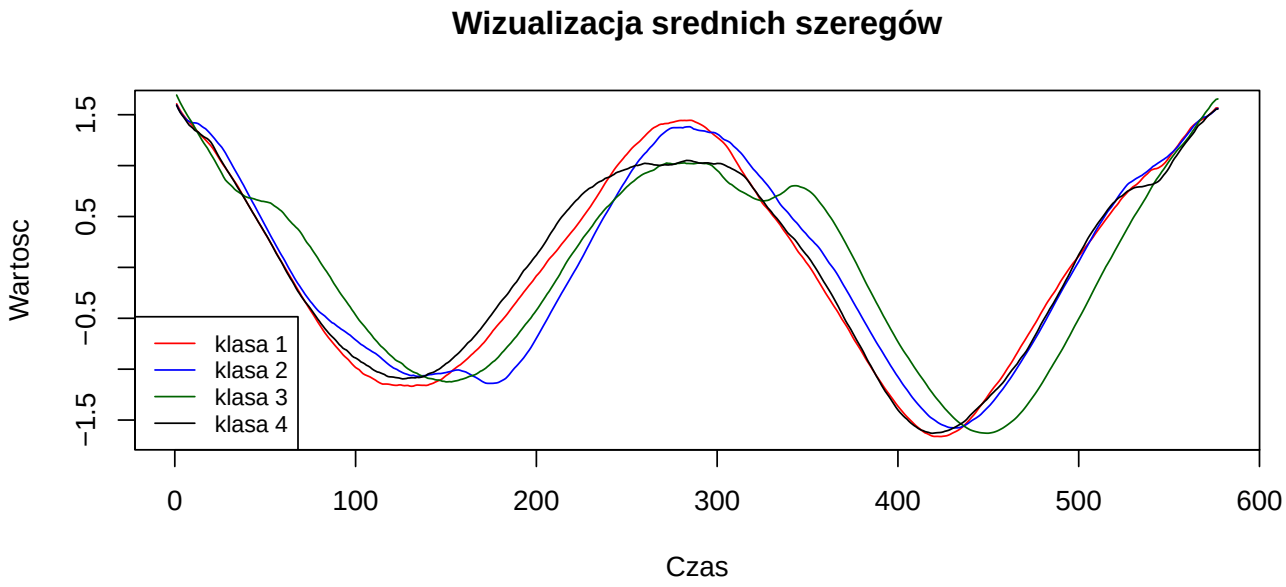
Wykorzystano dane pozyskane z nagrań wideo przedstawiających ruch uliczny w warunkach plenerowych. Nagrania te zostały zarejestrowane w rozdzielczości 320 × 240 pikseli. Na podstawie informacji o ruchu wyodrębniono kontury poruszających się pojazdów. ZAstosowana metoda umożliwia ekstrakcję sylwetek pojazdów, jednak nie uwzględnia wpływu cieni, co skutkuje częściowym zniekształceniem dolnych fragmentów uzyskanych kształtów.

Dane obejmują cztery klasy pojazdów: sedan, pickup, minivan oraz SUV. Dla każdej z klas pozyskano po 30 próbek, co łącznie daje 120 obserwacji. W celu ograniczenia wpływu szumu na dalsze przetwarzanie, każdy wyodrębniony kształt został poddany filtracji z wykorzystaniem filtra Gaussa o odchyleniu standardowym równym 5. Dane są znormalizowane.

Zestaw danych został następnie losowo podzielony na równoliczne zbiory treningowy oraz testowy.



Rysunek 6: Wizualizacja losowych szeregów ze wszystkich klas



Rysunek 7: Wizualizacja średnich szeregów ze wszystkich klas

3.2 Ogólny przegląd metod

3.3 Wpływ parametrów na poszczególne metody

3.3.1 Metoda Symbolic Aggregate approXimation

3.3.2 Metoda oparta na funkcji autokorelacji

3.4 Analiza stabilności metod