Inteligentna Analiza Danych
Zadanie 2

***Damian Rudnicki* 203983**
***Wojciech Różycki* 203982**

1. Cel

Celem zadania jest zaimplementowanie algorytmu k-średnich, oraz sprawdzenie wpływu parametrów metody, takich jak metoda inizjalizacyjna i ilość centrów na uzyskane wyniki pomiarów.

2.Wyniki

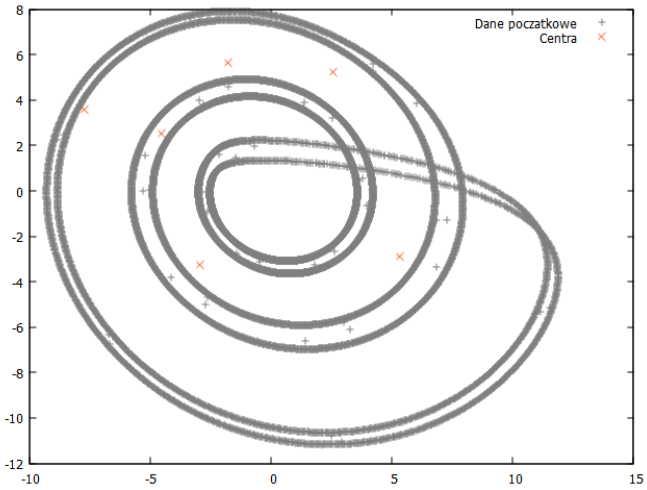
Napisany przez nas program umożliwia ustawienie następujących parametrów:

- ilość centr,
- metoda inicjalizacyjna: Forgy'iego lub Random Partition,
- ilość powtórzeń algorytmu dla danego przypadku.

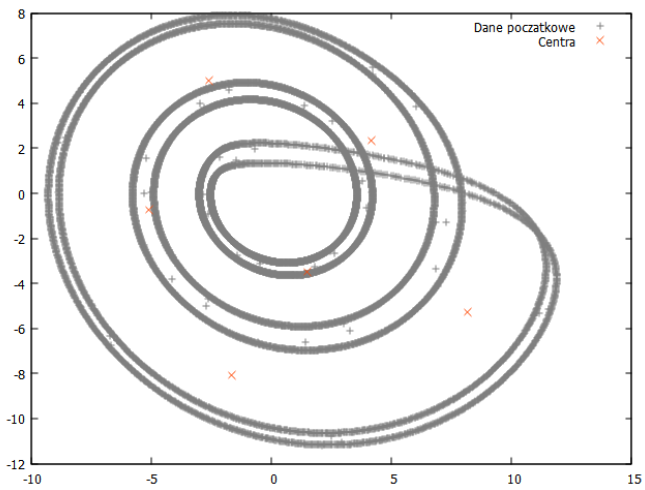
Ilość centr decyduje o ilości grup, na które zostanie podzielony zbiór początkowy. Zbiór początkowy zawarty jest w pliku „Dane testowe". Metoda inicjalizacyjna służy określeniu początkowych grup, od których rozpoczęty zostanie proces nauki. Program umożliwia wybór jednej z dwóch metod inicjalizacji: metoda Forgy'ego lub Random Partition. Metoda Forgy'ego polega na wybraniu odpowiedniej ilości losowych punktów początkowych, które będą stanowiły centra podczas kolejnego pomiaru. W przypadku metody Random Partition wszystkie punkty początkowe przydzielane są do losowych grup. Ilość grup uzależniona jest od ilości centrów. Parametr określający ilość powtórzeń algorytmu służy rejestrowaniu danych, aby wyciągnąć z nich średnią błędu i odchylenie standardowe wyników.

Algorytm przerywa pracę w momencie kiedy żadne z centrów po kolejnej iteracji nie zmieniło swojej poprzedniej pozycji. Przykład działania programu został przedstawiony na rysunku Rysunek 1.

Rysunek 1.
Dane testowe z zaznaczonymi 6 centrami:



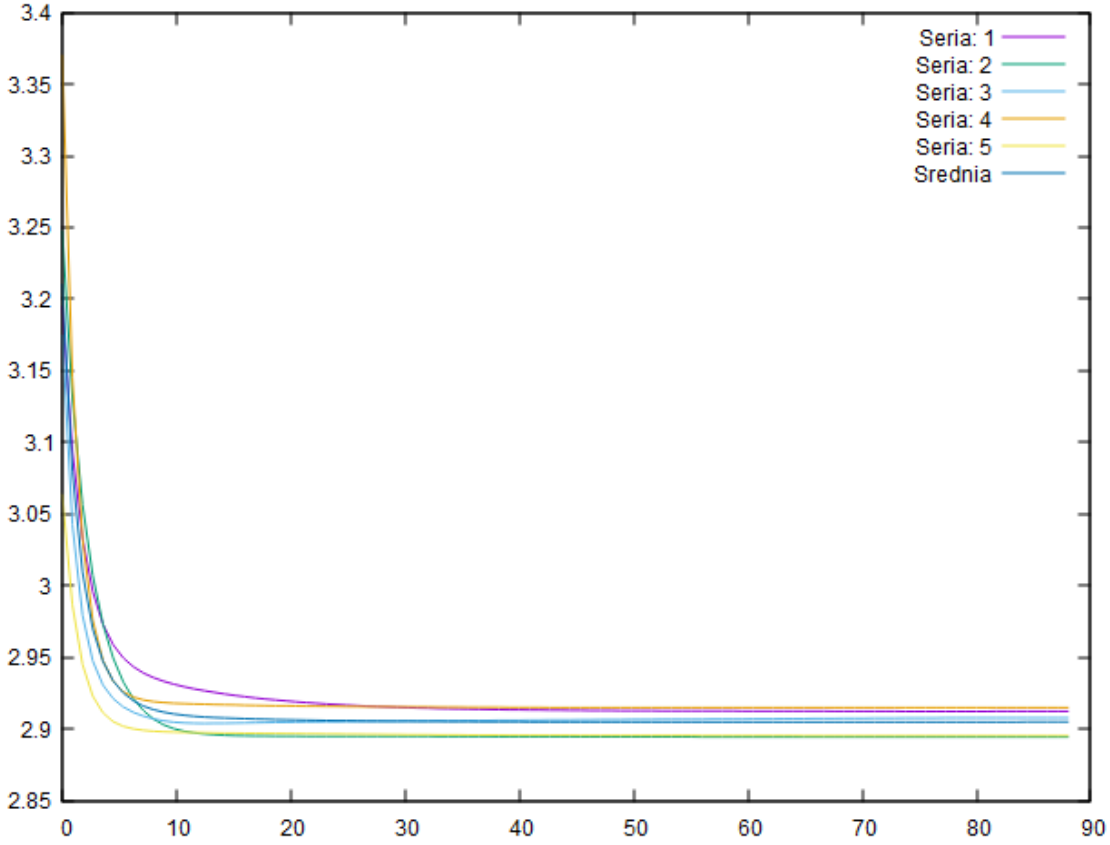
1a) Przed procesem nauki.



