Porównanie algorytmów analizy skupień

PDU 2018/19 - praca domowa nr 2

Konrad Komisarczyk

Contents

W pracy zbadam jakość 4 różnych grup algorytmów analizy skupień:

- własnej implementacji algorytmu spektralnego (kod załączony w pliku `spectral.R`)
- Algorytmu *Genie* z pakietu `genie`.
- Metod hierarchicznych z bazowej funkcji `hclust`.
- Metody k-średnich z funkcji `kmeans`.

Zbadam jakość poszczególnych metod w zależności od udostępnianych przez funkcję parametrów tuningujących, czy różnych wariantów metody. Przyjrzę się wpływowi standaryzacji wejściowych danych za pomocą bazowej funkcji `scale` na jakość poszczególnych algorytmów. Ponadto porównam między sobą średnie wyniki różnych grup metod i wyniki najlepszych spośród wariantów każdej metody.

Badania przeprowadzę na wszystkich dostarczonych zbiorach benchmarkowych i przygotowanych przeze mnie zbiorach danych poza `by_powiat.data`.

Jakość algorytmów będę oceniał na podstawie oceny podobieństwa generowanych przez nie podziałów zbioru do danych podziałów referencyjnych przy użyciu indeksu Fowlkesa-Mallowsa (za pomocą funkcji `clues::adjustedRand` z parametrem `randMethod = "FM"`) i skorygowanego indeksu Randa (funkcja `mclust::adjustedRandIndex`).

1. Porównanie różnych wariantów poszczególnych algorytmów analizy skupień

1.1 Własna implementacja algorytmu spektralnego

Porównałem jakość algorytmu dla kilku kolejnych małych wartości parametru `M` (3-9), oraz kilku wiekszych (16, 20, 32, 44) oznaczającego liczbę krawędzi do najbliższych sąsiadów punktu uwzględnianych w grafie sąsiedztwa używanym w algorytmie.

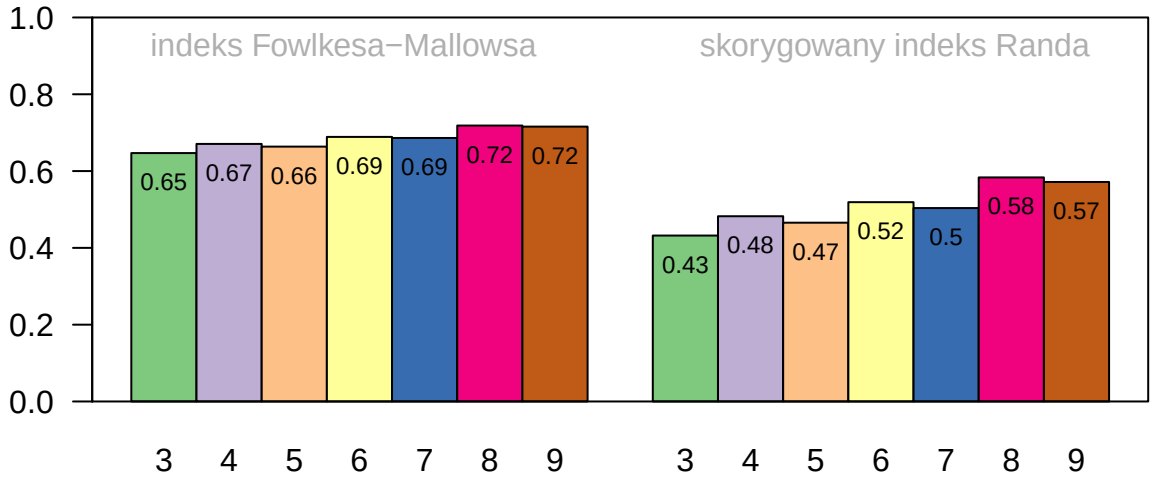

Figure 1: Ocena średnich jakości wyników algorytmu spektralnego w zależności od parametu M. (Dla małych wartości parametru M.)

