METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XIX/3, 2018, s. 263 – 273

Received: 25.06.2018
Accepted: 05.10.2018

ZASTOSOWANIE ANALIZY SKUPIEŃ I LASÓW LOSOWYCH W KLASYFIKACJI GMIN W POLSCE NA SKALI POZIOMU ROZWOJU SPOŁECZNO-GOSPODARCZEGO[1]

Robert Perdał https://orcid.org/0000-0002-2585-6898
Wydział Nauk Geograficznych i Geologicznych
Uniwersytet im. Adama Mickiewicza w Poznaniu
e-mail: r.perdal@amu.edu.pl

Streszczenie: W artykule przedstawiono algorytm klasyfikacji gmin na skali poziomu rozwoju społeczno-gospodarczego. Algorytm ten obejmuje cztery etapy: (1) dobór i redukcja zmiennych, (2) konstrukcja miernika syntetycznego i uszeregowanie liniowe gmin na skali poziomu rozwoju społeczno-gospodarczego, (3) grupowanie gmin metodą analizy skupień wg algorytmu *k*-średnich na podstawie wartości miernika syntetycznego, (4) weryfikacja klasyfikacji metodą lasów losowych. W wyniku procedury klasyfikacyjnej zidentyfikowano dywergencję rozwoju społeczno-gospodarczego w Polsce.

Słowa kluczowe: analiza skupień, lasy losowe, klasyfikacja, gminy, rozwój społeczno-gospodarczy

JEL classification: C38, C44, C55, O18

WSTĘP

Od lat 50. i 60. XX w., czyli tzw. „rewolucji ilościowej" w badaniach ekonomiczno-przestrzennych, geograficznych i regionalnych [Burton 1963], metody matematyczno-statystyczne stały się jednym z podstawowych narzędzi analitycznych, eksplanacyjnych i predykcyjnych, głównie dla zwolenników

[1] Praca powstała w ramach projektu badawczego finansowanego ze środków Narodowego Centrum Nauki (2015/19/B/HS5/00012) pt. „Nowe wyzwania polityki regionalnej w kształtowaniu czynników rozwoju społeczno-ekonomicznego regionów mniej rozwiniętych".

https://doi.org/10.22630/MIBE.2018.19.3.24

modelu empirycznego w ujęciu tradycyjno-empirycznym i empiryczno-indukcyjnym [Chojnicki 1985]. Dokonujący się postęp w zakresie technik i mocy obliczeniowych sprawił, że wzrosły możliwości analityczne. Jednakże nie wyeliminowało to, a wręcz można sądzić, że spotęgowało, problem efektywności, skuteczności i przydatności tego typu metod zwłaszcza do wyjaśniania i predykcji zjawisk społeczno-ekonomicznych w ujęciu przestrzennym. Stąd też jednym z podstawowych wyzwań współczesnych badań geograficzno-ekonomicznych na gruncie taksonomii numerycznej jest zagadnienie klasyfikacji, rozumianej dwojako. Po pierwsze, jako czynność wyodrębniania w ramach *n*-elementowego zbioru X takich niepustych jego *k*-podzbiorów, że spełnione są dwa warunki: (1) warunek adekwatności – suma wyodrębnionych podzbiorów (grup, klas) jest identyczna ze zbiorem: $X_1 \cup X_2 \cup \cdots \cup X_k = X$; (2) warunek rozłączności – podzbiory te nie zawierają elementów wspólnych: $X_i \cap X_j = \emptyset$ dla $i \neq j = 1, 2, \ldots, k$; po drugie – efekt tej czynności – konkretny podział przestrzenny [por. Chojnicki, Czyż 1973]. Częstokroć od wyników klasyfikacji zależą kolejne czynności badawcze oraz ich skuteczność, a także decyzje podejmowane przez przedstawicieli świata praktyki, w tym m.in. decyzje dotyczące polityki rozwoju prowadzonej i realizowanej w różnych jednostkach terytorialnych [por. Nijkamp 1986; Duranton 2015].

