

Hierarchiczne metody analizy skupień

Marzena Nowakowska

Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Istota metod hierarchicznych

Tworzą hierarchię skupień w postaci drzewa, zwanego dendrogramem. Nie wymagają podania liczby klastrow z góry; pozwalają na analizę struktury na różnych poziomach szczegółowości.

Zalety

- Brak potrzeby określania liczby klastrow z góry
- Wizualizacja wyników w postaci dendrogramu
- Dobrze działa dla małych zbiorów danych

Wady

- Wysoka złożoność obliczeniowa (1 000 obiektów → ok. $1000^2 = 1000000$ operacji)
- Raz scalone obiekty nie ulegają przemieszczeniu (aktualizacji)
- Wrażliwość na szum i obserwacje odstające
- Trudność interpretacji przy dużych zbiorach danych

Metody hierarchiczne podziałowe (divisive, top-down)

Algorytm

1. Cały zbiór danych stanowi jedno skupienie.
2. Następnie wyodrębniane są dwa skupienia najbardziej niepodobne do siebie (najbardziej heterogeniczne).
3. Proces podziału jest powtarzany dla wynikowych klastrów. W każdej iteracji wybierany jest najbardziej heterogeniczny klaster i dzielony ponownie.
4. Proces jest kontynuowany do momentu osiągnięcia żądanej liczby klastrów lub spełnienia innego warunku zatrzymania, w szczególności aż każdy punkt danych będzie stanowił oddzielne skupienie.

Przykładowe metody podziałowe

- DIANA (Divisive Analysis)
- Bisecting k-means
- DIVCLUS-T
- Spectral clustering

Metody hierarchiczne aglomeracyjne (agglomerative, bottom-up)

Algorytm

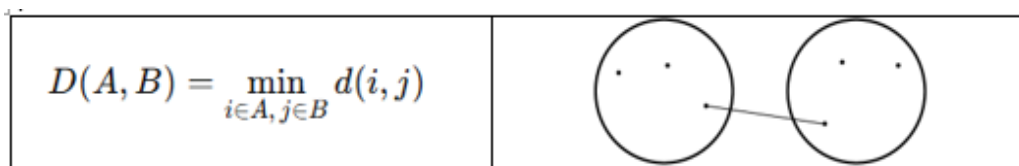
1. Każda obserwacja stanowi jedno skupienie; na początku liczba skupień jest N – taka jak liczba obserwacji.
2. Wykorzystując miarę podobieństwa (odległości) łączy się dwa klastry najbliższe sobie, uzyskując liczbę klastrów równą $N-1$.
3. Procedura łączenia jest powtarzana do chwili, gdy wszystkie obserwacje stanowią jedno skupienie lub gdy spełniony jest inny warunek zatrzymania (np. docelowa liczba skupień lub założone odległości wewnątrz- i między-skupieniowe).

Najpopularniejsze metody aglomeracyjne

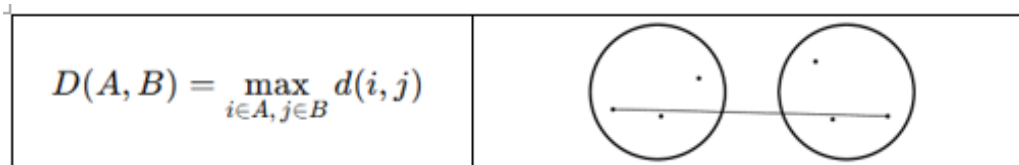
- Najbliższego sąsiada
- Najdalszego sąsiada
- Średnich odległości
- Centroidów
- Warda

Koncepcje metod aglomeracyjnych

Metoda najbliższego sąsiada (single linkage, pojedynczego wiązania); minimalizuje odległość między najbliższymi punktami w łączonych klastrach