1b) Po procesie nauki.

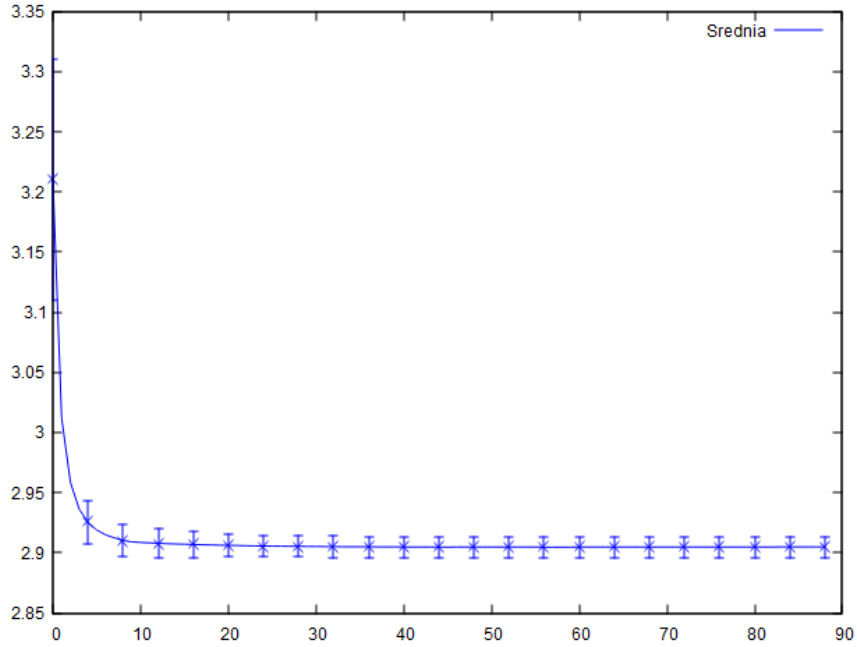
Na rysunkach (Rysunek 1a i 1b) widoczny jest zestaw danych testowych. Przyjmuje on kształt skorupy ślimaka. Poza danymi testowymi przedstawiony został zestaw centrów przed i po procesie nauki. Na podstawie rysunków widoczna jest zmiana położenia centrów po procesie nauki.
Wyniki uzyskanych błędów MSE z przeprowadzonego zadania zostały zestawione w formie wykresów Wykres 1-15. Wykresy (Wykres 1-12) zestawiają różne serie pomiarów przy zmieniającej się liczbie centrów. Uzyskane zależności zostały przedstawione w formie f(x) = y, gdzie x – epoka nauki, f(x) – błąd średniokwadratowy MSE.

Wykres 1.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 6 centr. Zastosowana metoda inicjalizacji: Forgy.



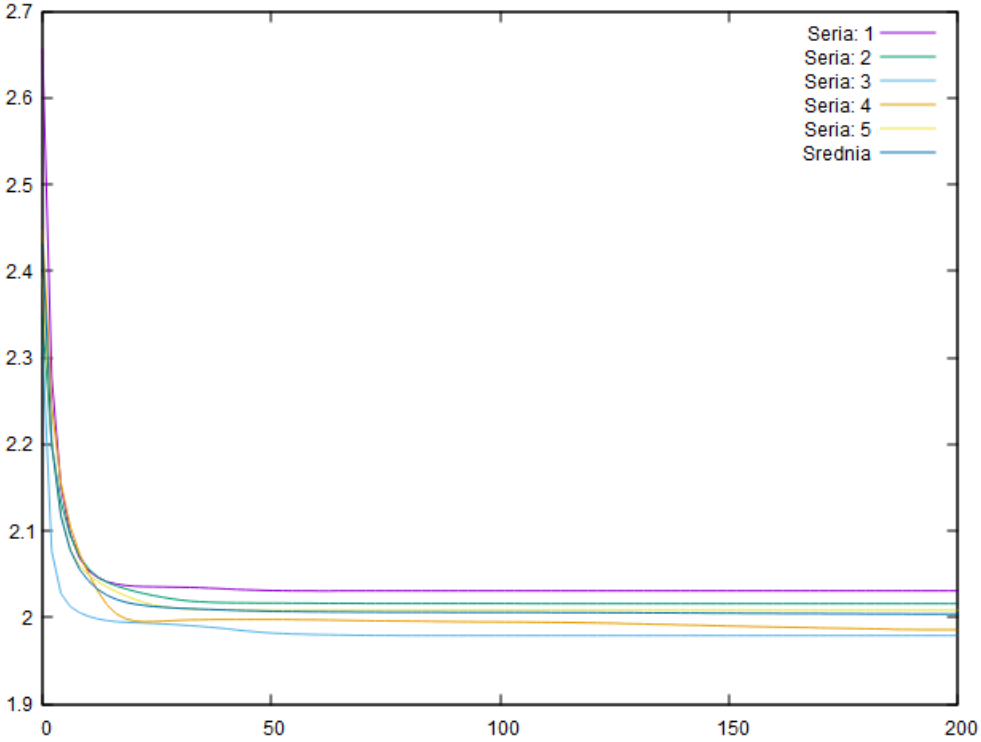
Na wykresie Wykres 1 widoczny jest spadek błędu średniokwadratowego wraz z kolejnymi epokami nauki. Najszybsze ustabilizowanie wartości nastąpiło podczas serii 5, a najwolniejsze podczas serii 1.

Wykres 2.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 6 centr z zaznaczonymi słupkami błędów. Zastosowana metoda inicjalizacji: Forgy.



Na wykresie Wykres 2 widoczny jest spadek wartości zarówno błędu, jak i odchylenia standardowego wraz z kolejnymi krokami nauki.