Celem niniejszej pracy jest przedstawienie nowego algorytmu klasyfikacji jednostek przestrzennych z zastosowaniem analizy skupień i lasów losowych. Zadaniem tego algorytmu jest uzyskanie możliwie wysokiego stopnia poprawności i efektywności klasyfikacji dużej liczby jednostek przestrzennych na podstawie wskaźników opisujących wybrane aspekty rozwoju społeczno-gospodarczego. Algorytm ten wykorzystano do klasyfikacji 2478 gmin w Polsce na skali poziomu rozwoju społeczno-gospodarczego. Składa się on z czterech etapów. W pierwszym etapie dokonano wyboru i redukcji zmiennych opisujących wybrane aspekty rozwoju (wykorzystano dane statystyczne publikowane w BDL GUS, dla lat 2004 i 2016). W drugim etapie, na podstawie wartości zmiennych diagnostycznych wyznaczono hipotetyczną gminę stanowiącą wzorzec rozwoju, a następnie za pomocą miary Braya-Curtisa wyznaczono stopień niepodobieństwa do wzorca każdej z badanych gmin. W wyniku tej procedury uzyskano syntetyczny miernik poziomu rozwoju. W trzecim etapie, na podstawie wartości syntetycznego miernika poziomu rozwoju, dokonano grupowania gmin metodą analizy skupień wg algorytmu *k*-średnich. W czwartym etapie za pomocą metody lasów losowych zweryfikowano uzyskane w trzecim etapie wyniki. Obliczenia wykonano za pomocą oprogramowania Statistica 13.1, a wizualizacje za pomocą ArcMap 10.5.1.

METODY BADAWCZE – ALGORYTM KLASYFIKACJI

Dobór i redukcja zmiennych

Początkowy zbiór zmiennych stanowiło 38 wskaźników opisujących poziom rozwoju społeczno-gospodarczego w układzie pięciu aspektów: kapitał ludzki (KL), kapitał społeczny (KS), kapitał materialny (KM), kapitał finansowy (KF), innowacje technologiczne i organizacyjne (IT). Z uwagi na postulat ograniczenia liczby zmiennych powielających tę samą informację [Zeliaś 2002] na zmiennych z 2016 r. przeprowadzono redukcję w oparciu o wartość krytyczną współczynnika korelacji liniowej Pearsona (r^*=0,3 dobrano subiektywnie, gdyż dla α=0,0000, przy t=4,426, r^*=0,089). Przy redukcji kierowano się także zdolnością dyskryminowania obiektów wyrażoną wielkością współczynnika zmienności (powyżej 15%) oraz merytoryczną oceną współzmienności i współzależności. Po redukcji zostały 23 zmienne diagnostyczne, które posłużyły do konstrukcji syntetycznego miernika poziomu rozwoju społeczno-gospodarczego dla 2004 jak i 2016 r.:

1. KL1: osoby w wieku nieprodukcyjnym na 100 osób w wieku produkcyjnym (DS – destymulanta)
2. KL2: przyrost naturalny na 1000 ludności (S – stymulanta)
3. KL3: współczynnik salda migracji wewnętrznych i zagranicznych w ‰ (S)
4. KL4: przychodnie na 10 tys. ludności (S)
5. KL5: liczba osób bezrobotnych na 100 osób w wieku produkcyjnym (DS)
6. KL6: pracujący na 1000 osób w wieku produkcyjnym (S)
7. KS1: fundacje, stowarzyszenia, organizacje na 1000 ludności (S)
8. KS2: osoby fizyczne prowadzące działalność gospodarczą na 1000 ludności (S)
9. KS3: udział przedstawicieli władz publicznych, wyższych urzędników, kierowników oraz specjalistów w ogóle radnych (S)
10. KS4: współczynnik skolaryzacji netto (gimnazja) (S)
11. KS5: liczba dodatków mieszkaniowych na 1000 ludności (DS)
12. KM1: udział osób korzystających z instalacji gazowej w ogóle populacji (S)
13. KM2: udział obszarów prawnie chronionych w ogólnej powierzchni gminy (S)
14. KM3: różnica udziałów osób korzystających z wodociągu i z kanalizacji (DS)
15. KM4: przeciętna powierzchnia użytkowa mieszkania na 1 osobę w m^2 (S)
16. KM5: odsetek mieszkań posiadających centralne ogrzewanie (S)
17. KF1: wydatki majątkowe inwestycyjne na 1 mieszkańca w zł (S)
18. KF2: dochody z podatku PIT na 1 mieszkańca w zł (S)
19. KF3: dochody z podatku CIT na 1 mieszkańca w zł (S)
20. KF4: dochody z podatku rolnego na 1 mieszkańca w zł (S)
21. KF5: dochody własne na 1 mieszkańca w zł (S)
22. KF6: podmioty finansowe i ubezpieczeniowe na 10 tys. ludności (S)