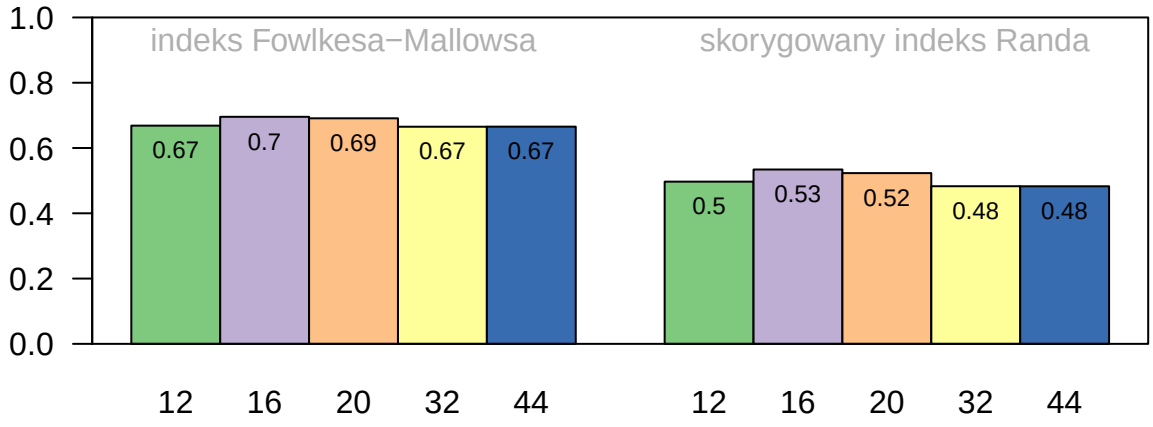

Figure 2: Ocena średnich jakości wyników algorytmu spektralnego w zależności od parametu M. (Dla większych wartości parametru M.)

Spośród porównanych parametrów algorytm najlepsze wyniki osiąga dla parametru M = 8.

Wpływ standaryzacji

W przypadku algorytmu spektralnego standaryzacja danych wejściowych ma mały, ale, w przeciwieństwie do wszystkich pozostałych badanych algorytmów, czasami pozytywny wpływ na jakość podziału.

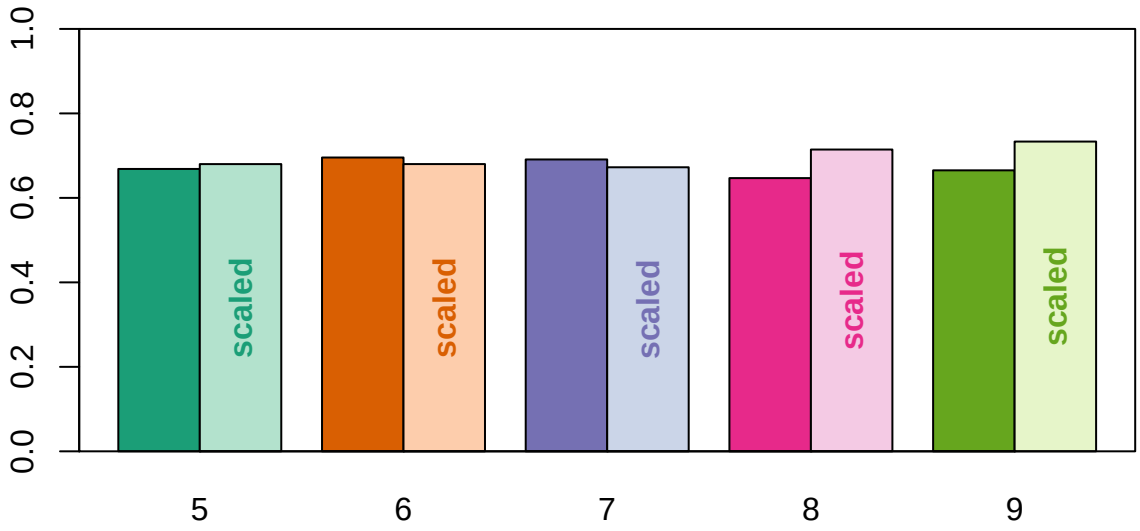

Figure 3: Średnia indeksów Randa dla własnej implementacji algorytmu spektralnego na nie standaryzowanych i standaryzowanych zbiorach, w zależności od wartości parametru M.

