TESTY

Elżbieta Jowik

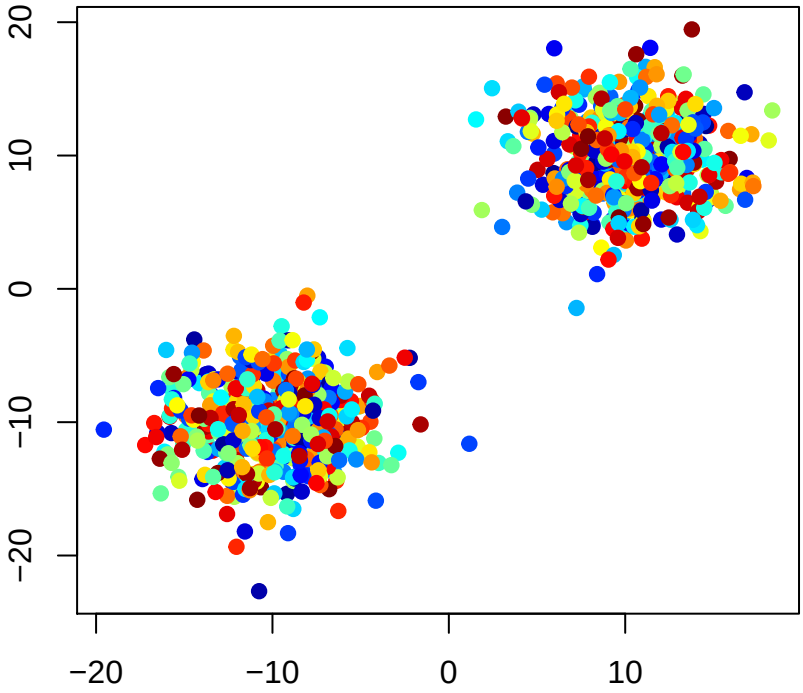
Ten plik zawiera testy poprawności metod: Mnn(), Mnn_graph(), Laplacian_eigen(), spectral_clustering(), zaimplementowanych w pliku **spectral.R**.
Testy są przeprowadzone na trzech własnych zbiorach danych zarówno z R^2, jak i R^3.
Oczywiście zbiory różnią się licznością danych, przyjmowanym kształtem, a w niektórych przypadkach liczbą wymiarów. Wszystkie z rozważanych zbiorów mają skoncentrowaną strukturę, gdyż umożliwia to sporządzenie wiarygodnych etykiet.

FIGURY 2D

1. SKUPISKA

Najprostszy do analizy skupień zbiór, wygenerowany przy pomocy rozkładu normalnego. Skupiska wygenerowane zostały dla 500 punktów każde.

a) INTERPRETACJA GRAFICZNA



b) TEST

Kryterium, jakie będziemy stosować do oceny poprawności zaimplementowanych metod będą etykiety, przygotowane na etapie tworzenia wymienionych w pracy zbiorów.
Do porównania zgodności wynikowego ciągu przypisań funkcji spectral_clustering() i danych z góry etykiet referencyjnych będziemy stosować:

1. **indeks Fowlkesa-Mallowsa** (dendextend::FM_index)

2. **skorygowany indeks Randa** mclust::adjustedRandIndex

Każdy z powyższych indeksów zwraca wartość równą **1**, jeśli dane podziały są równoważne. Im ich wartość jest dalej od 1, tym bardziej są one od siebie różne.

UWAGA! Na potrzeby testów będziemy rozważać maksymalne wartości zwróconych indeksów dla określonego (jednego) parametru M. Pełna analiza algorytmu z uwzględnieniem różnych parametrów znajduje się w raporcie.

