

Analiza wyników polskich matur rozszerzonych względem województw

Wydział Zastosowań Informatyki i Matematyki
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

Autorzy:
O. Deviatkin
K. Kochański
R. Kuligowski
M. Laszczka
P. Majewski

Streszczenie: Celem badań jest analiza wyników polskich matur dla poszczególnych województw za pomocą metod liniowego porządkowania obiektów i analizy skupień. Do realizacji tego celu pobrano oficjalną statystykę średnich wyników matur rozszerzonych pisanych w poszczególnych województwach. Następnie wykorzystując metody Hellwiga, TOPSIS, Nowak. Na podstawie wyników danych metod dokonano podziału województw na cztery grupy: bardzo wysokich wyników, wysokich, przeciętnych i niskich. Również przeprowadzono analizę skupień metodą k-średnich dla podzielonych danych na dwie podgrupy: wyników matur z przedmiotów ścisłych oraz nie ścisłych. Na podstawie wyników klasteryzacji ujawniono cztery charakterystyczne grupy: wysokie wyniki w obu podgrup, wysokie wyniki tylko z przedmiotów ścisłych, tylko z przedmiotów nieścisłych oraz niskie wyniki z obu podgrup przedmiotów.

Słowa kluczowe: matura rozszerzona, porządkowanie liniowe, klasteryzacja, boxplot, k-średnich, Hellwig, TOPSIS, Nowak

Wprowadzenie

Metody porządkowania liniowego są wykorzystywane w badaniach ekonomicznych w celu ustalenia kolejności lub klasyfikacji obiektów, takich jak kraje (ze względu na poziom rozwoju gospodarczego), przedsiębiorstwa (ze względu na kondycję finansową), produkty (ze względu na walory użytkowe) itp. Idea porządkowania liniowego obiektów wielowymiarowych opiera się na pojęciu porządkującej relacji binarnej (zwrotnej, antysymetrycznej, przechodniej i spójnej). Z aksjomatów tej relacji wynika, że jest możliwe stwierdzenie, który z dwóch dowolnych obiektów zbioru jest pierwszy (lepiej), a który drugi (gorszy), a także – czy są one identyczne.

W Polsce w 2021 roku do matur przystąpiło 273 419 studentów. Postawiono tezę, że średnie wyniki matur rozszerzonych są zróżnicowane i zależą od województw w których zostały przeprowadzone. Celem artykułu jest prezentacja wyników badań, wykonanych za pomocą metod porządkowania liniowego oraz klasteryzacji, dotyczących oceny jakości poszczególnych województw pod względem średnich wyników matur.

Dane i metody

Analizę wyników matur przeprowadzona na podstawie danych wtórnych Centralnej Komisji Egzaminacyjnej (tabela 1).

Tabela 1. Wartości cech diagnostycznych

Województwa	Cechy diagnostyczne [%]								
	X1	X2	X3	X4	X5	X6	X7	X8	X9
Dolnośląskie	31,20	31,37	30,83	30,57	33,92	39,49	65,60	49,96	28,07
Kujawsko-Pomorskie	30,29	30,97	36,45	30,84	35,65	41,05	65,03	44,49	27,77
Lubelskie	33,05	36,91	38,04	31,21	35,11	33,74	61,29	46,15	31,20
Lubuskie	31,98	33,78	31,14	31,63	31,97	33,37	64,60	54,33	25,27
Łódzkie	33,46	32,78	37,91	34,82	36,41	33,13	62,89	41,91	32,37
Małopolskie	35,00	38,67	42,64	34,68	42,69	44,50	64,33	49,29	35,13
Mazowieckie	35,25	37,49	44,74	38,57	40,27	40,65	68,80	51,77	36,80
Opolskie	32,25	27,82	27,72	31,54	32,44	33,82	64,50	48,21	27,41
Podkarpackie	32,82	35,50	34,35	30,91	35,21	31,64	60,81	44,63	30,84
Podlaskie	35,13	42,04	35,83	33,59	36,81	43,42	64,37	52,55	32,33
Pomorskie	32,87	36,41	38,15	34,14	33,13	38,31	67,02	43,11	31,67
Śląskie	32,67	34,38	35,55	34,28	35,16	37,45	65,16	51,96	29,05
Świętokrzyskie	37,17	36,39	34,23	31,18	36,84	33,47	61,69	49,73	29,43
Warmińsko-Mazurskie	30,39	30,40	36,55	29,96	31,50	30,51	63,78	46,82	26,22
Wielkopolskie	32,22	32,81	35,28	32,15	36,45	39,17	64,25	52,65	28,58
Zachodniopomorskie	32,85	33,54	39,14	30,48	34,55	35,51	65,76	54,57	27,86

