Stosowanie analizy skupień ma na celu określenie czy wartości zmiennych, które zostały zmierzone, tworzą jakąkolwiek strukturę obserwacji w zbiorze danych, oraz czy grupy obserwacji są blisko siebie. Jeśli istnieje zmienna pomocnicza (np. kontynent na przykładzie grup krwi), można sprawdzić, czy naturalne grupowanie danych odpowiada tej zmiennej, ale ważne jest, że głównym zadaniem klastrowania jest wyznaczenie naturalnego podziału, a następnie sprawdzić, czy odpowiada on narzuconym kryteriom. Najważniejszym celem analizy skupień jest możliwość przewidywania na podstawie stworzonego podziału.

Dane można grupować hierarchicznie- metodą najbliższego sąsiada lub niehierarchicznie (np. metoda K-średnich) klasyfikacja najbliższego sąsiada jest prosta i nieco intuicyjna - znajduje obserwacje w danych, które są bliskie, podobne sobie i klasyfikuje je razem. Osoba wprowadzająca klasyfikację musi określić sposób klasyfikowania sąsiadów (np. grupa krwi) oraz liczbę sąsiedzkich klas „k” (np. od 2 do 10). Przed wyborem liczby klas powinna zostać stworzona macierz odległości pomiędzy obserwacjami. Należy znaleźć ***k*** najmniejszych wartości w każdym rzędzie macierzy, po ich znalezieniu należy określić, do których obserwacji należą, i stwierdzić kryterium, po którym zostały sklasyfikowane.

Optymalną liczbą grup (skupień) można znaleźć metodą dwustopniowej analizy skupień. Miara odległości oparta na wiarygodności zakłada, że zmienne w modelu skupień są niezależne (testowanie: korelacje parami dla zmiennych ilościowych, tabele krzyżowe dla zmiennych jakościowych).

Analizy>Analizy klasyfikacyjne> dwustopniowa analiza skupień

Analizy>Analizy klasyfikacyjne> grupowanie met. K-średnich