Skorygowany indeks Randa

```
mclust::adjustedRandIndex(set_labels, result)
```

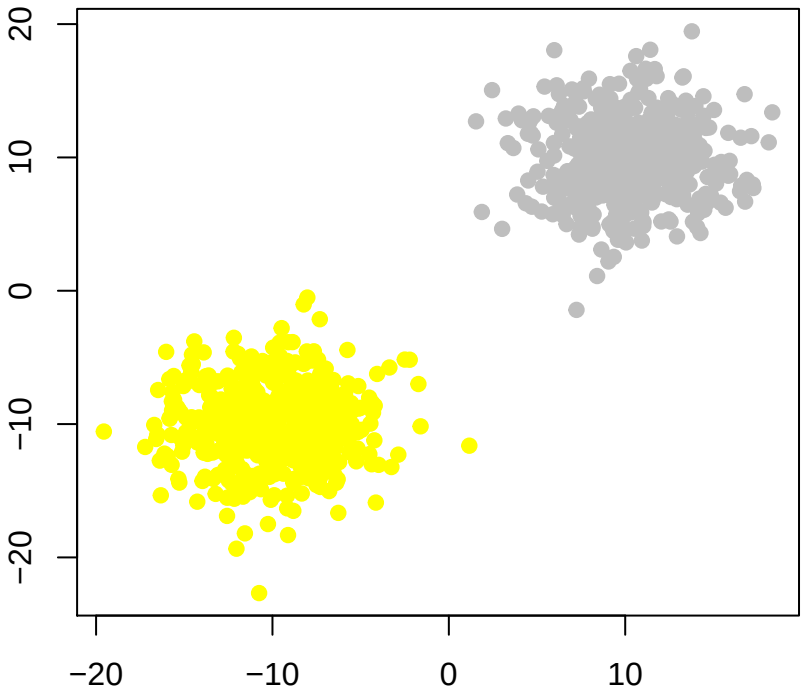
```
## [1] 1
```

Indeks Fowlkesa-Mallowsa

```
dendextend::FM_index(set_labels, result)
```

```
## [1] 1
## attr(,"E_FM")
## [1] 0.4994995
## attr(,"V_FM")
## [1] 5.02509e-07
```

WIZUALIZACJA OTRZYMANEGO PODZIAŁU



WNIOSKI:

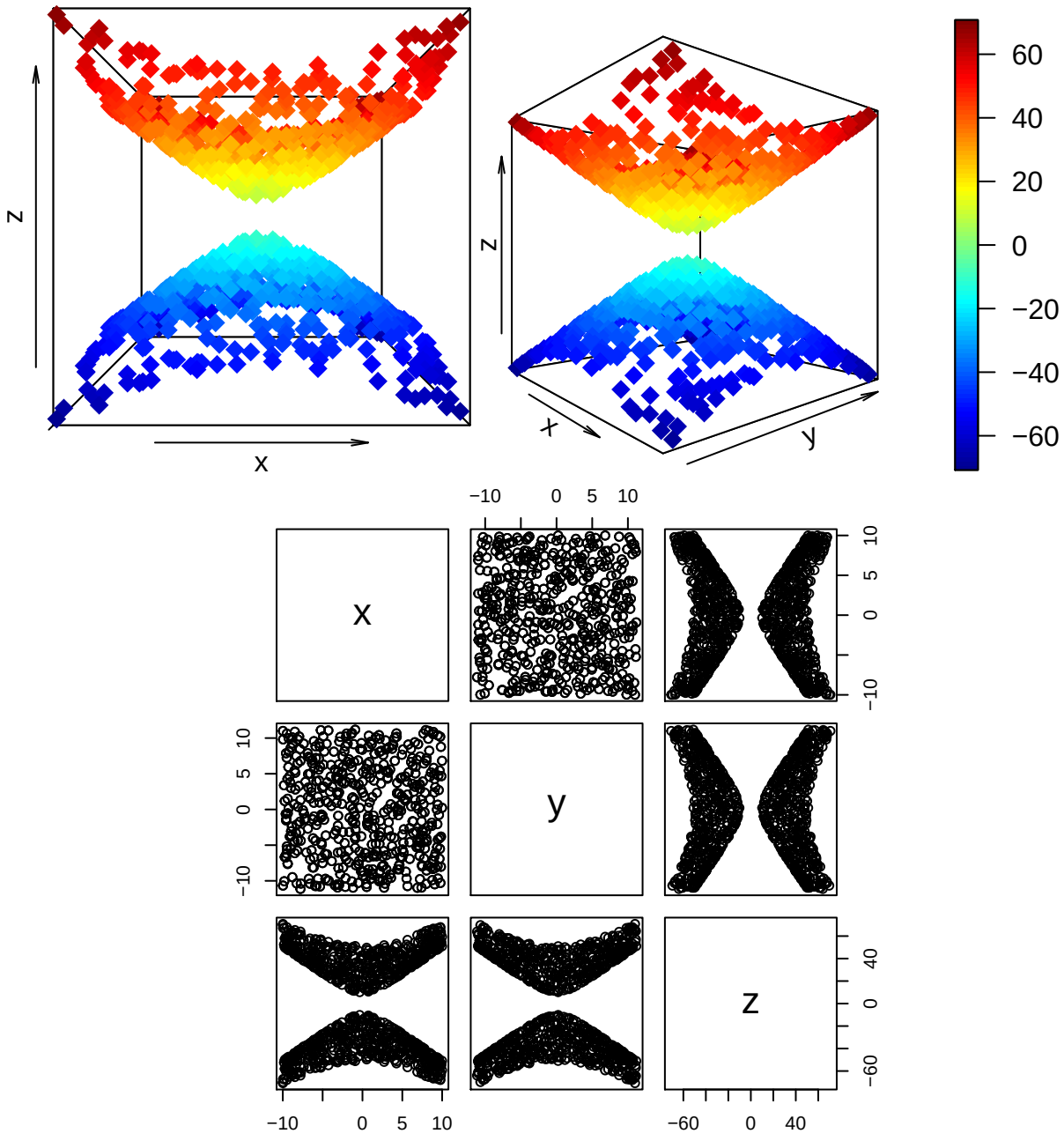
Oczywiście zarówno indeks Randa, jak i Fowlkesa-Mallowsa przyjmują różne wartości przy kolejnych kompilacjach. W większości jednak wartości tych indeksów zawierają się w przedziale [0.7, 1]. Na podstawie otrzymanych rezultatów możemy stwierdzić, że dla **prostych** zbiorów danych, w których występuje **wyraźny podział** punktów na podzbiory, algorytm działa zadowalająco, co jasno wskazuje na poprawność działania zaimplementowanych metod.