Do analizy przyjęto następujące cechy diagnostyczne:

- X1 – wyniki śr. z rozsz. biologii
- X2 – wyniki śr. z rozsz. chemii
- X3 – wyniki śr. z rozsz. fizyki
- X4 – wyniki śr. z rozsz. geografii
- X5 – wyniki śr. z rozsz. historii
- X6 – wyniki śr. z rozsz. informatyki
- X7 – wyniki śr. z rozsz. j. angielskiego
- X8 – wyniki śr. z rozsz. j. polskiego
- X9 – wyniki śr. z rozsz. matematyki

Do analizy jako cechy diagnostyczne wybrano wyniki procentowe z 8 najpopularniejszych matur rozszerzonych oraz rozszerzenie z informatyki z podziałem na województwa.

Także powyższe dane podzielono na dwie podgrupy według kategorii matur ścisłych oraz nieścisłych (tabela 2).

Tabela 2. Wartości cech diagnostycznych według podgrup

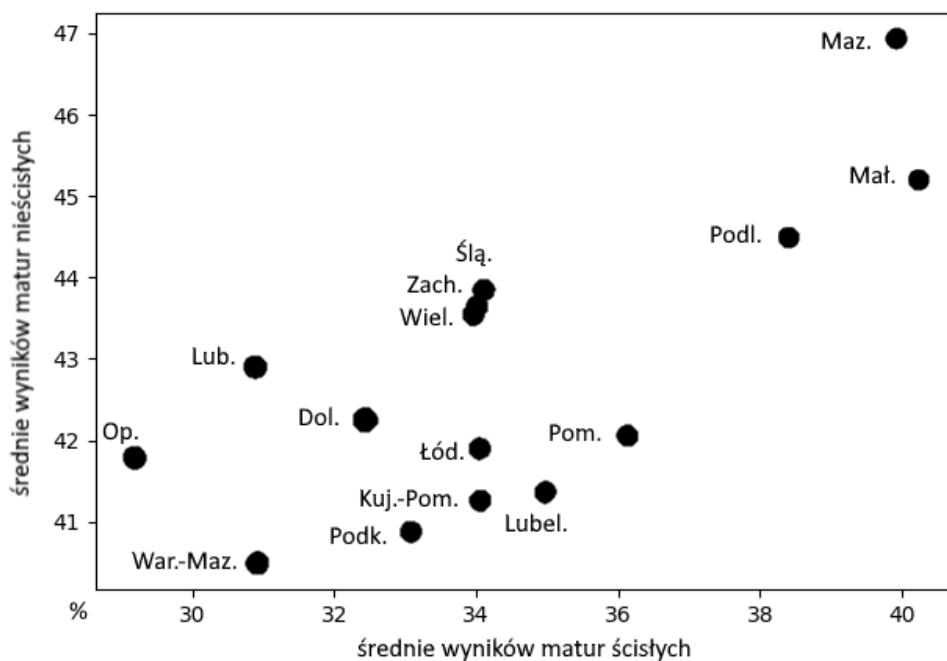
Województwa	Cechy diagnostyczne [%]	
	ściśle	nieściśle
Dolnośląskie	32,44	42,25
Kujawsko-Pomorskie	34,06	41,26
Lubelskie	34,97	41,36
Lubuskie	30,89	42,90
Łódzkie	34,05	41,90
Małopolskie	40,24	45,20
Mazowieckie	39,92	46,93
Opolskie	29,19	41,79
Podkarpackie	33,08	40,88
Podlaskie	38,41	44,49
Pomorskie	36,14	42,05
Śląskie	34,11	43,85
Świętokrzyskie	33,38	43,32
Warmińsko-Mazurskie	30,92	40,49
Wielkopolskie	33,96	43,54
Zachodniopomorskie	34,01	43,64

Do analizy przyjęto następujące cechy diagnostyczne:

- ściśle – śr. z cech X2, X3, X6, X9
- nieściśle – śr. z cech X1, X4, X5, X7, X8

Również, na podstawie wyników matur według podgrup sporządzono wykres.