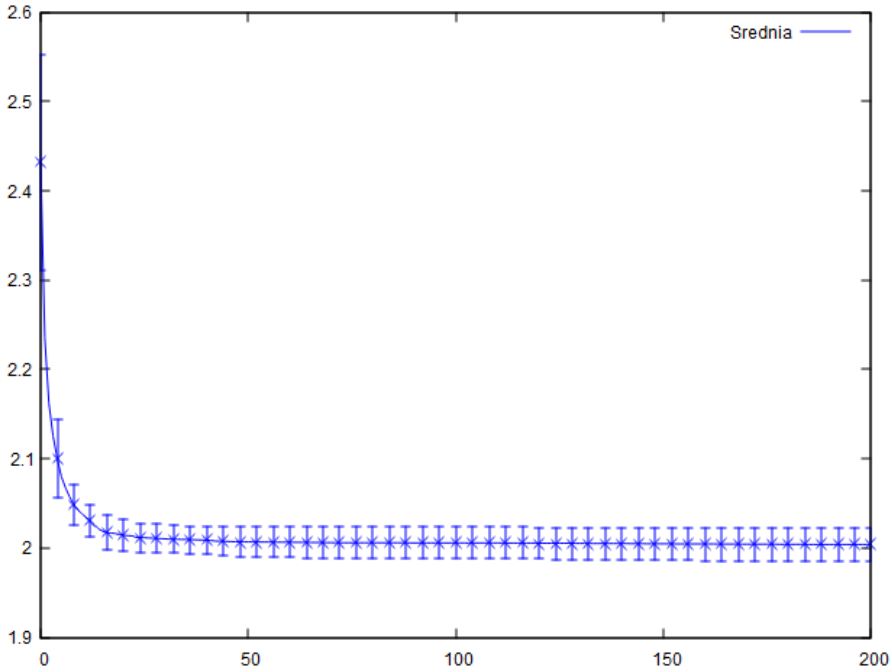
Wykres 3.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 12 centrów. Zastosowana metoda inicjalizacji: Forgy.



Na wykresie Wykres 3. widoczny jest spadek błędu średniokwadratowego wraz z kolejnymi epokami nauki. Tym razem najszybciej ustabilizowane pomiary pochodzą z serii 3.

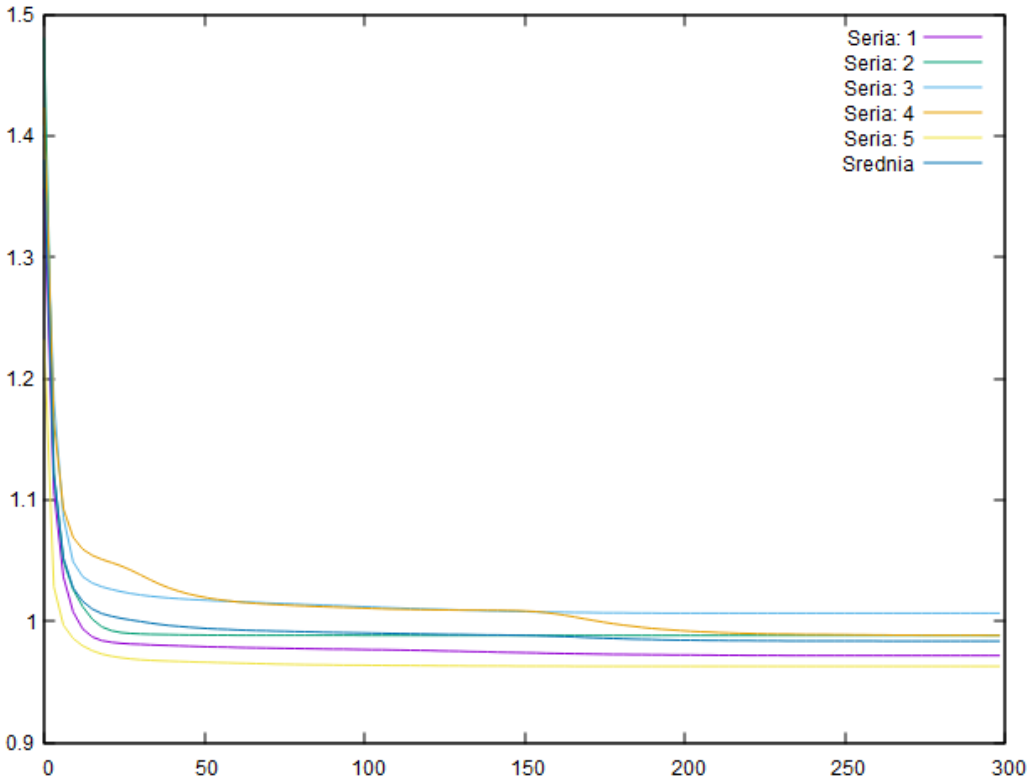
Wykres 4.

Zmiana wartości błędu MSE po kolejnych epokach nauki dla 12 centrów z zaznaczonymi słupkami błędu. Zastosowana metoda inicjalizacji: Forgy.