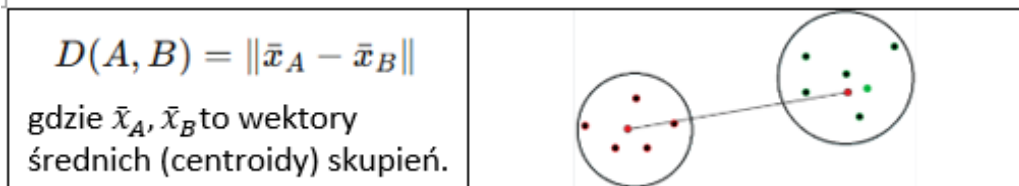
Metoda najdalszego sąsiada (complete linkage, pełnego wiązania); minimalizuje odległość między najdalszymi punktami w łączonych klastrach



Metoda średnich odległości (average linkage, średniego wiązania); minimalizuje średnią odległość między wszystkimi parami punktów dwóch klastrów



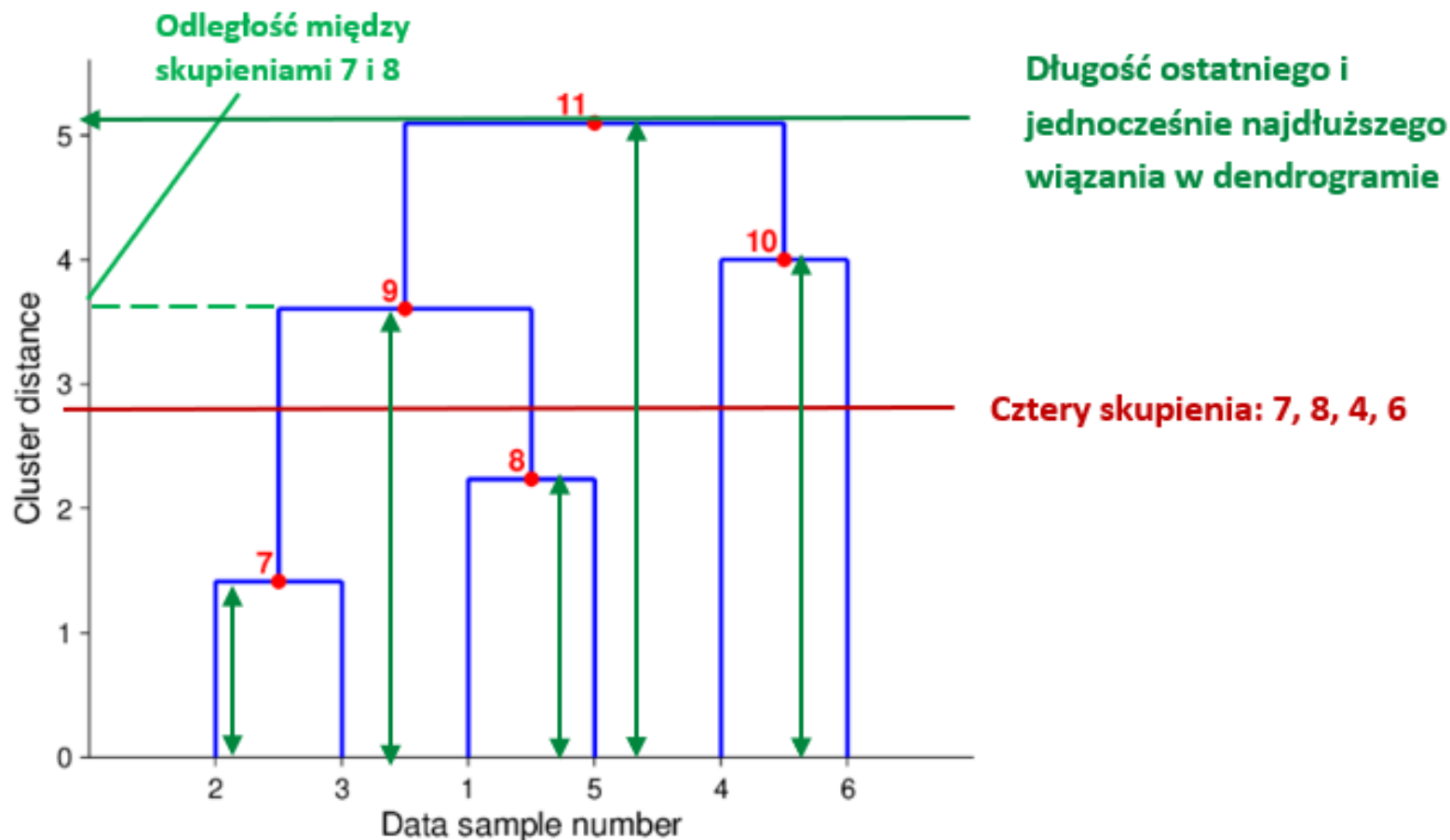
Metoda centroidów (centroid linkage); minimalizuje odległość między centroidami (środkami masy) łączonych klastrów