Wykres 1. Wartości cech diagnostycznych według podgrup



Wybrane matury rozszerzone obejmują wszystkie najbardziej znaczące przedmioty, które są najczęściej brane pod uwagę na uczelniach w procesie rekrutacji. Nie brano pod uwagę mniej popularnych rozszerzeń ze względu na małą próbkę statystyczną oraz brak danych dla niektórych województw (jeżeli w danym województwie mniej niż 10 osób napisało maturę to ich średni wynik nie był publikowany). Starano się też zachować podobną liczbę cech w obu podgrupach. Wszystkie cechy są stymulantami. Wartości charakterystyk liczbowych cech diagnostycznych przedstawiono niżej (tabela 3).

Tabela 3. Wartości cech diagnostycznych

Charakterystyki liczbowe	Cechy diagnostyczne [%]								
	X1	X2	X3	X4	X5	X6	X7	X8	X9
Max	37,17	42,04	44,74	38,57	42,69	44,50	68,80	54,57	36,80
Min	30,29	27,82	27,72	29,96	31,50	30,51	60,81	41,91	25,27
Średnia	33,04	34,45	36,16	32,53	35,51	36,83	64,37	48,88	30,00
Mediana	32,83	34,08	36,14	31,58	35,18	36,48	64,44	49,51	29,24
Odchylenie standardowe	1,84	3,54	4,23	2,29	2,91	4,27	2,05	4,01	3,15
Współczynnik zmienności	0,06	0,10	0,12	0,07	0,08	0,12	0,03	0,08	0,10
Max/Min	1,23	1,51	1,61	1,29	1,35	1,46	1,13	1,30	1,46

Zauważmy, że cechy X1-X9 w badanej grupie obiektów charakteryzuje bardzo małe zróżnicowanie, stąd prawie identyczna średnia oraz mediana dla każdej z podgrup.

Wartości charakterystyk liczbowych cech diagnostycznych podzielonych na podgrupę przedstawiono niżej (tabela 4).

Tabela 4. Analiza cech diagnostycznych według podgrup

Charakterystyki liczbowe	Cechy diagnostyczne [%]	
	ściśle	nieściśle
Max	40,24	46,93
Min	29,19	40,49
Średnia	34,36	42,87
Mediana	34,03	42,58
Odchylenie standardowe	3,08	1,72
Współczynnik Zmienności	0,09	0,04

W pierwszym etapie badań wytypowano trzy procedury porządkowania liniowego:

- metoda Topsis
- metoda Hellwiga
- metoda Nowaka

Szczegółowy opis każdej z metod zamieszczony jest niżej:

Metoda TOPSIS

- Normalizacja

zmiennych:

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^n x_{ij}^2}},$$

gdzie

- x_{ij} obserwacja j-tej zmiennej dla obiektu i;
- Współrzędne wzorca: $z_{0j}^+ = \max_i \{z_{ij}\}$ dla zmiennych stymulant;
 - Współrzędne antywzorca: $z_{0j}^- = \min_i \{z_{ij}\}$ dla zmiennych stymulant;
 - Odległości obiektów od wzorca: $d_{i0}^+ = \sqrt{\sum_{j=1}^m (z_{ij} - z_{0j}^+)^2}$;
 - Odległości obiektów od antywzorca: $d_{i0}^- = \sqrt{\sum_{j=1}^m (z_{ij} - z_{0j}^-)^2}$;
 - Wartości zmiennej agregatowej:
 $q_i = \frac{d_{i0}^-}{d_{i0}^+ + d_{i0}^-}$, gdzie
 $0 \leq q_i \leq 1$, $\max_i \{q_i\}$ - element najlepszy, $\min_i \{q_i\}$ - element najgorszy;