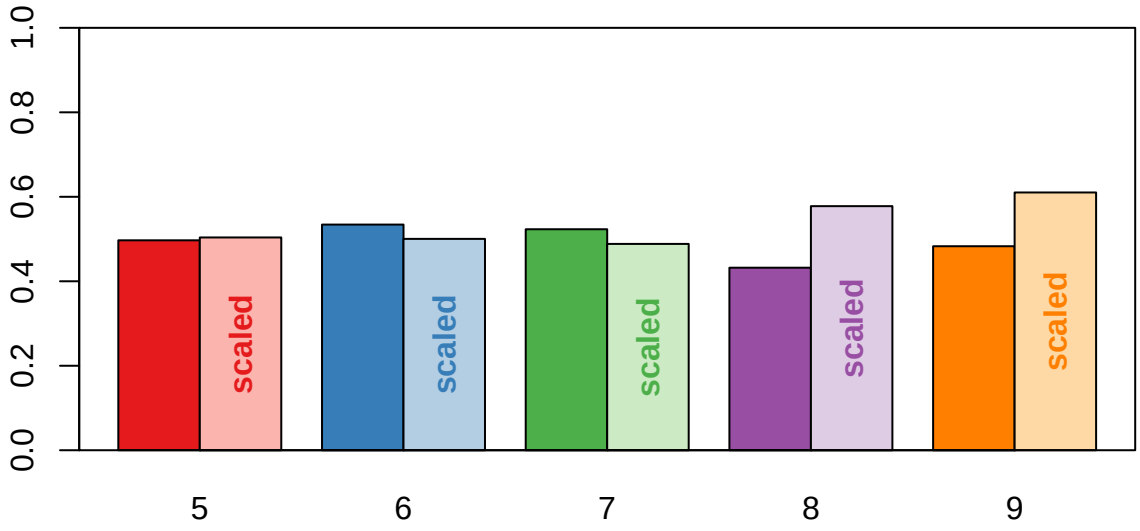

Figure 4: Średnia indeksów Fowlkesa-Mallowsa dla własnej implementacji algorytmu spektralnego na nie standaryzowanych i standaryzowanych zbiorach, w zależności od wartości parametru M.

1.2 Algorytm *Genie* z pakietu `genie`

Zbadałem w jakość algorytmu w zależności od parametru `thresholdGini` (progu dla współczynnika Giniego). Współczynnik może przyjmować wartości rzeczywiste z przedziału `(0, 1]`. Domyślną wartością jest `0.3`.

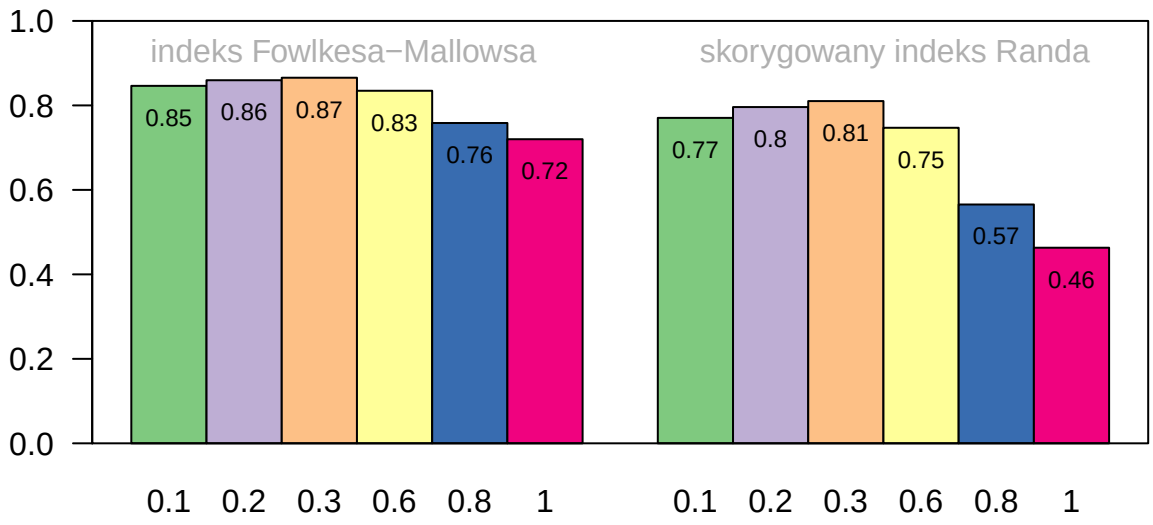

Figure 5: Ocena średnich jakości wyników algorytmu Genie w zależności od parametu thresholdGini.

Najlepsze wyniki algorytm osiąga w okolicach współczynnika równego `0.3` i gorsze im dalej od `0.3`. Oznacza to że jego domyślna wartość jest dobrze dobrana.

Wpływ standaryzacji

Standaryzacja danych wejściowych ma nieistotnie negatywny wpływ na jakość algorytmu.