FIGURY 3D

1. HIPERBOLOIDA DWUPOWŁOKOWA

Trójwymiarowa hiperboloida dwupowłokowa, wygenerowana dla 1000 punktów przy użyciu równania powierzchni stożkowej wyrażonego we współrzędnych kartezjańskich.

a) Interpretacja graficzna



b) TEST

Skorygowany indeks Randa

```
mclust::adjustedRandIndex(something_labels, result)
```

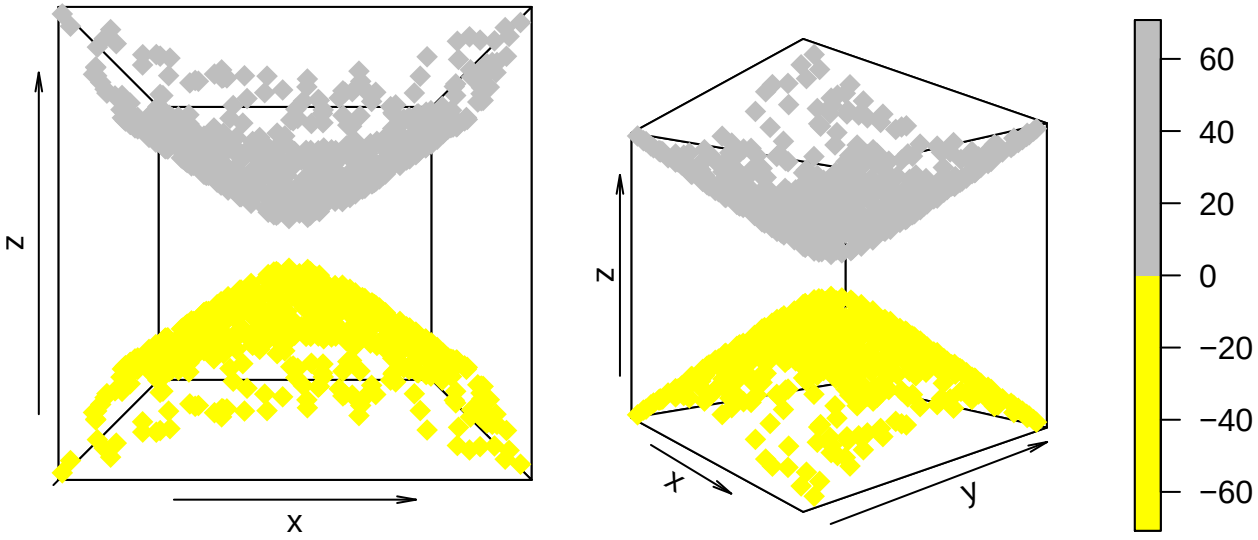
```
## [1] 1
```

Indeks Fowlkesa-Mallowsa

```
dendextend::FM_index(something_labels, result)
```

```
## [1] 1
## attr(,"E_FM")
## [1] 0.4994995
## attr(,"V_FM")
## [1] 5.02509e-07
```