Metoda Hellwiga

- Normalizacja zmiennych: $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$, gdzie
 x_{ij} obserwacja j-tej zmiennej dla obiektu i,
 \bar{x}_j średnia arytmetyczna obserwacji j-tej zmiennej
 s_j odchylenie standardowe obserwacji j-tej zmiennej
- Współrzędne wzorca: $z_{0j} = \max_i (z_{ij})$ dla zmiennych stymulant;
- Odległości obiektów od wzorca: $d_{i0} = \sqrt{\sum_{j=1}^m (z_{ij} - z_{0j})^2}$;
- Wartości zmiennej agregatowej:
 $q_i = 1 - \frac{d_{i0}}{d_0}$, gdzie:
 $0 \leq q_i \leq 1$, $\max \{q_i\}$ - element najlepszy, $\min \{q_i\}$ - element najgorszy;
 $d_0 = \bar{d}_0 + 2s_d$;
 $s_d = \sqrt{\frac{1}{n} \sum_{i=1}^n (d_{i0} - \bar{d}_0)^2}$.

Metoda Nowaka

- Normalizacja zmiennych: $z_{ij} = \frac{x_{ij}}{\bar{x}_j}$, gdzie
 x_{ij} obserwacja j-tej zmiennej dla obiektu i
 \bar{x}_j średnia arytmetyczna obserwacji j-tej
- Wartość zmiennej agregatowej:

$$Q_i = \frac{1}{m} \sum_{j=1}^m z_{ij}$$

Następnie sporządzono na ich podstawie tabelę wyników badanych obiektów dla poszczególnych metod (tabela 5).

Tabela 5. Wyniki obliczeń według poszczególnych metod porządkowania liniowego

Województwa	Metody		
	Topsis	Hellwiga	Nowaka
Dolnośląskie	0,156025	0,243164	0,961287
Kujawsko-Pomorskie	0,101793	0,246206	0,97072
Lubelskie	0,407757	0,30354	0,991583

Lubuskie	0,253371	0,208139	0,949524
Łódzkie	0,457662	0,339088	0,991258
Małopolskie	0,692229	0,667013	1,112928
Mazowieckie	0,726448	0,795715	1,130051
Opolskie	0,276107	0,140076	0,917066
Podkarpackie	0,369057	0,225525	0,96226
Podlaskie	0,701574	0,561972	1,075035
Pomorskie	0,387032	0,408323	1,012024
Śląskie	0,355362	0,421173	1,008253
Świętokrzyskie	0,857819	0,339127	0,998824
Warmińsko-Mazurskie	0,072413	0,112509	0,91913
Wielkopolskie	0,294574	0,36864	1,001056
Zachodniopomorskie	0,378537	0,3473	0,999002

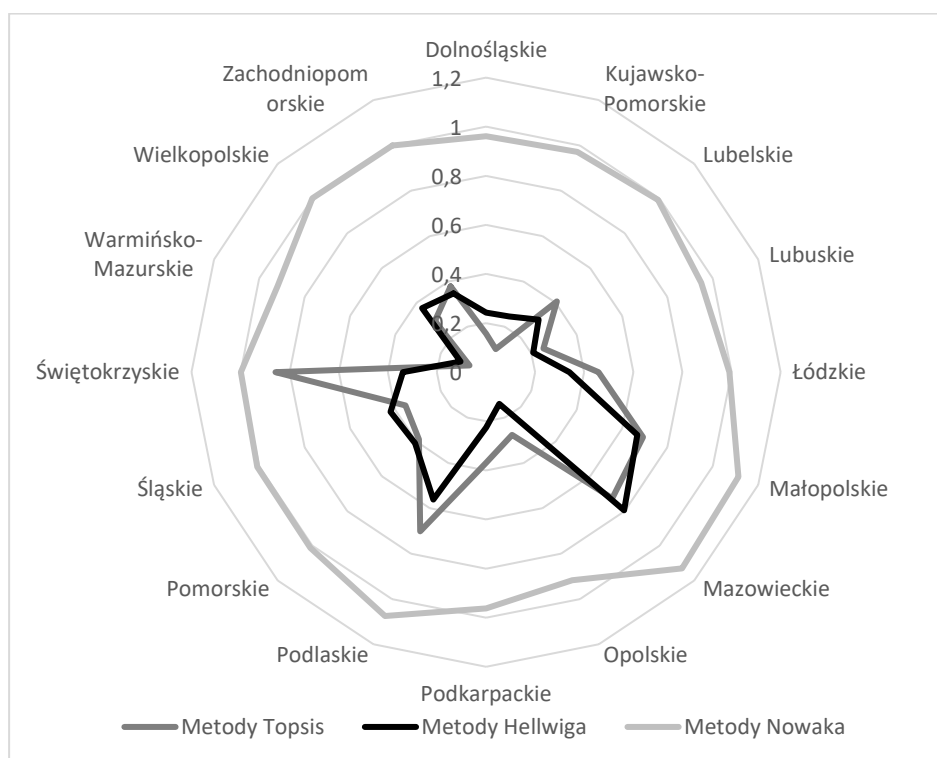
Także stworzono ranking województw według wyników dla poszczególnych metod (tabela 6)