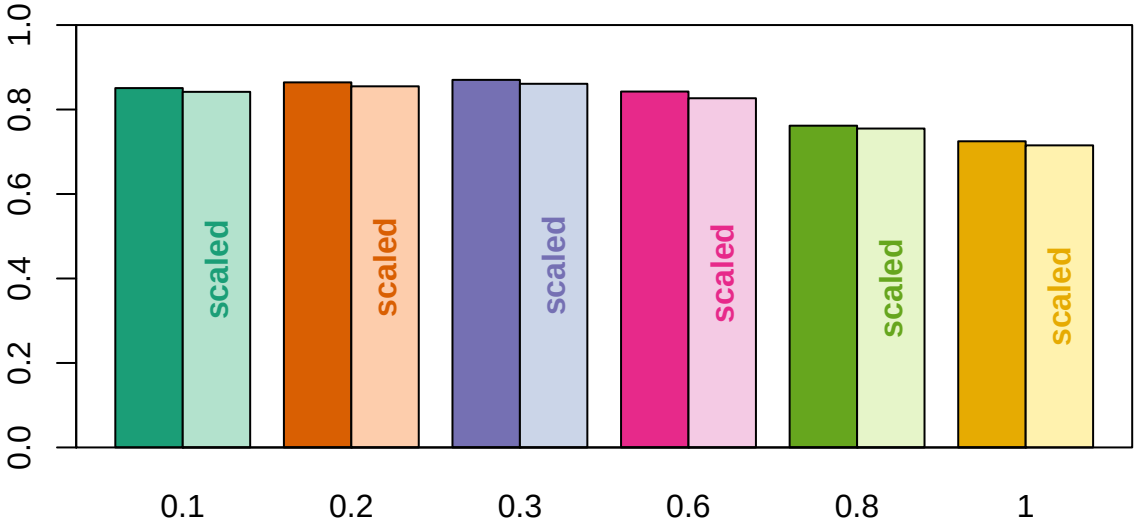

Figure 6: Średnia indeksów Randa dla algorytmu Genie na nie standaryzowanych i standaryzowanych zbiorach, w zależności od wartości parametru thresholdGini.

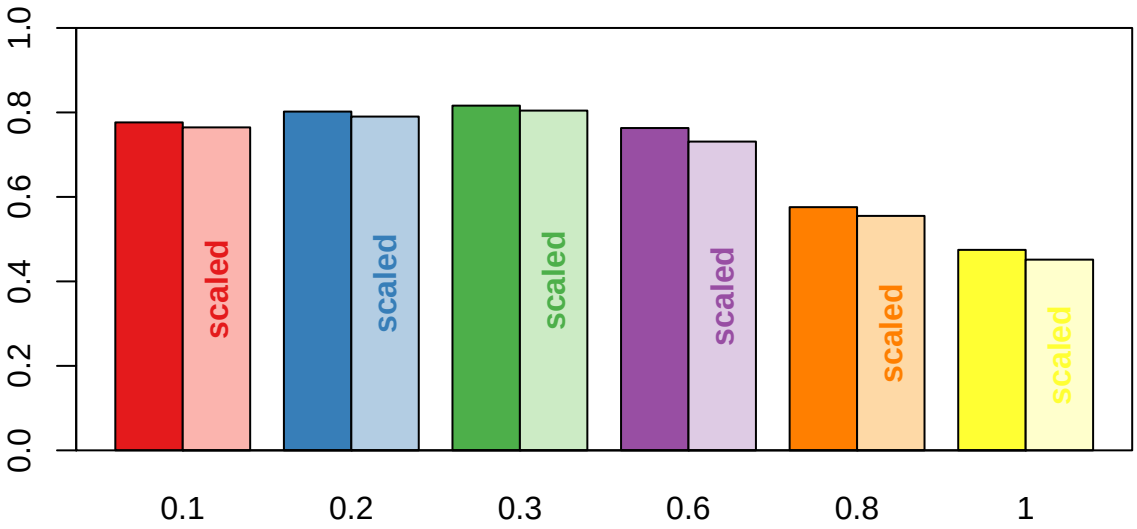

Figure 7: Średnia indeksów Fowlkesa-Mallowsa dla algorytmu Genie na nie standaryzowanych i standaryzowanych zbiorach, w zależności od wartości parametru thresholdGini.

1.3 Algorytmy hierarchiczne z funkcji `hclust()`

Zbadałem jakość algorytmu w zależności od parametru `method`, czyli używanej przez algorytm metody aglomeracji. Zbadałem wszystkie udostępniane przez funkcję metody.