Metoda Warda (single linkage, pojedynczego wiązania); minimalizuje przyrost sumy kwadratów odchyleń wewnątrz klastrów (wariancji), łącząc te klastry, które powodują najmniejszy wzrost błędu



Dendrogram



Dendrogram przedstawia:

- sposób łączenia obiektów w skupienia
- kolejność tych połączeń
- odległości (lub niepodobieństwa) między skupieniami

Objaśnienie dendrogramu

Krok 1. **Obiekty (liście drzewa)**

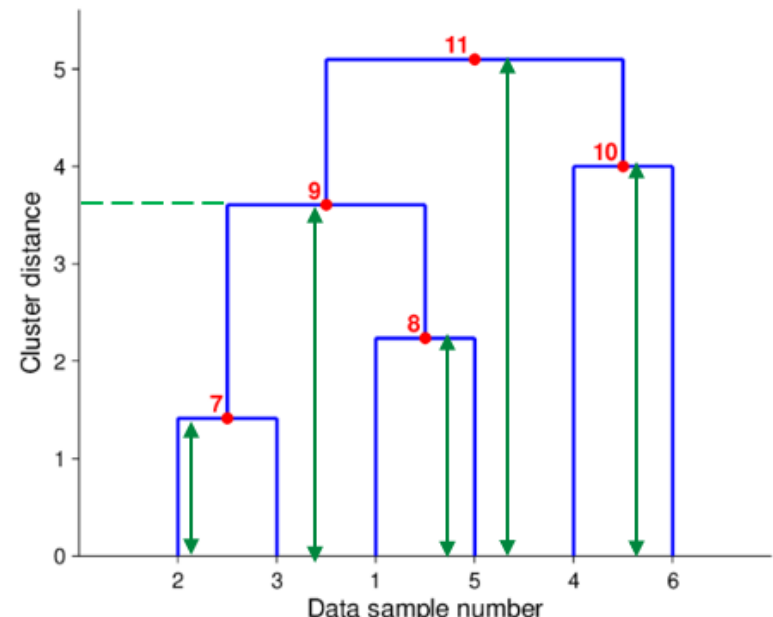
- Na dolnej osi (w poziomym dendrogramie: na dole) znajdują się pojedyncze obiekty (np. obserwacje).
- Każdy obiekt to oddzielny punkt początkowy analizy.

Krok 2. **Gałęzie i łączenia**

- Linie (gałęzie) łączą obiekty lub grupy obiektów w większe skupienia.
- Im wcześniej (niżej na osi odległości) dwa obiekty się połączą, tym bardziej są do siebie podobne.

Krok 3. **Wysokość połączenia**

- Wysokość, na której łączą się gałęzie, oznacza odległość/niepodobieństwo między skupieniami.
- Mała wysokość → bardzo podobne obiekty.
- Duża wysokość → połączenie odległych skupień.



Dane do analizy

Zbiór danych zawiera zmienne ilościowe niosące informację o tym, na jakim poziomie jest zielona transformacja w wybranych krajach Europy. Opis metadanych oraz dane są zawarte w pliku *WDI_ZInTrns.xlsx*.