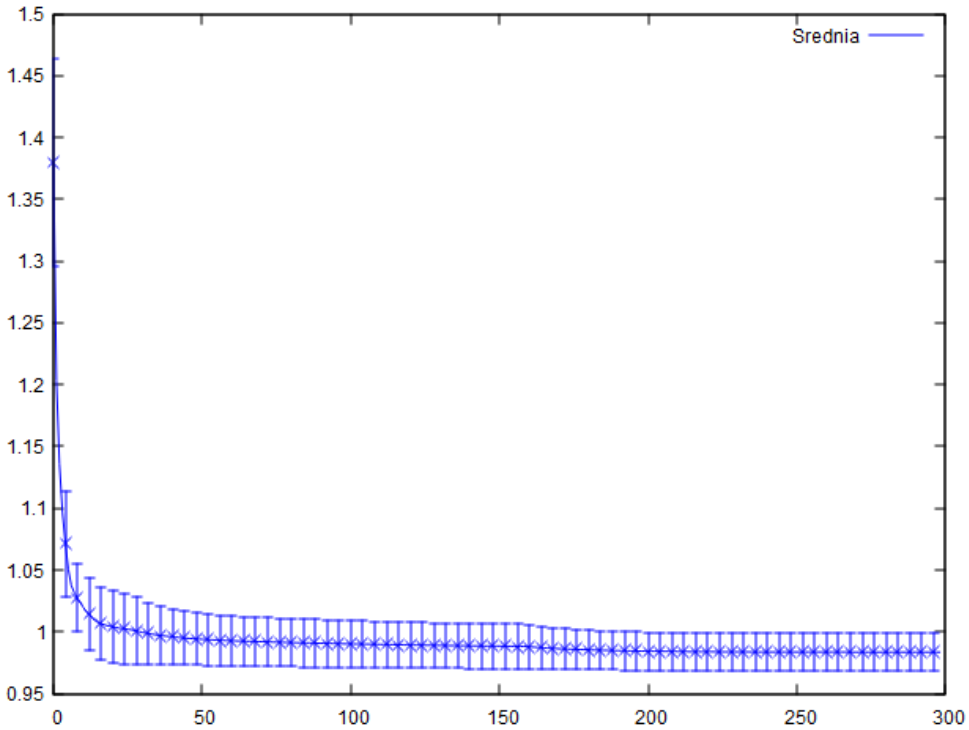


Wykres 5.

Zmiana wartości błędu MSE po kolejnych epokach nauki dla 30 centrów. Zastosowana metoda inicjalizacji: Forgy.



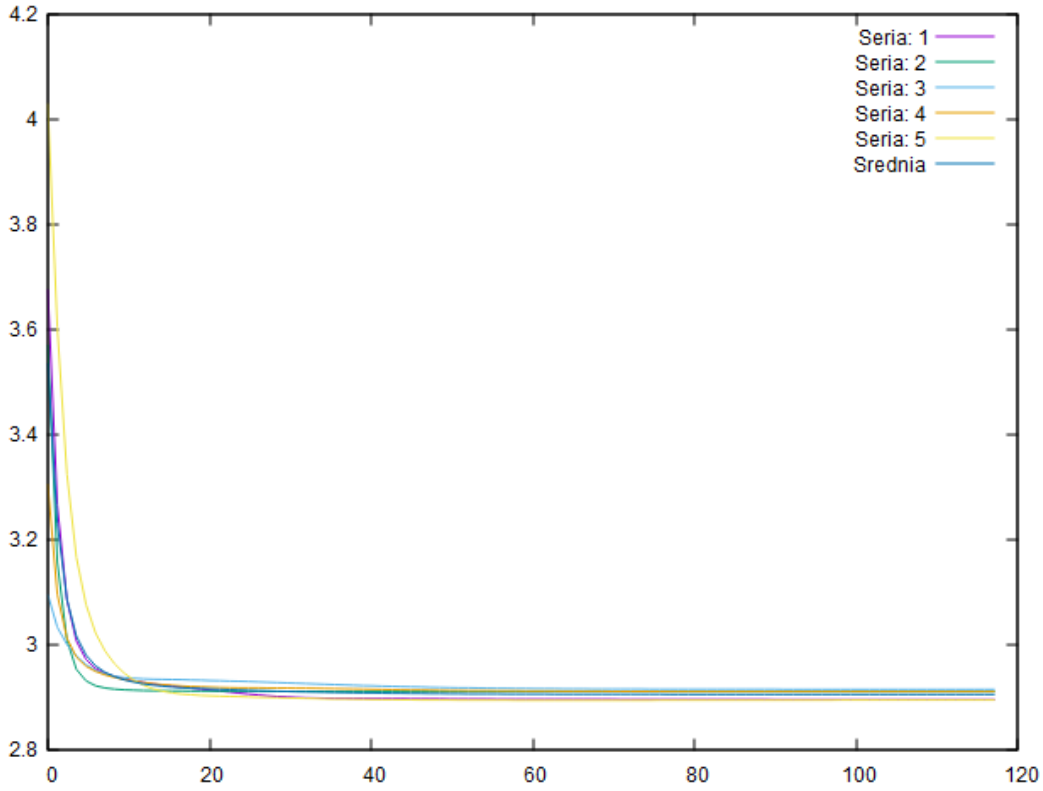
Wykres 6.



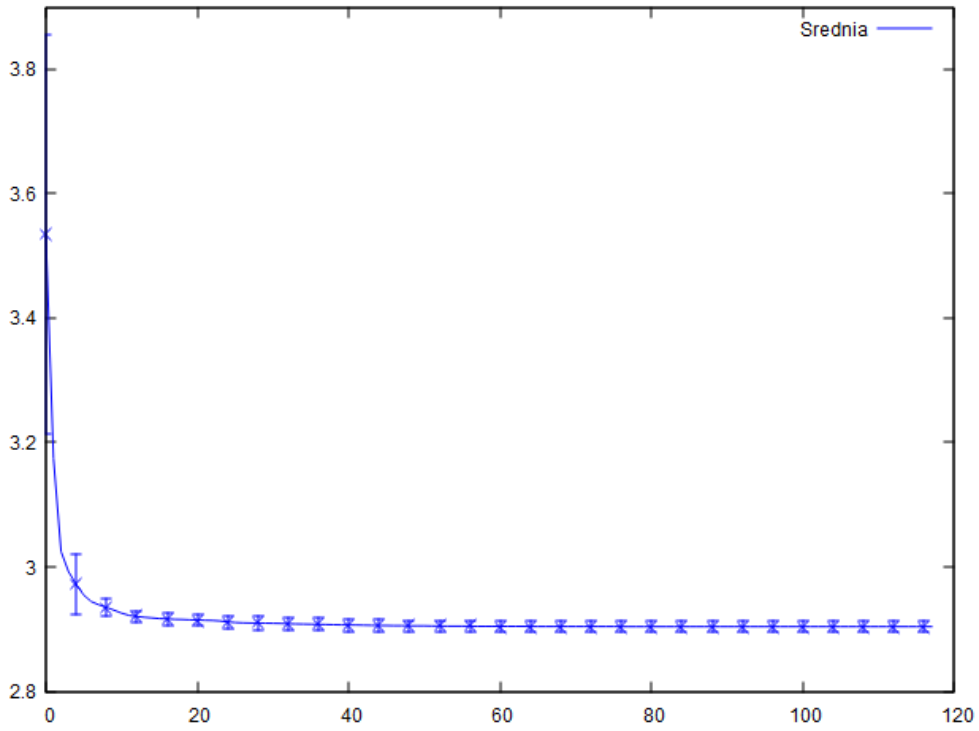
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 30 centrów z zaznaczonymi słupkami błędu. Zastosowana metoda inicjalizacji: Forgy.