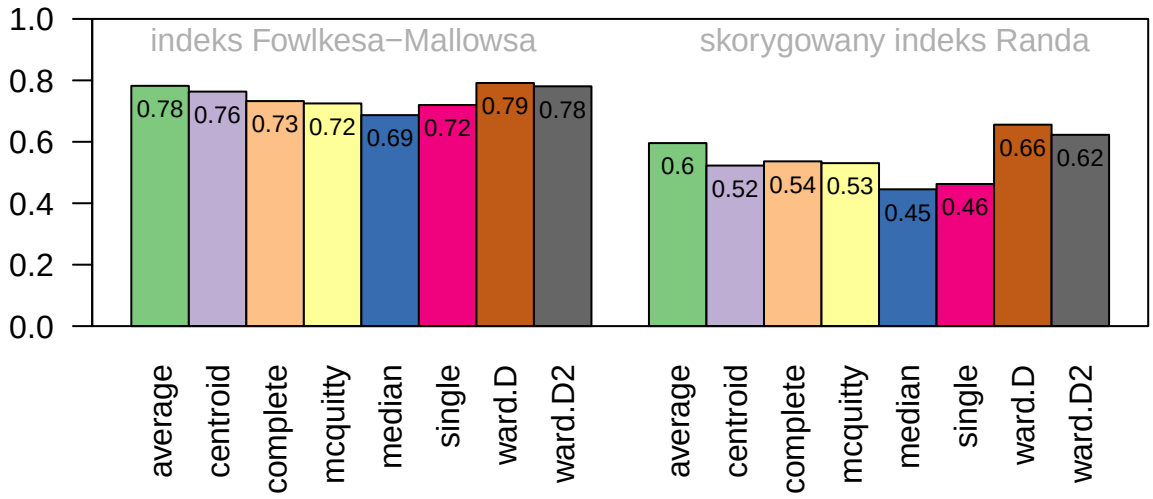

Figure 8: Ocena średnich jakości wyników poszczególnych algorytmów hierarchicznych z funkcji hclust.

Różnice pomiędzy jakościami poszczególnych metod są niewielkie, ale najlepszą metodą okazuje się "ward.D", "ward.D2" jest niewiele słabsza. Najgorszą za to okazała się metoda "median".

Wpływ standaryzacji

Standaryzacja danych wejściowych także w przypadku tego algorytmu, dla wszystkich jego wariantów, ma bardzo mały negatywny wpływ na jego jakość.

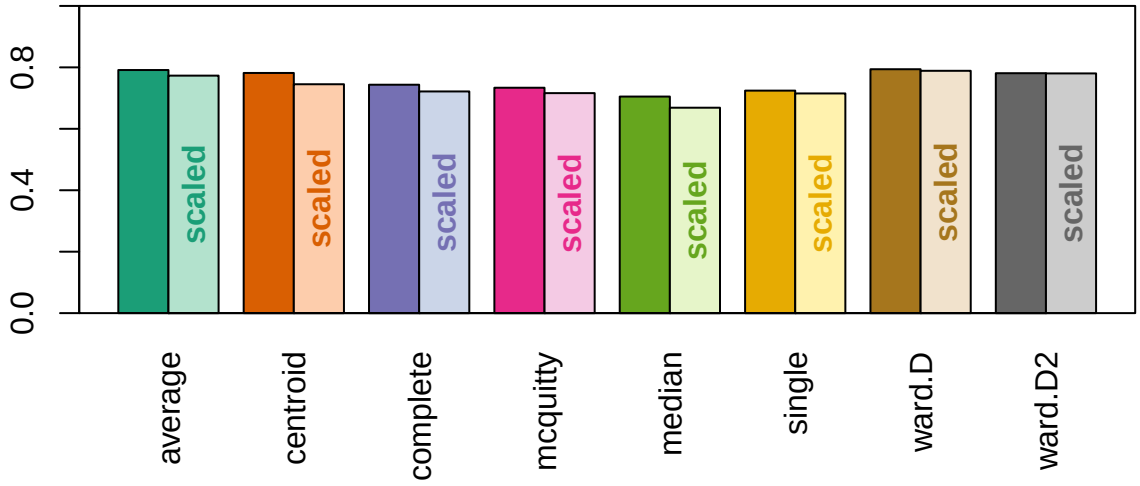

Figure 9: Średnia indeksów Randa dla poszczególnych algorytmów z funkcji hclust na nie standaryzowanych i standaryzowanych zbiorach.