Źródło danych: Bank Światowy (WORLD BANK GROUP):

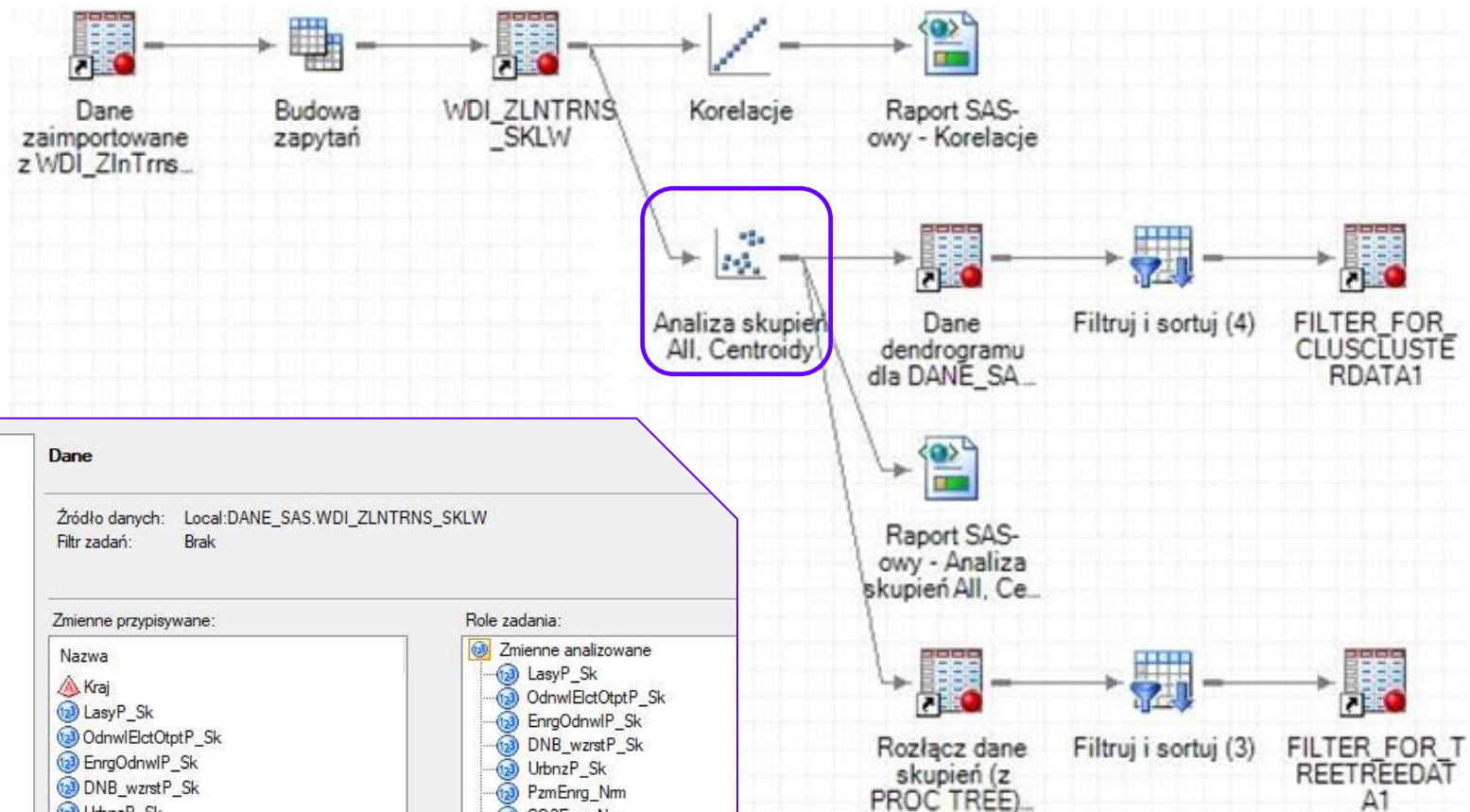
<https://databank.worldbank.org/source/world-development-indicators>

WDI (World Development Indicator) wybrane z bazy danych oferowanej na ww. stronie. Pobrano dane wg założeń:

- **Database** (dostępnych 87 tematów); World Development Indicator (liczebność dalszych pozycji zależy od wybrania tematu bazy)
- **Country** (266); zaznaczono kraje europejskie (jest ok. 50-51 państw, wliczając terytoria uznawane geograficznie za część Europy), np. kraje duże (Niemcy, Francja, Włochy, Hiszpania, Polska, Ukraina), małe kraje i mikro-państwa (Lichtenstein, Monako, San Marino, Watykan), transkontynentalne (Rosja, Turcja, Azerbejdżan, Gruzja (dwa ostatnie leżą w Azji, ale ze względu na uwarunkowania i powiązania historyczne są czasami zaliczane do krajów europejskich), po czym wybrano 26 krajów - por. skoroszyt Excela
- **Variables** (1516); wstępnie 8 zmiennych (szeregów statystycznych) - por. skoroszyt Excela
- **Time** (65 lat, od 1960) 3 ostatnie lata: 2018-2024 - - por. skoroszyt Excela (zauważyć braki w danych)