Wykres 7.

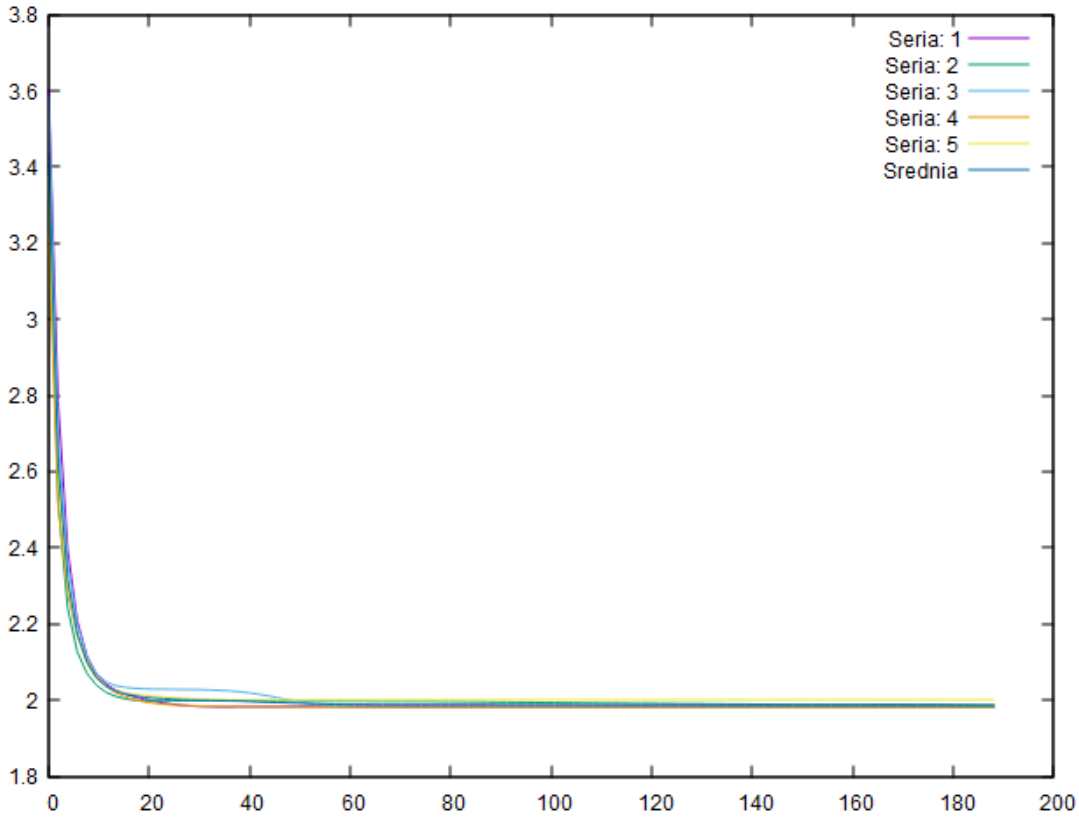
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 6 centr. Zastosowana metoda inicjalizacji: Random Partition.



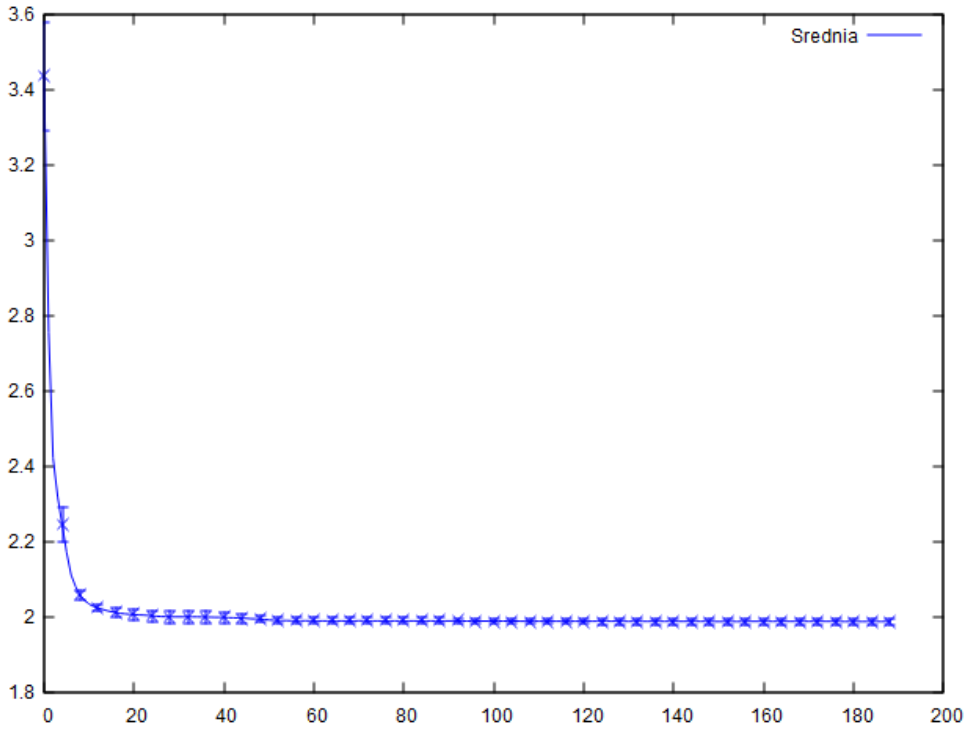
Wykres 8.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 6 centr z zaznaczonymi słupkami błędu. Zastosowana metoda inicjalizacji: Random Partition.