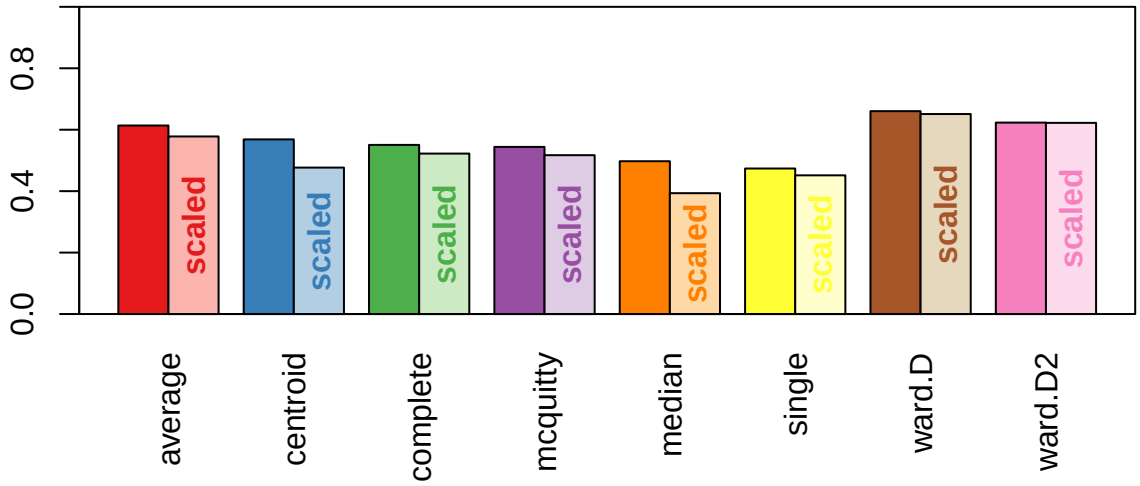

Figure 10: Średnia indeksów Fowlkesa-Mallowsa dla poszczególnych algorytmów z funkcji hclust na nie standaryzowanych i standaryzowanych zbiorach.

1.4 metoda k-średnich `kmeans()`

Zbadałem jakość metody w zależności od parametru `algorithm` udostępniającego 3 warianty.

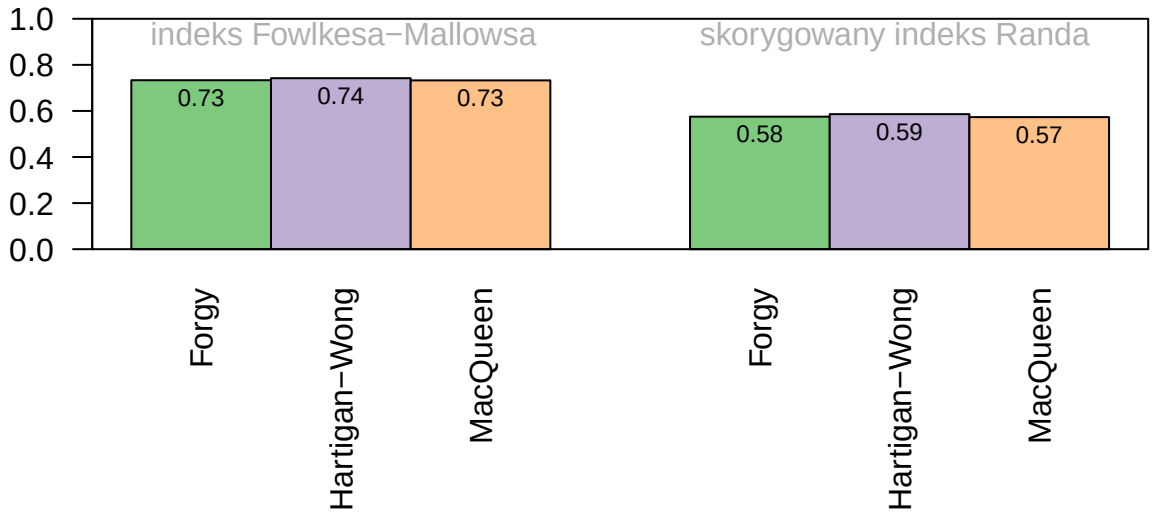

Figure 11: Ocena średnich jakości wyników poszczególnych wariantów algorytmu k-średnich z funkcji kmeans.

Dla wszystkich algorytmów metoda osiąga bardzo bliskie wyniki.

Wpływ standaryzacji

Tutaj także standaryzacja okazuje się mieć niski, ale nieistotny negatywny wpływ na jakość metody dla wszystkich wariantów.

