

Marzena Nowakowska, Wydział Zarządzania i Modelowania Komputerowego, Politechnika Świętokrzyska

Podobieństwa i różnice między krajami Europy w zakresie zielonej transformacji z wykorzystaniem aglomeracyjnej analizy skupień

Należy pamiętać, aby przy ponownym uruchomieniu projektu EG uruchomić węzeł przypisania biblioteki.

Zbiór danych zawiera zmienne ilościowe niosące informację o tym, na jakim poziomie jest zielona transformacja w wybranych krajach Europy. Sprawdzić zawartość pliku z danymi *WDI_ZlnTrns.xlsx*. Zaznajomić się z opisem zmiennych w arkuszu *Series - Metadata*.

Raport zawiera nie więcej 5 stron.

Do każdego umieszczonego w raporcie wykresu należy się w treści odnieść.

Przygotowanie środowiska pracy

Utworzyć plik Excela z danych oryginalnych zawartych w pliku *WDI_ZlnTrns.xlsx*. Nadać plikowi nazwę *WDI_ZlnTrns_<nr>.xlsx*. Plik docelowy zawiera dane do analizy skupień będące średniki dwóch wybranych lat ustalonych jak następuje:

- 1) 2018+2019,
- 2) 2018+2020,
- 3) 2018+2021,
- 4) 2019+2020,
- 5) 2019+2021,
- 6) 2020+2021.

Zdefiniować w EG diagram procesu analizy skupień dla tych danych wykorzystując wybraną metodę aglomeracyjnej analizy skupień. Nadać plikowi projektu nazwę *WDI_ZlnTrns_<nr>.egp*.

W oknie procesu skojarzyć bibliotekę z folderem plików SAS; menu *Narzędzia / Przypisz bibliotekę projektu...*

Wykonać operację importu swoich danych Excel-owych do środowiska projektu; menu *Plik / Importuj dane*.

Preprocessing danych

Przekształcić wszystkie wskaźniki procentowe (zmienne określające procent, które są identyfikowane literą *P* na końcu nazwy), tak aby ich wartości zostały wyskalowane do przedziału [0, 1] (przez jaką liczbę należy podzielić wartości tych zmiennych?). Znormalizować zmienne niebędące wskaźnikami procentowymi (są trzy takie zmienne – nie mają litery *P* na końcu nazwy).

Wykonanie

Wprowadzić do diagramu węzeł *Budowa zapytań* z opcji *Zadania / Dane*. Wybrać zakładkę *Kolumny wyliczane*, a następnie kliknąć przycisk *Nowa*.

- Skalowanie. W kolejnych oknach kreatora nowej kolumny określić: 1) *Typ – Wyrażenie zaawansowane*, 2) w edytorze wyrażeń wyskalować zmienną *LasyP*, 3) nadać nowobudowanej zmiennej nazwę *LasyP_Sk*, 4) zakończyć tworzenie zmiennej i wrócić do okna *Kolumny wyliczane*. Powtarzać operacje skalowania dla pozostałych zmiennych procentowych.
- Normalizacja. Utworzyć nowe zmienne wg zależności: $x_{nowe} = (x - \min(x)) / (\max(x) - \min(x))$. W celu wprowadzenia agregatów skorzystać z funkcji wbudowanych kreatora.
- Dodać do zestawu zmiennych *Kraj* (przeciągnąć do zakładki *Wybierz dane*).

Analiza korelacji

Wykonać analizą korelacji zmiennych kandydujących do analizy skupień. Zdiagnozować największe korelacje i uwzględnić je w budowie modelu analizy skupień.

Wykonanie

- Wprowadzić do diagramu węzeł *Korelacje* z opcji *Zadania / Wielowymiarowe*.
- Jako zmienne analizowane wprowadzić wszystkie zmienne otrzymane po przekształceniu (pozycja *Dane*).
- Wybrać typ korelacji liniowej Pearsona (*Opcje*).

Wstępna analiza skupień

Wykonać wstępną analizę skupień uwzględniając wszystkie zmienne i wszystkie obserwacje. Sprawdzić wynik analizując korelacje, dendrogram oraz wykresy diagnostyczne. Na tej podstawie określić ostateczny zbiór do analiz poprzez wykluczenie punktów odstających oraz skorelowanych zmiennych.

Wykonanie

- Wprowadzić do diagramu węzeł *Analiza skupień* z opcji *Zadania / Wielowymiarowe*.
- Ustawienia węzła
 - *Dane*. Wprowadzić wszystkie zmienne otrzymane po przekształceniu do analizy oraz zmienną *Kraj* jako etykietę identyfikacyjną.
 - *Skupienie*. Wybrać metodę analizy skupień.
 - *Wykresy*. Ustalić wykresy i orientację dendrogramu.
 - *Rezultaty*. Zaznaczyć opcję *Wynik drzewiasty*. Dzięki temu będzie możliwa identyfikacja przynależności poszczególnych krajów do klastrów.

Wyniki

- *Raport SAS-owy* Tabela z historią skupień wraz z wartościami statystyk, wykresy liniowe wybranych miar oraz wykres dendrogramu.
- *Rozłącz dane skupień* Zbiór danych zawierający wykaz skupień (w liczbie podanej w pozycji *Rezultaty* okna *Analiza skupień*) zawierający nazwę i numer klastra oraz nazwę kraju (etykietującego rekordy), który należy do klastra. Po otwarciu węzła zbiór można uporządkować wg numeru klastra następująco:
 - wybrać opcję menu poziomego *Filtruj i sortuj*
 - wprowadzić do panelu z prawej strony wszystkie zmienne
 - w zakładce *Sortuj* wybrać *Sortuj wg* → *CLUSTER*.
- *Dane dendrogramu* Zbiór danych zawierający wyniki analizy skupień (historię skupień). Dane z tego zbioru można wykorzystać do utworzenia wykresów punktowych wybranych miar w zależności od liczby skupień. Wcześniej można uporządkować zbiór wg liczby skupień (*_NLC_*). Wykres punktowy dla zilustrowania miary R-kwadrat tworzy się wybierając menu *Wykres* → *Wykres punktowy* oraz kolejno:
 - *Dane*: *_NLC_* dla osi poziomej, *_RSQ_* dla osi pionowej
 - Pozostałe pozycje umożliwiają ustawienie wyglądu wykresu.

Alternatywna analiza skupień

Korzystając z wyników poprzednich punktów odrzucić rekordy odstające (jeżeli są) oraz nadmiarowe zmienne (stosunkowo mocno skorelowane z innymi zmiennymi). Wykonać dla tak zmodyfikowanego zbioru analizę skupień. Wyznaczyć liczbę klastrów i przeprowadzić ocenę jakości ilustrując ją odpowiednimi wykresami.

Porównać wyniki z dwóch przeprowadzonych analiz i jako docelowe klastry wybrać jedną z nich

Uwaga 1. Jeżeli potrzeba, można niektóre wyniki wyeksportować do Excela i w nim zrealizować uzupełnienia do analiz (wykresy dodatkowe obliczenia).

Uwaga 2. Raport SAS-owy można zapisać w formacie pdf: Menu podręczne i *Eksportuj* → *Eksportuj Raport SAS-owy – Analiza skupień* i zapisać plik wybierając typ zapisu pdf.