23. IT1: spółki handlowe z kapitałem zagranicznym na 1000 podmiotów gospodarczych (S).

Konstrukcja syntetycznego miernika poziomu rozwoju społeczno-gospodarczego

Przy konstrukcji syntetycznego miernika poziomu rozwoju społeczno-gospodarczego wymagane jest by zmienne były porównywalne, głównie dzięki pozbyciu się mian i ujednoliceniu rzędów wielkości. Z uwagi na fakt, że większość zmiennych diagnostycznych cechowała się rozkładem skośnym prawostronnie oraz brakiem rozkładu normalnego dokonano normalizacji zmiennych poprzez przekształcenie ilorazowe w postaci [Walesiak 2016]:

$$z_{ij} = 1 \Big/ \sqrt{\sum_{i=1}^{n} x_{ij}^2} \times x_{ij} \tag{1}$$

gdzie: z_{ij} – znormalizowana wartość zmiennej j dla gminy i (n = 1, 2, …, 2478), x_{ij} – oryginalna wartość zmiennej j dla gminy i.
Następnie wyznaczono syntetyczny miernik poziomu rozwoju społeczno-gospodarczego w oparciu o miarę Braya-Curtisa [Bray, Curtis 1957]:

$$d_{kj}^{BC} = 1 - \frac{\sum_{j=1}^{m} |z_{ij} - z_{kj}|}{\sum_{j=1}^{m} (z_{ij} - z_{kj})} \tag{2}$$

gdzie: z_{ij} – znormalizowana wartość zmiennej j dla gminy i (i = 1, 2, …, 2478), k – gmina „wzorzec” (hipotetyczna gmina, w której wskaźniki przyjmują wartości najbardziej pożądane, tzn. wartości maksymalne dla stymulant i wartości minimalne dla destymulant), j = 1, 2,…, m – numer zmiennej (m = 23).

Miara ta jest pewną modyfikacją metryki Manhattan, ale z uwagi na brak spełnienia kryterium dodatniej wartości odległości między obiektami nie jest metryką. Jednakże dzięki przekształceniu ilorazowemu wyznaczona miara (2) przyjmuje wartości w przedziale od 0 do 1. Tym samym można traktować ją jako unormowaną miarę podobieństwa do wzorca rozwoju społeczno-gospodarczego, gdzie 0 oznacza maksymalne niepodobieństwo, a 1 – maksymalne podobieństwo.

Analiza skupień wg algorytmu *k*-średnich – wstępna klasyfikacja gmin

W celu wydzielenia klas gmin podobnych pod względem poziomu rozwoju społeczno-gospodarczego zastosowano iteracyjną niehierarchiczną metodę grupowania – analizę skupień wg algorytmu *k*-średnich. Uproszczając można przyjąć, że celem metody jest utworzenie k niepustych, rozłącznych i względnie jednorodnych klas (tzw. skupień), w taki sposób, że w każdej iteracji część obiektów jest przenoszona między skupieniami tak aby maksymalizować wariancje międzygrupowe i minimalizować wariancje wewnątrzgrupowe [Hartigan 1975; Kaufman, Rousseeuw 1990]. Przyjęto k = 5 z trzech powodów: (1) w 2004 i 2016 r. zapewnia względnie najwyższy przyrost wariancji międzygrupowej oraz wysoki spadek wariancji wewnątrzgrupowej w relacji do k = 4 i k = 6 (tabela 1), (2) chęci

wydzielenia nieparzystej liczby klas (np.: wysoki, przeciętny, niski poziom rozwoju), (3) zapewnienia odpowiedniej liczności klas z uwagi na zastosowanie w kolejnym etapie metody lasów losowych (np. dla $k = 7$ jedna z klas składała się tylko z 5 gmin). Z uwagi na fakt, że grupowanie odbywało się na podstawie syntetycznego miernika rozwoju, klasy i gminy uszeregowano liniowo na jego podstawie. Utworzono klasy gmin z bardzo wysokim, wysokim, przeciętnym, niskim i bardzo niskim poziomem rozwoju społeczno-gospodarczego. W tabeli 2 przedstawiono odległości euklidesowe między skupieniami, a w tabeli 3 – główne charakterystyki statystyczne wydzielonych skupień.

Tabela 1. Wielkość wariancji wewnątrzgrupowej (WG) i międzygrupowej (MG) dla k liczby skupień

k	2004 r.			2016 r.		
	WG	MG	F	WG	MG	F
3	1,255	3,347	3298,75	1,379	3,878	3478,65
4	0,857	3,745	3603,80	0,914	4,343	3917,23
5	0,611	3,991	4036,66	0,661	4,596	4298,29
6	0,483	4,119	4213,47	0,446	4,811	5331,83
7	0,410	4,192	4209,14	0,352	4,905	5744,20