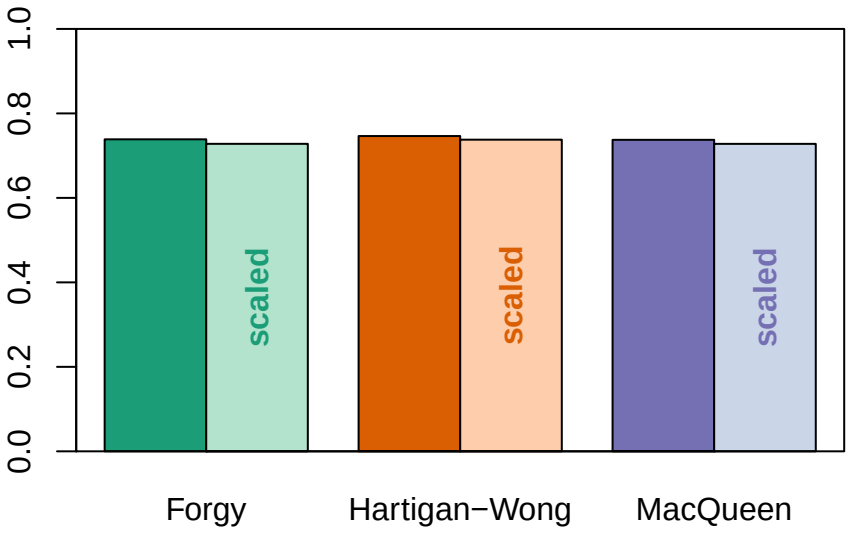

Figure 12: Średnia indeksów Randa dla poszczególnych wariantów funkcji kmeans na nie standaryzowanych i standaryzowanych zbiorach.

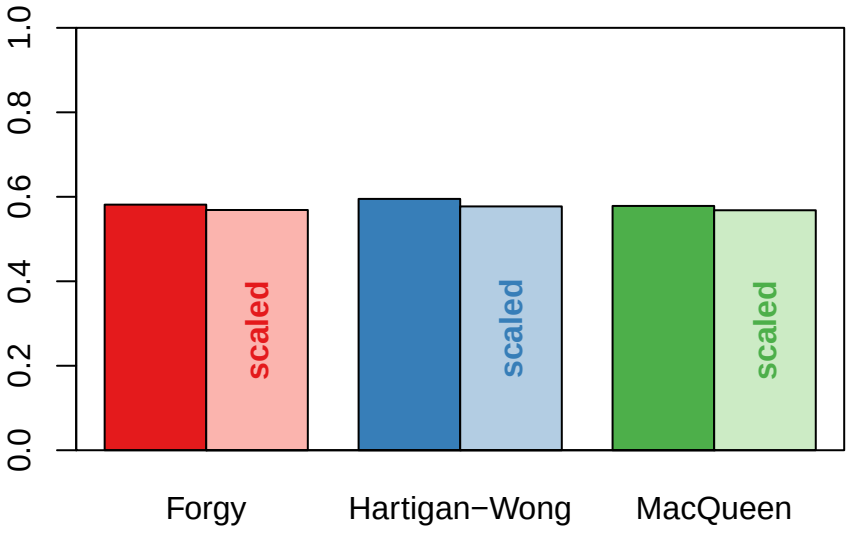

Figure 13: Średnia indeksów Fowlkesa-Mallowsa dla poszczególnych wariantów funkcji kmeans na nie standaryzowanych i standaryzowanych zbiorach.