Tabela 6. Ranking województw według poszczególnych metod porządkowania liniowego

Województwa	Metody		
	Topsis	Hellwiga	Nowaka
Świętokrzyskie	1	8	8
Mazowieckie	2	1	1
Podlaskie	3	3	3
Małopolskie	4	2	2
Łódzkie	5	9	10
Lubelskie	6	10	9
Pomorskie	7	5	4
Zachodniopomorskie	8	7	7
Podkarpackie	9	13	12
Śląskie	10	4	5
Wielkopolskie	11	6	6
Opolskie	12	15	16
Lubuskie	13	14	14
Dolnośląskie	14	12	13
Kujawsko-Pomorskie	15	11	11
Warmińsko-Mazurskie	16	16	15

Oraz wizualizację graficzną (wykres 2), obrazującą wyniki dla każdej z metod.

Wykres 2. Wyniki obliczeń według poszczególnych metod porządkowania liniowego



W drugim etapie analizy spośród sporządzonych rankingów wybrano ten, który jest najbardziej podobny do pozostałych, czyli ten dla którego wartość w wektorze wartości miary podobieństwa $\overline{u_p}$ jest największa. Szczegółowy opis metody wyliczania $\overline{u_p}$:

$$- \quad \overline{u_p} = \frac{1}{v-1} \sum_{\substack{q=1 \\ p \neq q}}^v m_{pq} \quad p, q = 1, 2, \dots, v$$

gdzie:

- v - liczba rankingów

$$- \quad m_{pq} = 1 - \frac{2 \sum_{i=1}^n |c_{ip} - c_{iq}|}{n^2 - z} \quad \text{takie, że:}$$

c_{ip}, c_{iq} – odpowiednio pozycja “i” obiektu w rankingu o numerze p, q

$$z = \begin{cases} 0, & n \text{ jest parzyste} \\ 1, & n \text{ jest nieparzyste} \end{cases}$$

Wybrana w wyżej opisany sposób metoda stanowi podstawę do sporządzenia i interpretacji rankingu województw za względu na średnie wyniki matur rozszerzonych.

Tabela 6. Wartości przejściowe dla metody porównań

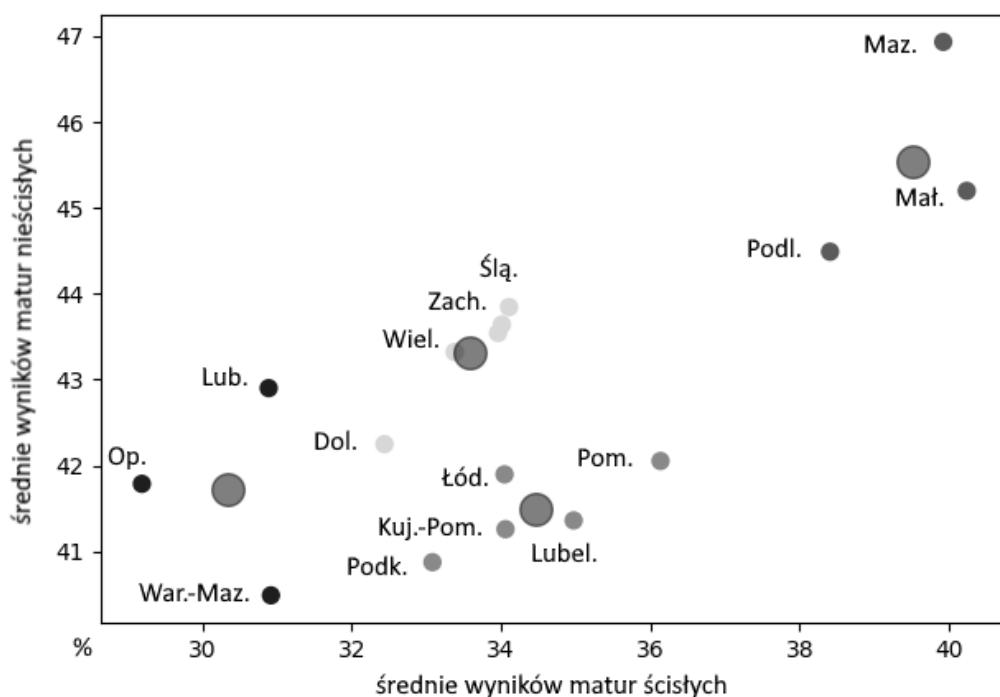
MPQ		
1	0,986215486	0,925685629
	1	0,919746185
		1