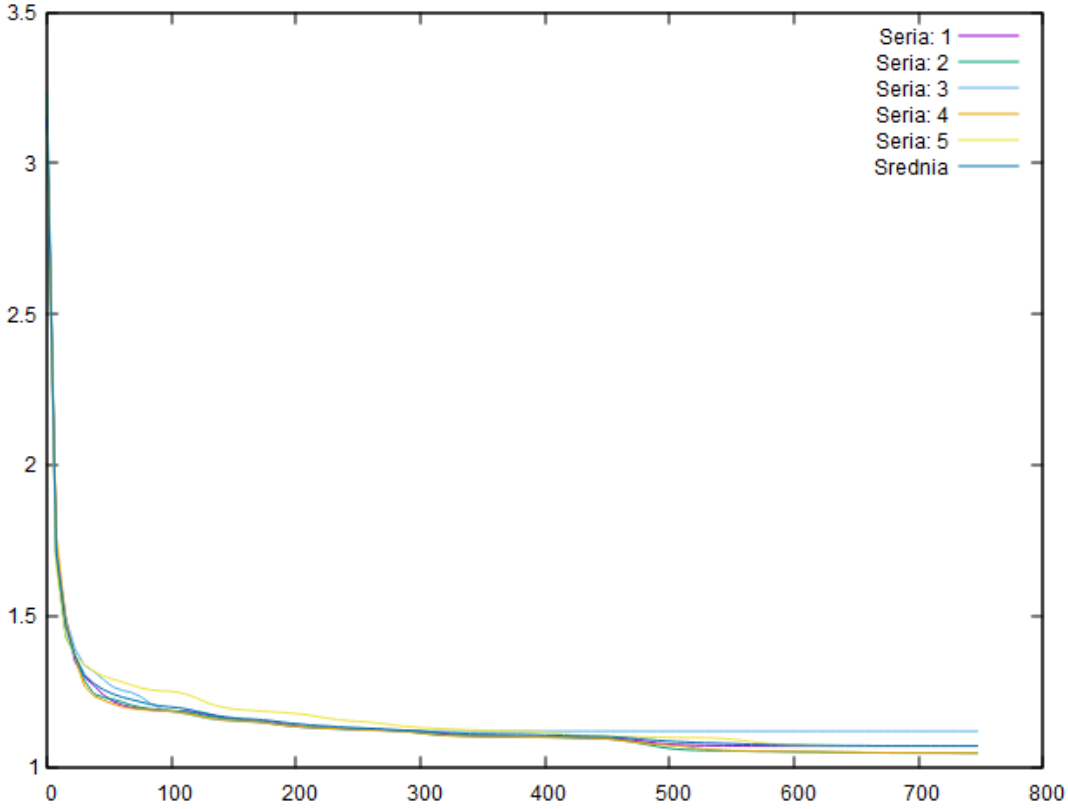
Wykres 9.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 12 centrów. Zastosowana metoda inicjalizacji: Random Partition.



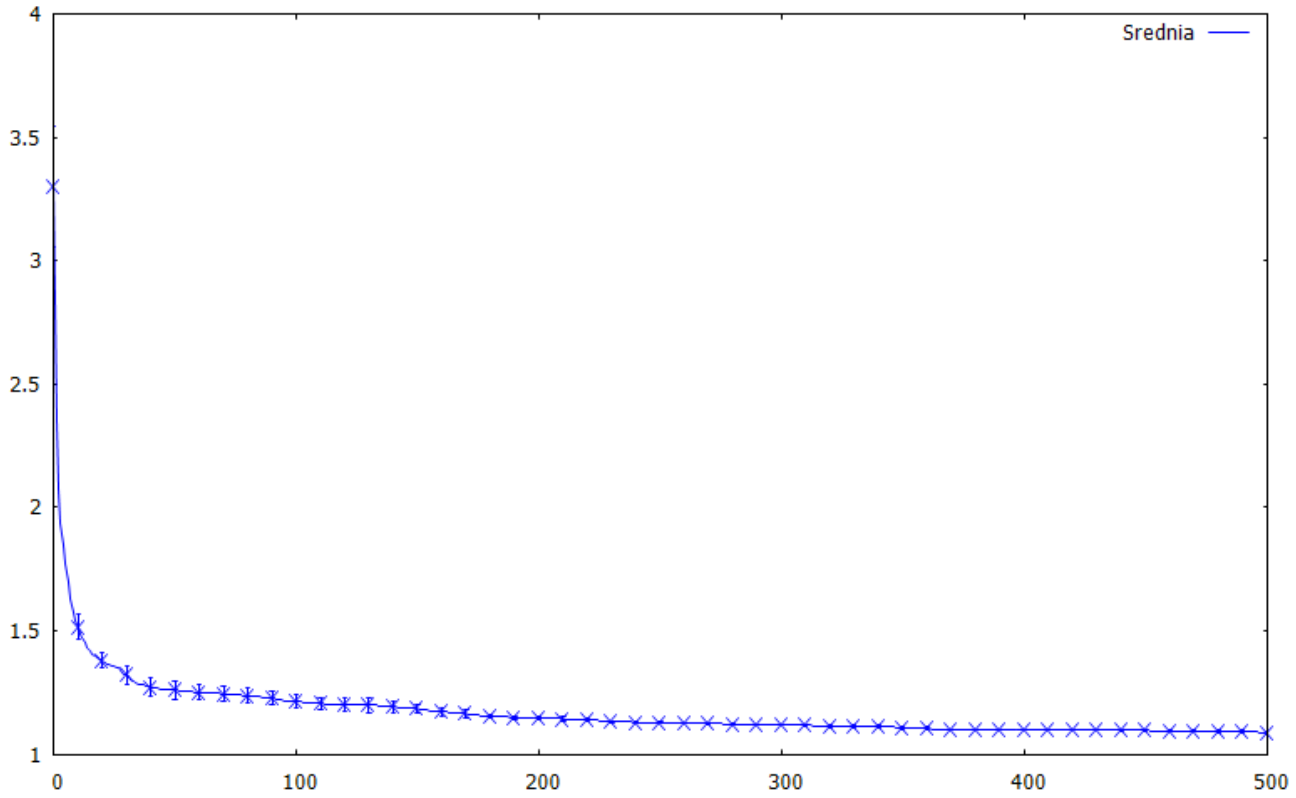
Wykres 10.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 12 centrów z zaznaczonymi słupkami błędu. Zastosowana metoda inicjalizacji: Random Partition.



Wykres 11.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 30 centrów. Zastosowana metoda inicjalizacji: Random Partition.