2. Porównanie algorytmów

2.1. Porównanie uśrednionych wyników

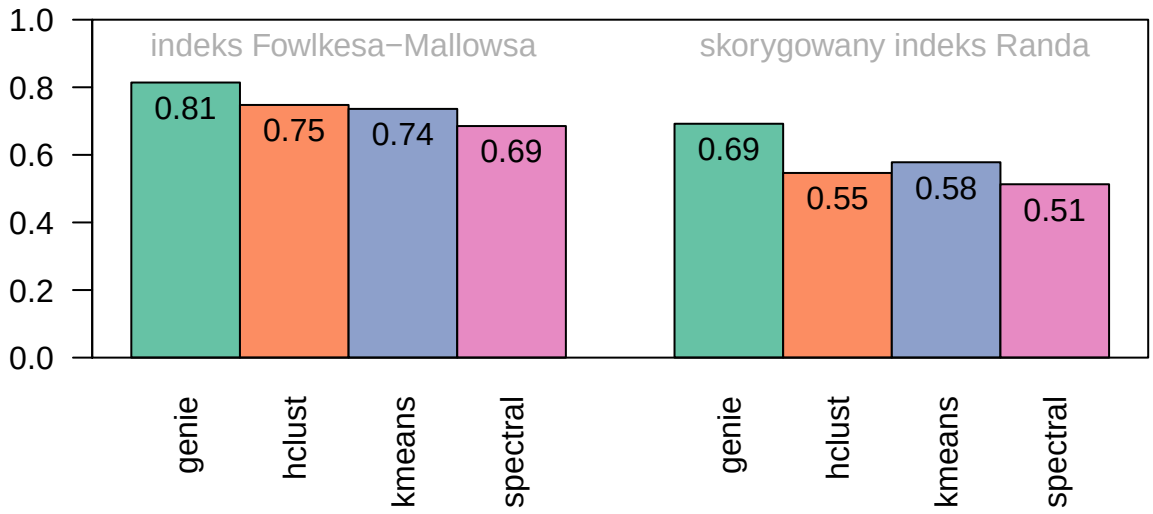

Figure 14: Porównanie średnich jakości wyników poszczególnych badanych metod.

Algorytm *Genie* wyraźnie wyróżnia się pozytywnie na tle pozostałych. Algorytm spektralny okazał się najsłabszy, słabszy nawet od wykorzystywanego przez niego algorytmu k-średnich.

2.2. Porównanie wyników najlepszych wariantów

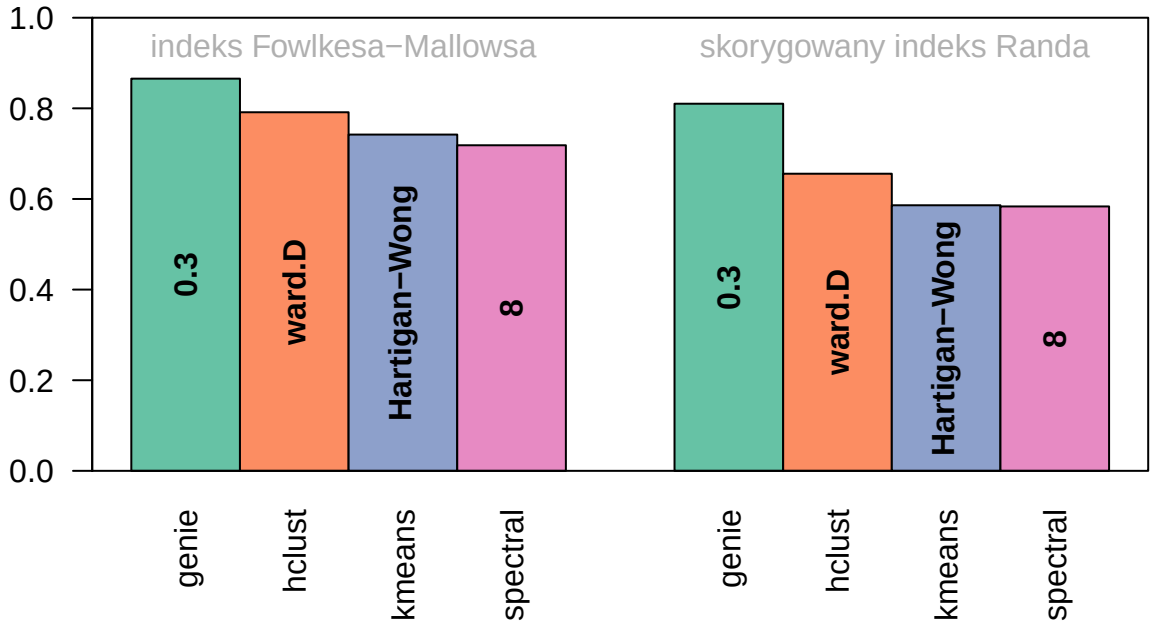

Figure 15: Porównanie średnich jakości wyników najlepszych (o największej średniej z danego indeksu na wszystkich zbiorach) wariantów badanych metod.

Dla każdej z metod ten sam spośród badanych wariantów okazał się najlepszym pod względem zarówno indeksu Fowlkesa-Mallowsa, jak i Randa. Słaby średni wynik algorytmu spektralnego w poprzednim porównaniu raczej nie wynika z doboru badanych parametrów M, ponieważ nawet dla dającego najlepsze wyniki parametru M = 8 nadal `spectral` otrzymuje najniższe indeksy.