Wybrana w wyżej opisany sposób metoda stanowi podstawę do sporządzenia i interpretacji rankingu województw za względu na średnie wyniki matur rozszerzonych.

Za pomocą metody k-średnich udało się podzielić zbiór danych na cztery klastry (wykres 3). Zostało to przeprowadzone w następujących krokach:

- Krok pierwszy: Wybranie ilości centroidów (w naszym przypadku 4) i ich początkowego ułożenia.
- Krok drugi: Ustalenie przynależności punktów do konkretnych centroidów, wyliczamy średnie odległości poszczególnych punktów i przypisujemy najbliższy im centroid.
- Krok trzeci: Ustalenie nowego położenia centroidów. Nowe współrzędne centroidów to średnia arytmetyczna współrzędnych wszystkich punktów mających jego grupę.
- Krok czwarty: Powtarzanie krok 2 i 3 do momentu, w którym nie zmieniła się przynależność punktów do klastr.

Wykres 3. Wyniki obliczeń według k-średnich

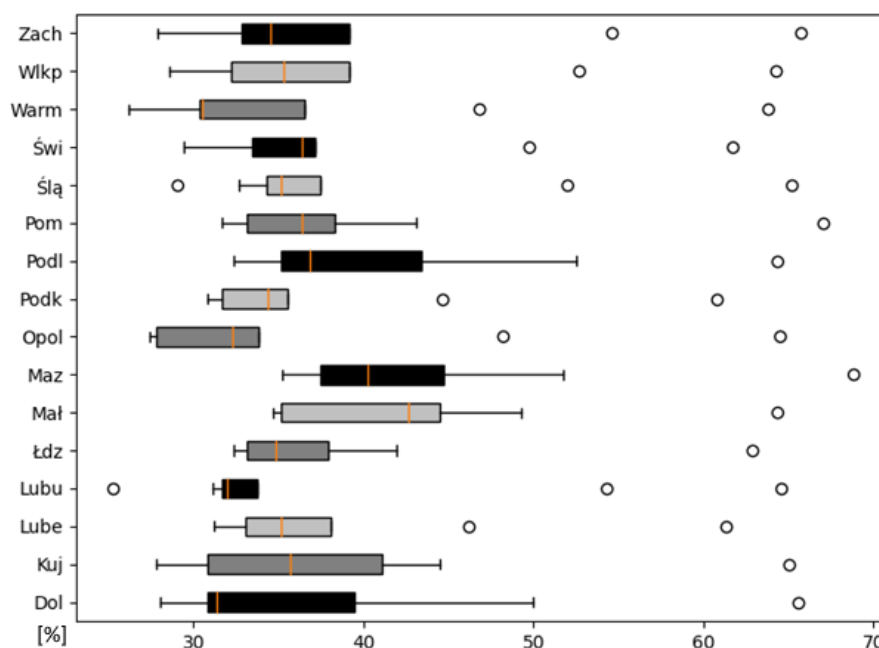


Wyniki badań

- Na podstawie metody k-średnich oraz podziału danych na dwie podgrupy według typu zdawanego przedmiotu ścisłe/nieścisłe wyłoniono cztery podgrupy województw według jakości edukacji w nich. Są to:
 - grupa najlepsza:
 - mazowieckie
 - małopolskie
 - podlaskie,
 - grupa dobra z przedmiotów ścisłych:
 - pomorskie
 - lubelskie