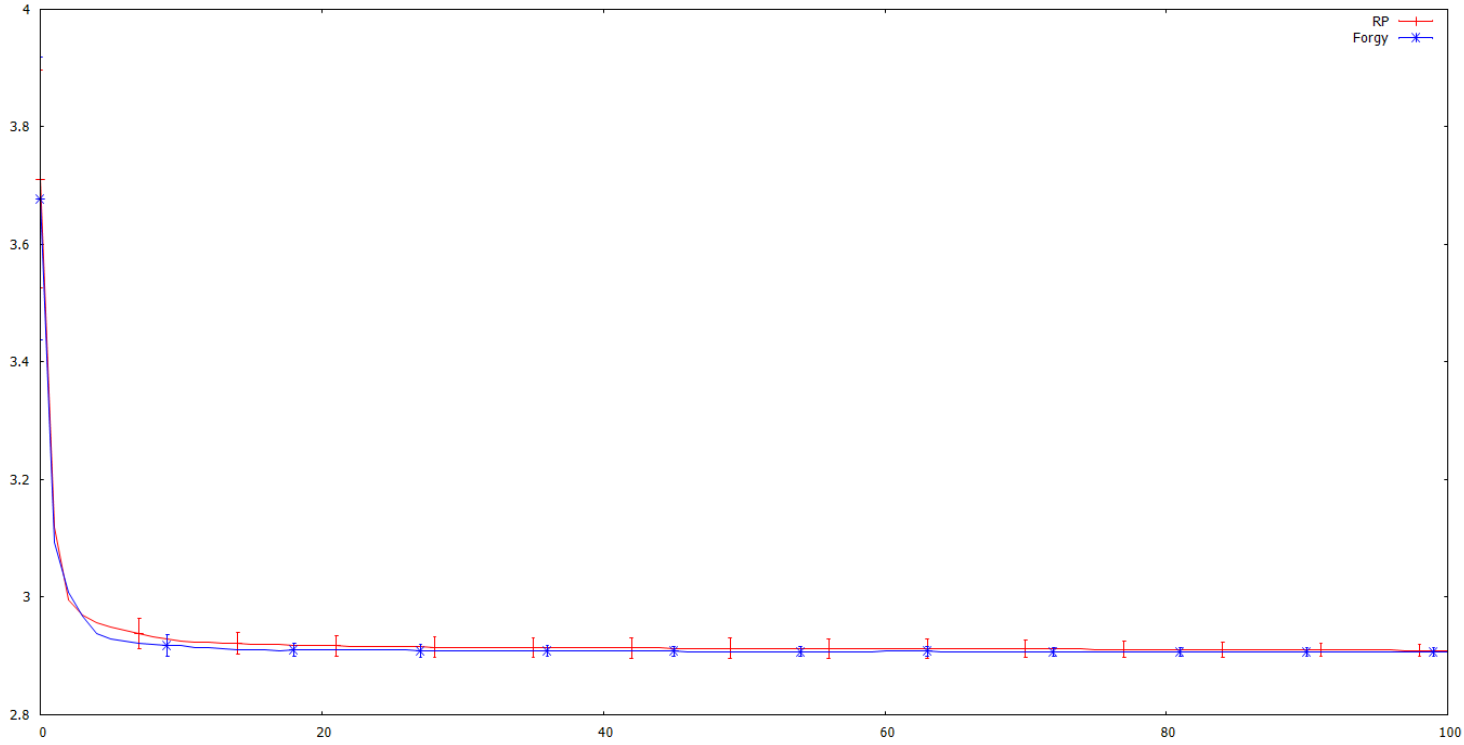
Wykres 12.
Zmiana wartości błędu MSE po kolejnych epokach nauki dla 30 centrów z zaznaczonymi słupkami błędu. Zastosowana metoda inicjalizacji: Random Partition.



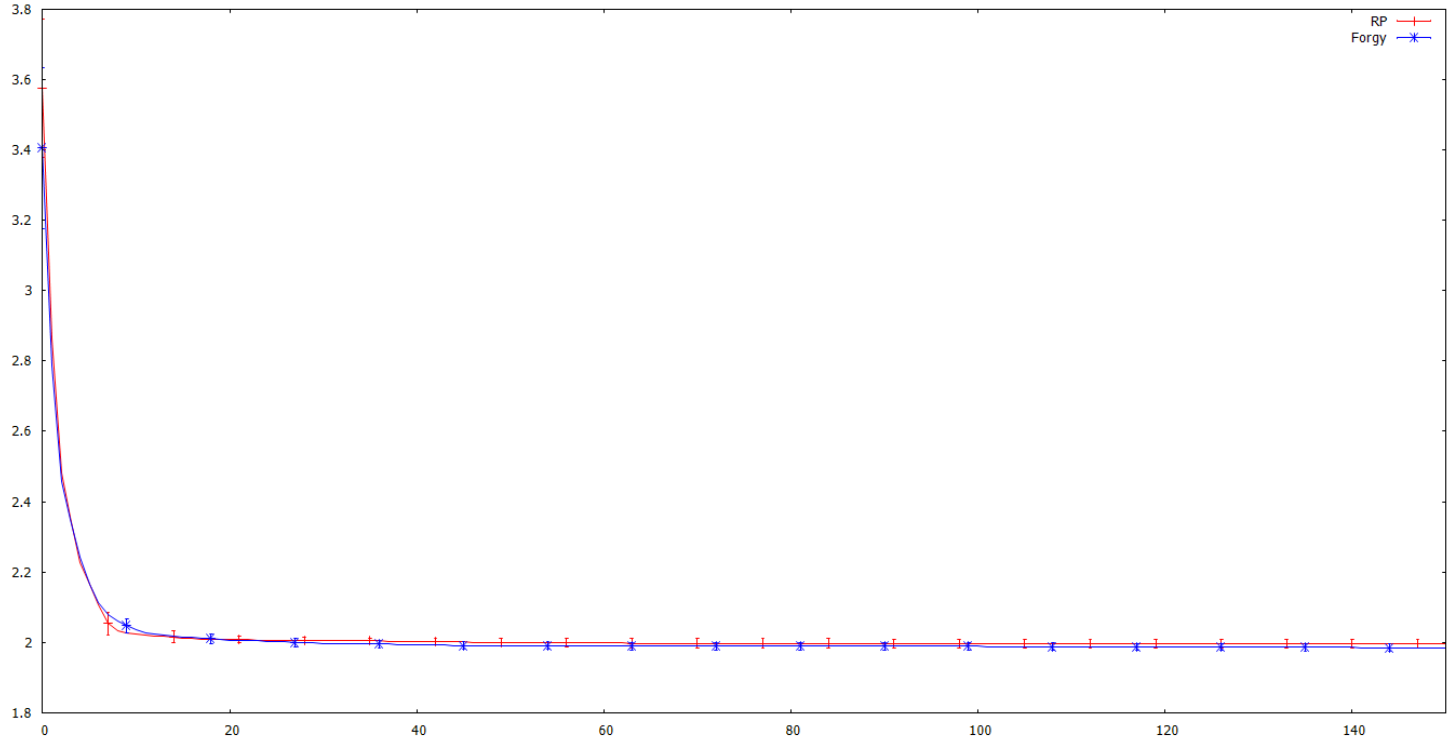
Zauważyć można, że w każdym z przypadków wartości błędów zmniejszają się wraz ze wzrostem epoki nauki. Taka sama tendencja dotyczy odchyleń wartości pomiarów od wartości średnich – maleją wraz z kolejnymi epokami nauki. Wraz ze zwiększającą się liczbą centrów wartość błędu dążyła do niższych wartości, a szybkość nauki wzrastała (wartości ulegały stabilizacji przy mniejszej ilości epok nauki), co widoczne jest na wykresach, np. Wykres 8, 10, 12.