WIZUALIZACJA OTRZYMANEGO PODZIAŁU



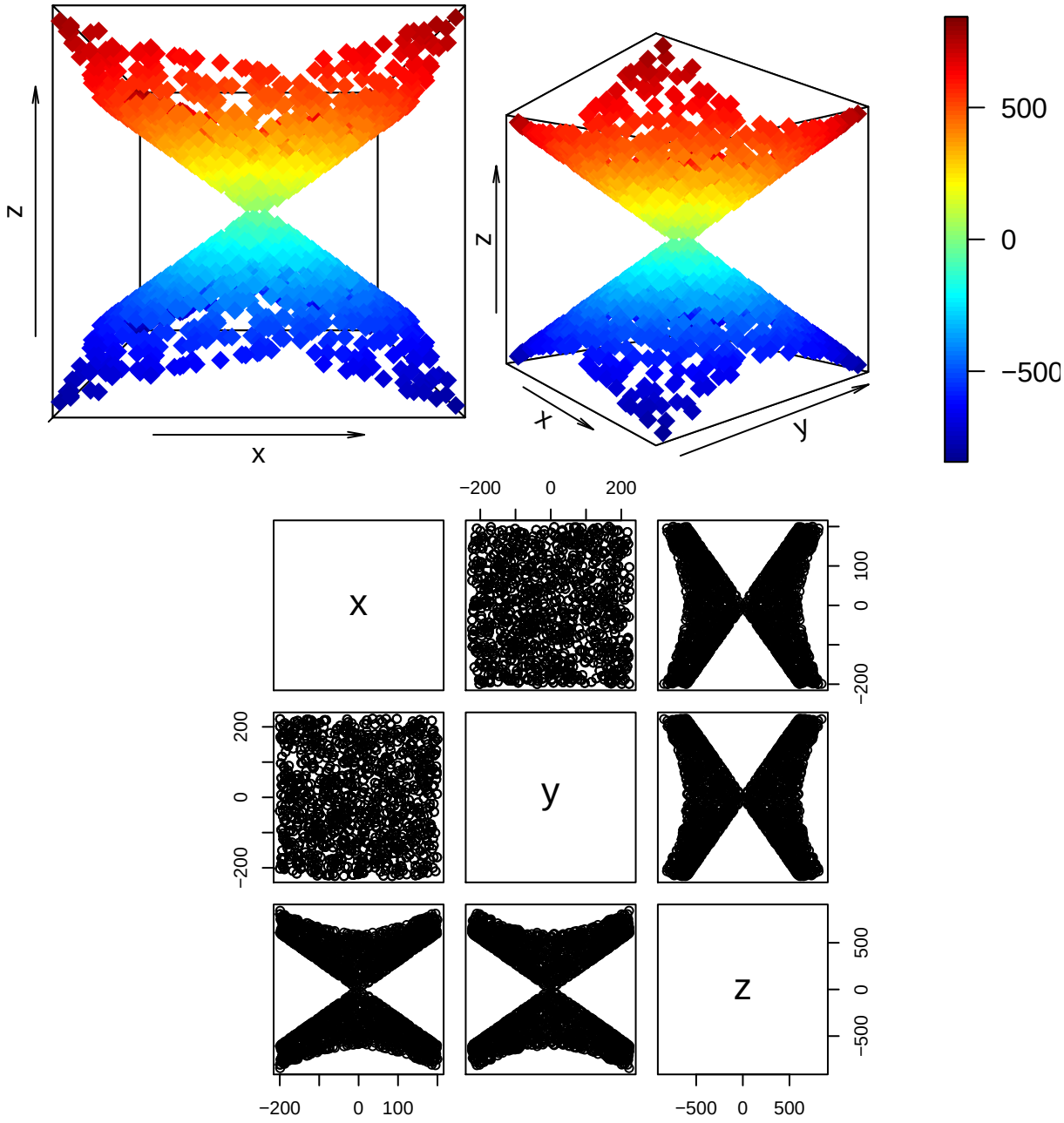
WNIOSKI:

Zbiór danych, reprezentujący hiperboloidę dwupowłokową, jest w pewnym sensie analogiczny do pierwszego z omawianych przykładów - skupisk. Bowiem zarówno tutaj, jak powyżej gołym okiem można wyodrębnić pewien podział. Podobnie jak w pierwszym przypadku zwracane indeksy są dość jednoznaczne, co sugeruje, że testowany algorytm dobrze radzi sobie z rozważanym zbiorem.