Aglomeracyjna analiza skupień w SAS EG

Diagram projektu



Dane

Skupienie
Wykresy
Rezultaty
Tytuły
Właściwości

Dane

Źródło danych: Local:DANE_SAS.WDI_ZLNTRNS_SKLW
Filtr zadań: Brak

Zmienne przypisywane:

Nazwa
▲ Kraj
🔍 LasyP_Sk
🔍 OdnwlElctOptP_Sk
🔍 EnrgOdnwlP_Sk
🔍 DNB_wzrstP_Sk
🔍 UrbnzP_Sk
🔍 PzmEnrg_Nm
🔍 CO2Ems_Nm
🔍 DNB_oS_Nm
🔍 minim
🔍 maxim
🔍 Rozstep
🔍 Licznik

Role zadania:

- 🔍 Zmienne analizowane
 - 🔍 LasyP_Sk
 - 🔍 OdnwlElctOptP_Sk
 - 🔍 EnrgOdnwlP_Sk
 - 🔍 DNB_wzrstP_Sk
 - 🔍 UrbnzP_Sk
 - 🔍 PzmEnrg_Nm
 - 🔍 CO2Ems_Nm
 - 🔍 DNB_oS_Nm
- 🔍 Grupuj analizowane wg
- 🔍 Liczebność (Limit: 1)
- 🔍 Etykieta identyfikacyjna (Limit: 1)
- ▲ Kraj
- 🔍 Zmienne kopiowane

Aglomeracyjna analiza skupień w SAS EG

Ustawienie węzła analizy skupień

Dane
Skupienie
Wykresy
Rezultaty
Tytuły
Właściwości

Skupienie

Metoda analizy skupień

☐ Łączenie średnich

☒ Metoda centroidów

☐ Algorytm k-średnich

☐ Metoda minimalnej wariancji Warda

Opcje metody k-średnich

Maksymalna liczba skupień:

Maksymalna liczba iteracji:

Zastępowanie ziarna:

☐ Podaj ziarno losowe:

Dane
Wykresy
Rezultaty
Tytuły
Właściwości

Wykresy

Wykresy i dendrogramy

 ☒ Statystyka pseudo-F

☒ Statystyka pseudo t-kwadrat

☒ Statystyka CCC

 ☒ Orientacja dendrogramu

☐ Pionowa

☒ Pozioma

Dane
Skupienie
Wykresy
Rezultaty
Tytuły
Właściwości

Rezultaty

Zapisz dane wynikowe

☒ Dane dendrogramu

Local:WORK.CLUSClusterData

☒ Wynik drzewiasty

☒ Skupienie:

Liczba skupień:

☐ Przycinanie:

Poziom:

Local:WORK.TREETreeData

☐ Skupienia k-średnich

Local:WORK.CLKMKMeansData

☒ Wyświetl wynik.

☐ Proste statystyki agregujące

☒ Sześciennne kryterium skupienia (CCC)

☒ Statystyki pseudo-F i t-kwadrat

Generacje skupień:

Aglomeracyjna analiza skupień w SAS

Tabela wyników

Nazwa	Typ	Etykieta	
 _NCL_	Numeryczna	Liczba skupień	Liczba klastrow pozostała po połączeniu Liczba obiektów w nowotworzonym klastrze Wysokość łączenia na dendrogramie
 _FREQ_	Numeryczna	Liczność skupienia	
 _HEIGHT_	Numeryczna	Odl. między centroidami skupień	
 _RMSSTD_	Numeryczna	Pierwiastek ze średniokwadratowego	
 _SPRSQ_	Numeryczna	Semicząstkowe R-kwadrat	
 _RSQ_	Numeryczna	R-kwadrat	
 _PSF_	Numeryczna	Statystyka pseudo-F	
 _PST2_	Numeryczna	Statystyka pseudo-t-kwadrat	
 _ERSQ_	Numeryczna	Przybliżone oczekiwane R-kwadrat	
 _RATIO_	Numeryczna	Przybliżone oczekiwane R-kwadrat	
 _LOGR_	Numeryczna	_LOGR_	
 _CCC_	Numeryczna	Sześcienne kryterium skupień	
 LasyP_Sk	Numeryczna	LasyP_Sk	
 OdnwlElctOtp...	Numeryczna	OdnwlElctOtpP_Sk	
 EnrgOdnwlP_...	Numeryczna	EnrgOdnwlP_Sk	
 DNB_wzrstP_...	Numeryczna	DNB_wzrstP_Sk	
 UrbnzP_Sk	Numeryczna	UrbnzP_Sk	
 PzmEnrg_Nm	Numeryczna	PzmEnrg_Nm	
 CO2Ems_Nm	Numeryczna	CO2Ems_Nm	
 DNB_oS_Nm	Numeryczna	DNB_oS_Nm	
 _DIST_	Numeryczna	Odległość między centroidami	
 _AVLINK_	Numeryczna	Średnie połączenia między skupieniami	
 Kraj	Znakowa	Kraj	

Aglomeracyjna analiza skupień w SAS

Wskaźniki jakości skupień - 1

SPRSQ (*Semi-Partial R-Squared*). Spadek wyjaśnionej wariancji po połączeniu dwóch klastrów. Pokazuje, ile wariancji w danych objaśnianej przez klastry zostaje utracone, gdy dwa skupienia zostają połączone. Małe wartości – połączenie nie pogarsza znacznie jakości. Duże wartości – połączenie grup bardzo różnych, utrata jakości – dobry moment (wskazane zatrzymanie aglomeracji).

RSQ (*R-Squared*). Procent całkowitej wariancji danych wyjaśnionej przez aktualny podział na klastry. Maleje przy kolejnych połączeniach (bo skupień coraz mniej). Im wyższy RSQ, tym bardziej homogeniczne skupienia. Często analizuje się nagły spadek RSQ (tzw. „łokieć”) w celu wyboru liczby klastrów.

ERSQ (*Expected R-Squared*). Oczekiwana wartość RSQ, jeśli dane byłyby losowe. Umożliwia porównanie jakości rzeczywistego podziału z przypadkowym. $RSQ \gg ERSQ$ oznacza, że klasteryzacja ma sens. $RSQ \approx ERSQ$ sugeruje, że podział nie jest lepszy niż losowy.

Aglomeracyjna analiza skupień w SAS

Wskaźniki jakości skupień - 2

CCC (*Cubic Clustering Criterion*). Wskaźnik proponowany przez SAS do wyboru liczby klastrów. Obliczany na podstawie RSQ, ERSQ i liczby obserwacji.

- $CCC > 2$ lub > 3 – dobre skupienia (wyraźna struktura).
- CCC w przedziale $[0, 2)$ – potencjalne skupieniem; traktować ostrożnie.
- $CCC < 0$ – brak wyraźnych skupień; duże wartości ujemne mogą wskazywać na obecność obserwacji odstających.

Lokalne maksima – często interpretowane jako optymalna liczba skupień.

PSF (*Pseudo F Statistic*) znane także jako indeks Calinskiego-Harabasza. Miara oceny, jak dobrze algorytm grupowania poradził sobie z podziałem danych.

- Duże PSF – klastry dobrze odseparowane, spójne; im wyższa wartość, tym lepsze rozdzielenie między klastrami.
- Spadek PSF – może sugerować utratę struktury.
- Lokalne maksimum PSF – często wskazuje sensowną liczbę klastrów.

PST² (*Pseudo t² Statistic*). Mierzy różnicę między dwoma ostatnimi etapami łączenia.

- Duża wartość PST² oznacza, że połączenie dwóch klastrów bardzo różnych – czyli prawdopodobnie należy zatrzymać się przed tym etapem.
- Małe PST² – łączenie grup podobnych (dalsze łączenia są jeszcze sensowne).