Wykresy (Wykres 13 - 15) przedstawiają zestawione wyniki obu metod inicjalizacji w celu ich porównania.

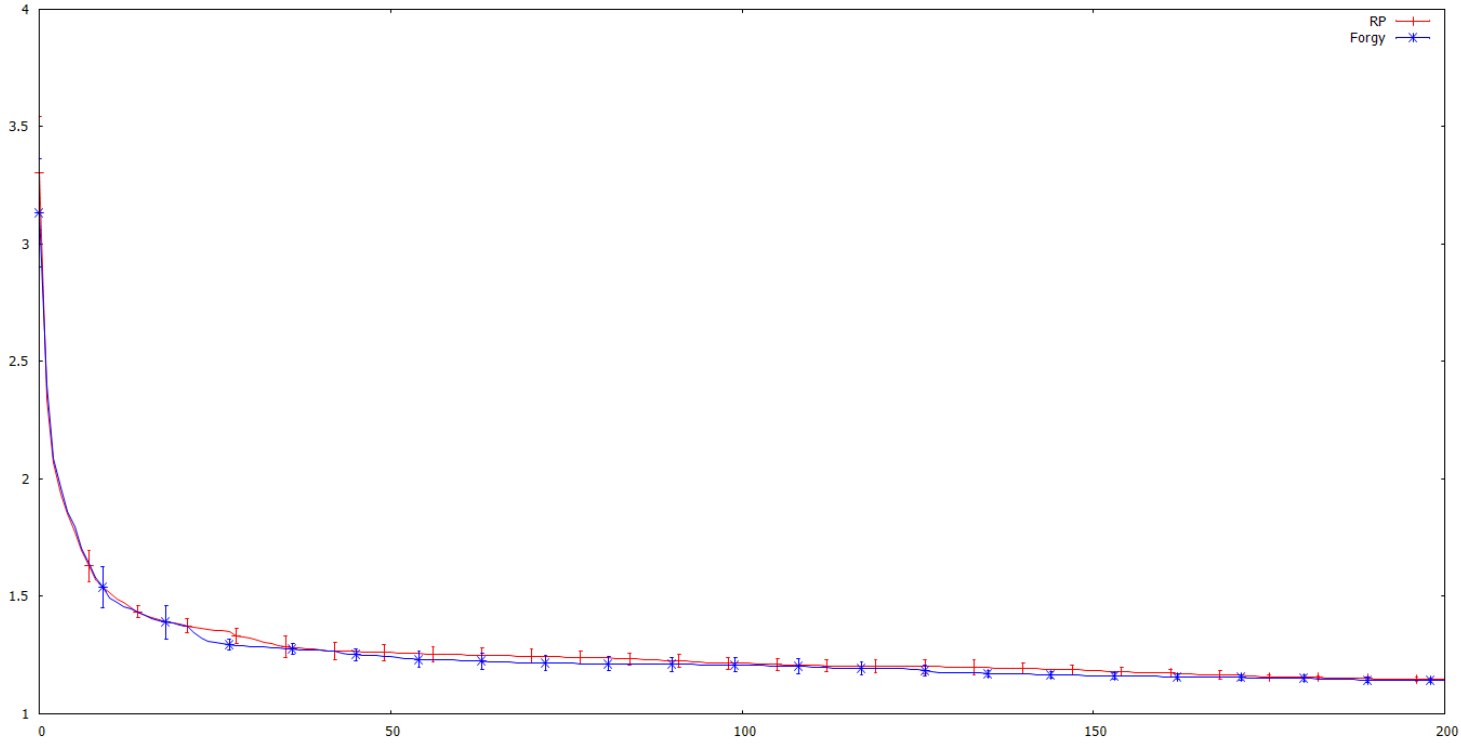
Wykres 13.
Zmiana wartości błędu MSE dla obu metod inicjalizacji po kolejnych epokach nauki dla 6 centrów z zaznaczonymi słupkami błędu.



Wykres 14.
Zmiana wartości błędu MSE dla obu metod inicjalizacji po kolejnych epokach nauki dla 12 centrów z zaznaczonymi słupkami błędu.



Wykres 15.
Zmiana wartości błędu MSE dla obu metod inicjalizacji po kolejnych epokach nauki dla 30 centrów z zaznaczonymi słupkami błędu.



Na wykresach Wykres 12-15 widoczny jest spadek błędu MSE dla obu metod inicjalizacji: Forgy'ego i Random Partition przy zwiększającej się epoce nauki. Obie zależności są porównywalne i mają zbliżone wartości. Momentami zauważane jest odejście krzywych względem siebie, jednakże parę epok później krzywe znowu idą ku sobie. Dla 6 i 30 centrów, dla metody Forgy'ego proces nauki przebiegał nieco szybciej. Dla metody Forgy'ego uzyskane finalne wartości błędu MSE są niższe niż dla metody Random Partition, choć różnica tych błędów jest niewielka (rzędu 0.02)

3.Dyskusja

Na podstawie uzyskanych wyników programu można stwierdzić, że zaimplementowana metoda k-średnich działa. Program w kolejnych krokach nauki przyporządkowuje punkty do innych grup, aż do momentu, w którym żadna z grup nie zyskuje nowego punktu. Ilość centr wpływa na ostateczną wartość uzyskanego błędu MSE. Im więcej centr jest rozpatrywane, tym wartość błędu osiąga niższe wartości. Kolejną zależnością jest wpływ ilości centr na szybkość nauki. Im więcej centr zostało zastosowanych, tym szybkość nauki rosła. Widać również, że wybór metody inicjalizacyjnej nie wpływa znacząco ani na szybkość nauki ani na końcową wartość błędu.