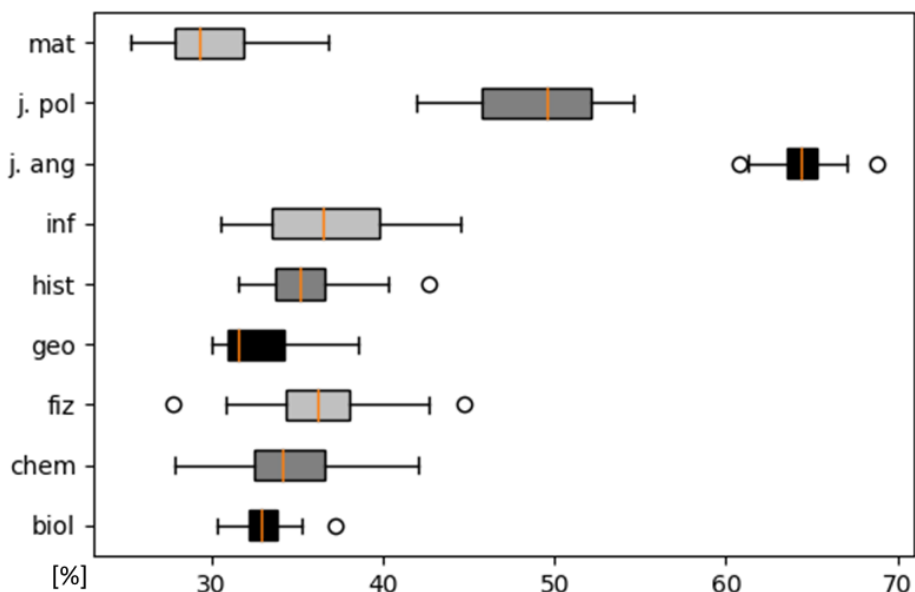
- łódzkie
 - kujawsko-pomorskie
 - podkarpackie
 - grupa dobra z przedmiotów nieściśle:
 - śląskie
 - zachodniopomorskie
 - wielkopolskie
 - dolnośląskie
 - grupa najgorsza:
 - lubuskie
 - opolskie
 - warmińsko-mazurskie
- Po wstępnej analizie wyjawiono, że średnie wyniki matur rozszerzonych jak i mediana w poszczególnych województwach na ogół są podobne i mieszczą się w granicach 25-45[%]. Podobnie wartości odstające układają się podobnie niezależnie od województwa. Jedną z takich wartości znacząco powyżej średniej jest w przedziale 60-70[%] i jest to język angielski.
 - W niektórych województwach można też zaobserwować inne wartości odstające takie jak język polski (lekko powyżej średniej) albo matematyka (lekko poniżej średniej). Jednak mimo podobieństwa średniej i mediany, odchylenie standardowe (rozstrzał wyników względem średniej) jest skrajnie różny, dla różnych województw np. województwo lubuskie oraz podlaskie (wykres 4).

Wykres 4. Boxplot średnich wyników poszczególnych matur dla województw



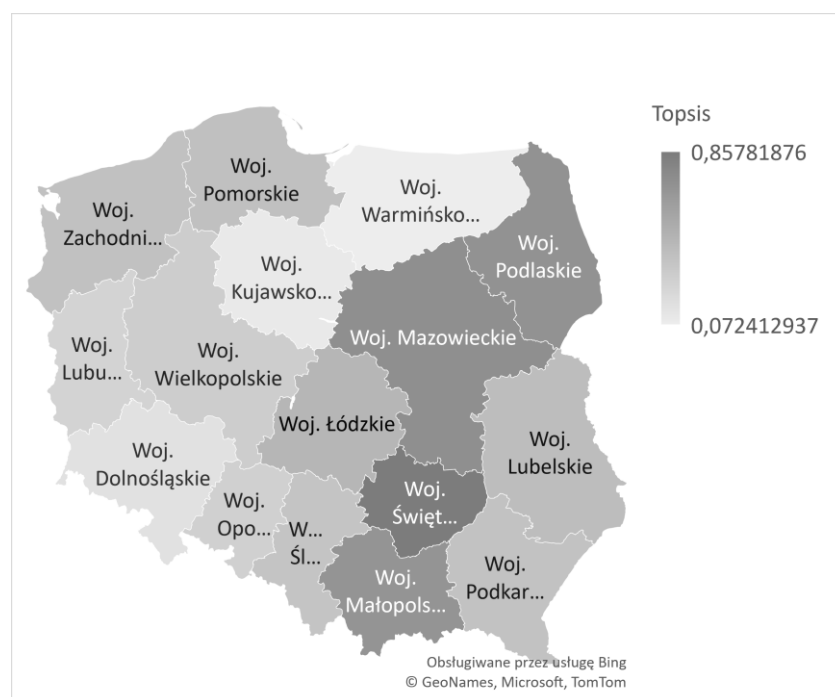
- Boxplot średnich matur rozszerzonych w poszczególnych województwach dla przedmiotów w całej Polsce dobrze obrazuje wyniki otrzymane z wcześniejszej analizy. Średni wynik matur z języka polskiego oraz jeszcze bardziej znacząco z języka angielskiego jest zdecydowanie wyższy od pozostałych przedmiotów (wykres 5).