Sposoby poprawy jakości skupień

Zmienić metodę aglomeracji

Zastosować inną miarę odległości

Zredukować liczbę zmiennych

Usunąć punkty (rekordy) odstające

Zapytanie: Budowa zapytań Filtry NOR Wynik: DANE_SAS.WDI_ZLNTRNS_FLT Zmień...

Kolumny wyliczane Menedżer podpowiedzi Podgląd Narzędzia Opcje

Dodaj tabelę Usuń Złącz tabelę Wybierz dane Filtruj dane Sortuj dane

t1 (WDI_ZLNTRNS_SKLW)

- Kraj
- LasyP_Sk
- OdnwIElctOptP_Sk
- EnrgOdnwIP_Sk
- DNB_wzrstP_Sk
- UrbnzP_Sk
- PzmEnrg_Nm
- CO2Ems_Nm
- DNB_oS_Nm
- minim
- maxim
- Rozstep
- Licznik

Nazwa kolumny	Kolumna źródłowa	Podsu
Kraj	t1.Kraj	
LasyP_Sk	t1.LasyP_Sk	
EnrgOdnwIP_Sk	t1.EnrgOdnwIP_Sk	
OdnwIElctOptP...	t1.OdnwIElctOptP_Sk	
DNB_wzrstP_Sk	t1.DNB_wzrstP_Sk	
UrbnzP_Sk	t1.UrbnzP_Sk	
PzmEnrg_Nm	t1.PzmEnrg_Nm	
CO2Ems_Nm	t1.CO2Ems_Nm	
DNB_oS_Nm	t1.DNB_oS_Nm	