2. POWIERZCHNIA STOŻKOWA

Trójwymiarowa powierzchnia stożkowa wygenerowana dla 1 000 punktów. Na pierwszy rzut oka zbiór ten wydaje się być w pełni analogiczny do poprzedniego. Różnica jednak jest zasadnicza. Otóż zarówno w przypadku skupisk, jak i hiperboloidy pozdbiory, na które dzieliliśmy zbiory, były rozłączne. W tym przypadku nieznacznie się one nakładają, przez co podział zwracany przez testowany algorytm jest obarczony większym błędem niż w poprzednich.

a) INTERPRETACJA GRAFICZNA



b) TEST

Skorygowany indeks Randa

```
mclust::adjustedRandIndex(funnel_labels, result)
```

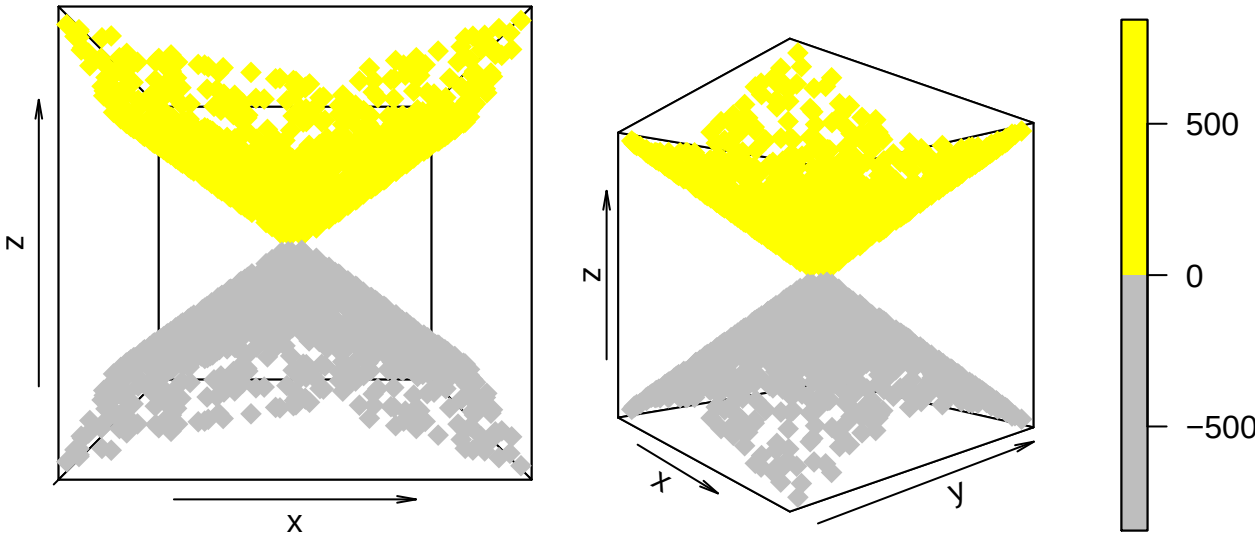
```
## [1] 1
```

Indeks Fowlkesa-Mallowsa

```
dendextend::FM_index(funnel_labels, result)
```

```
## [1] 1
## attr(,"E_FM")
## [1] 0.4997499
## attr(,"V_FM")
## [1] 1.253131e-07
```

WIZUALIZACJA OTRZYMANEGO PODZIAŁU



WNIOSKI:

Brak rozłączności podzbiorów powoduje, że wygenerowane indeksy nie są już tak jednoznaczne jak w poprzednich przypadkach. Jednocześnie odchylenia od oczekiwanego wyniku są na tyle nieznaczne, że nie poddają pod wątpliwość poprawności implementacji metod.

PODSUMOWANIE:

Na postawie przeprowadzonych testów mogę stwierdzić, że implementacja testowanych metod jest poprawna. Wartości zwracanych indeków są zadowalające, algorytm dobrze radzi sobie na zbiorach o skoncentrowanej strukturze. Pełna analiza algorytmu na zbiorach o odmiennych kształtach znajduje się w raporcie.