Wykres 5. Boxplot średnich wyników matur województw dla poszczególnych przedmiotów



- Najbardziej podobnym rankingiem porządków liniowych zbioru jest ranking sporządzony na podstawie metody TOPSIS. Niżej znajduje się zobrazowanie wyników tej metody (wykres 6)

Wykres 6. Mapa obliczeń dla metody TOPSIS



- Za pomocą metod porządkowania liniowego, z których następnie była wybrana najbardziej reprezentująca pozostałe, wyłoniono jednoznacznego lidera wśród wszystkich województw pod względem ogólnych wyników matur – województwo Świętokrzyskie

Podsumowanie

Wbrew temu, że wyniki matur w poszczególnych województwach na ogół są bardzo podobne, da się wyłonić spośród tych województw jednoznacznego lidera – najlepsze województwo spośród wszystkich pod względem średnich wyników matur. Niemniej jednak wynik jest na pierwszy rzut oka mniej oczywisty, gdyż metoda Topsis wybrana do porządkowania liniowego obiektów uwzględnia nie tylko średnie wyniki jako takie, lecz ogólniejszą analizę wyników, taką jak odbieganie wartości od średniej i inne.

Metoda k-średnich podzieliła województwa na grupy, które dość dobrze pokrywają się z grupami województw o podobnych poziomach rozwoju – oznacza to, że grupa województw najbardziej rozwiniętych otrzymała najlepsze średnie wyniki z matur rozszerzonych.

Literatura

- Jacek B. 2016 *Identyfikacja obserwacji odstających w analizie skupień* Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
- Andrzej B. 2016 *Porządkowanie liniowe obiektów metodą Hellwiga i TOPSIS* Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
- Informatyczne centrum edukacji i nauki 2020 *Raport z Wyników matur* Wydawnictwo Główny Urząd Statystyczny
- Kukuła K., Luty L. 2017 *Jeszcze o procedurze wyboru metody porządkowania* Wydawnictwo Główny Urząd Statystyczny
- Pawełek B. 2008 *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych* Wydawnictwo Uniwersytet Ekonomiczny w Krakowie
- Bożek J. 2002 *O niektórych metodach porządkowania liniowego* Wydawnictwo Uniwersytetu Rzeszowskiego
- Zeliau A. 2002 *Uwagi na temat wyboru metody normowania zmiennych diagnostycznych* Wydawnictwo Uniwersytet M. Kopernika w Toruniu
- Kufel T., Piłatowska M. (red) *Analiza szeregów czasowych na początku XXI wieku* Wydawnictwo Uniwersytet M. Kopernika w Toruniu
- Kolenda M. 2003 *Zmiana cechy na stymulantę, w: Ekonometria 11* Wydawnictwo Uniwersytet Ekonomiczny we Wrocławiu, Wrocław.
- Kolonko J. 1980 *Analiza dyskryminacyjna i jej zastosowania w ekonomii* Wydawnictwo PWN, Warszawa.
- Sztemberg-Lewandowska M. 2016 *Grupowanie danych funkcjonalnych w analizie poziomu wiedzy maturzystów* Wydawnictwo Uniwersytet Ekonomiczny we Wrocławiu
- Przychodzeń B. 2011 *Analiza wyników egzaminów maturalnych z matematyki oraz fizyki i astronomii w województwie pomorskim w latach 2010-2011 z uwzględnieniem płci zdających* Wydawnictwo Okręgowa Komisja Egzaminacyjna w Gdańsku – Wydział badań i analiz