☐ Wybierz tylko wiersze bez powtórzeń

Uruchom Zapisz i zamknij

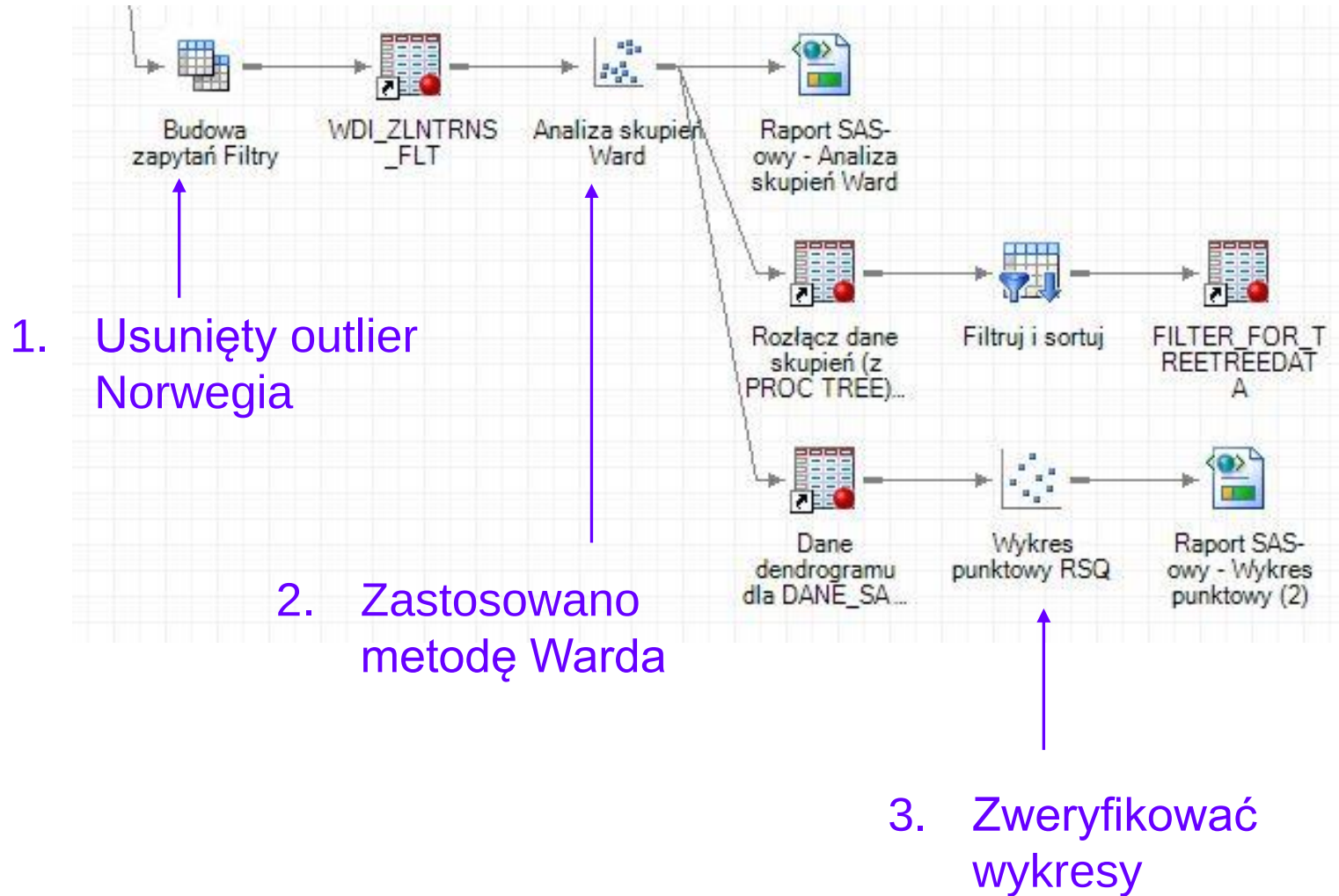
Filtruj dane surowe Operator

Where

t1.Kraj NOT = 'NOR'

Uruchom Zapisz i zamknij Anuluj Pomoc

Modyfikacja diagramu przepływu



Wyniki przykładowe – analiza

Historia skupień									
Liczba skupień	Skupienia połączone	Liczebn.	Półcząstkowe R-kwadrat	R-kwadrat	Przybliżone oczekiwane R-kwadrat	Sześciennne kryterium skupień	Statystyka pseudo-F	Pseudo t-kwadrat	Równe wartości
24	ESP ITA	2	0.0014	.999	.	.	30.5	.	
23	HRV LTU	2	0.0015	.997	.	.	30.8	.	
22	CL23 PRT	3	0.0039	.993	.	.	20.6	2.6	
21	CL24 GRC	3	0.0041	.989	.	.	18.1	2.9	
20	BEL NLD	2	0.0046	.984	.	.	16.6	.	
19	CZE POL	2	0.0057	.979	.	.	15.3	.	
18	SVK SVN	2	0.0065	.972	.	.	14.4	.	
17	CYP HUN	2	0.0069	.965	.	.	13.9	.	
16	CL22 LVA	4	0.0074	.958	.	.	13.6	2.7	
15	FRA GBR	2	0.0095	.948	.	.	13.1	.	
14	CL20 DEU	3	0.0111	.937	.	.	12.6	2.4	
13	CL21 CL15	5	0.0114	.926	.	.	12.5	2.3	
12	CHE DNK	2	0.0149	.911	.	.	12.1	.	
11	CL17 CL13	7	0.0189	.892	.	.	11.6	2.8	
10	CL19 CL18	4	0.0226	.869	.	.	11.1	3.7	
9	CL12 IRL	3	0.0312	.838	.	.	10.4	2.1	
8	FIN SWE	2	0.0313	.807	.	.	10.1	.	
7	CL10 EST	5	0.0352	.772	.	.	10.1	3.0	
6	AUT CL8	3	0.0399	.732	.	.	10.4	1.3	
5	CL11 CL16	11	0.0774	.654	.707	-1.5	9.5	10.7	
4	CL14 CL7	8	0.0826	.572	.641	-1.8	9.3	5.8	
3	CL6 CL4	11	0.1366	.435	.532	-2.0	8.5	5.1	
2	CL3 CL5	22	0.1835	.252	.351	-1.8	7.7	7.1	
1	CL2 CL9	25	0.2516	.000	.000	0.00	.	7.7	

Wskaźnik	Optymalny punkt cięcia
SPRSQ	Zatrzymać się przy 6
RSQ	Największy spadek po 6
PST ²	Największy skok przed 6
PSF	Ostatni wysoki poziom 6–7
CCC	Negatywny poniżej 6 → nie schodzić niżej