Źródło: opracowanie własne

Tabela 2. Odległości euklidesowe między skupieniami

2004 r.		kwadrat odległości euklidesowej					2016 r.		kwadrat odległości euklidesowej				
		1	2	3	4	5			1	2	3	4	5
odległość euklidesowa	1	0,000	0,019	0,036	0,051	0,066	odległość euklidesowa	1	0,000	0,003	0,018	0,008	0,015
	2	0,137	0,000	0,003	0,008	0,014		2	0,055	0,000	0,035	0,001	0,005
	3	0,190	0,053	0,000	0,001	0,004		3	0,133	0,188	0,000	0,050	0,066
	4	0,226	0,089	0,035	0,000	0,001		4	0,091	0,037	0,224	0,000	0,001
	5	0,257	0,120	0,066	0,031	0,000		5	0,124	0,069	0,257	0,033	0,000

Objaśnienia do tabeli: numery skupień odpowiadają klasom poziomu rozwoju społeczno-gospodarczego: 1 - bardzo wysoki, 2 - wysoki, 3 - przeciętny, 4 - niski, 5 - bardzo niski

Źródło: opracowanie własne

Tabela 3. Podstawowe charakterystyki statystyczne skupień

klasa	2004 r.					2016 r.				
	n	%	$\bar{x}$	σ	σ^2	n	%	$\bar{x}$	σ	σ^2
1	22	0,9	0,352	0,106	0,011	27	1,1	0,356	0,097	0,009
2	220	8,9	0,215	0,023	0,001	235	9,5	0,223	0,023	0,001
3	603	24,3	0,161	0,012	0,000	594	24,0	0,168	0,013	0,000
4	967	39,0	0,126	0,009	0,000	921	37,2	0,132	0,010	0,000
5	666	26,9	0,095	0,012	0,000	701	28,3	0,099	0,012	0,000

Objaśnienia do tabeli: jak w tabeli 2

Źródło: opracowanie własne

Lasy losowe – weryfikacja klasyfikacji gmin

Wstępną klasyfikację gmin na pięć klas poziomu rozwoju społeczno-gospodarczego poddano weryfikacji przy pomocy metody lasów losowych, która stanowi przykład nadzorowanego uczenia maszynowego. Metoda ta została utworzona przez Leo Breimana [2001] i stanowi próbę eliminacji wad klasycznych drzew klasyfikacyjnych [Breiman i in. 1998]. Jest hybrydą baggingu (Bootrap AGGregatING) i metody losowych podprzestrzeni (Random Subspace Method), która polega na łączeniu wielu drzew klasyfikacyjnych (las) bez przycinania, na wielu losowo dobranych próbach (losowo dobierane obiekty i zmienne) przy jednoczesnym podziale zbioru na zbiór uczący i zbiór testowy, w której ostateczna klasyfikacja powstaje w wyniku „głosowania" (wybór większościowy) zespołu drzew. Rozwiązanie to zapewnia minimalizację błędu modelu przy jednoczesnym utrzymaniu stosunkowo małego jego obciążenia (kwadratu różnicy między wartością oczekiwaną przewidywań modeli dla różnych prób a wartością obserwowaną) – takiego jak dla pojedynczego drzewa oraz stosunkowo niskiej wariancji modelu (poprzez tworzenie drzew w najmniejszym stopniu skorelowanych ze sobą dzięki uczeniu drzew klasyfikacyjnych na próbach losowanych ze zwracaniem oraz przez losowanie pewnej liczby zmiennych objaśniających spośród wszystkich zmiennych przed każdym podziałem w drzewie i tylko na tych zmiennych budowaniu klasyfikacji). Zaletami metody lasów losowych są: (1) odporność na: współliniowość zmiennych, wartości odstające i dużą liczbę zmiennych objaśniających, (2) możliwość odtworzenia złożonych zależności i wykrycia interakcji między zmiennymi, (3) możliwość określenia wskaźników determinujących klasyfikację oraz (4) odporność na „przeuczenie" klasyfikatora [Breiman 2001; Chen i in. 2004; Hastie i in. 2013].

W procedurze jako zmienną objaśniającą przyjęto klasę poziomu rozwoju społeczno-gospodarczego (zmienna nominalna), a zmiennymi objaśniającymi były 23 zmienne diagnostyczne. Prawdopodobieństwo a priori było równe (0,2), a nie szacowane na podstawie wielkści poszczególnych klas zmiennej objaśnianej. Średnia ocena ryzyka błędnych klasyfikacji (każdorazowo dla 200 drzew) dla 2004 i 2016 r. wynosiła odpowiednio – dla zbioru uczącego 0,222 i 0,198, a dla zbioru testowego 0,293 i 0,305 przy błędzie standardowym poniżej 0,017. Ocena jakości klasyfikatora [Powers 2011] z punktu widzenia miar typu czułość (TPR od 0,69 do 0,96), specyficzność (TNR od 0,87 do 0,99), precyzja (PPV od 0,47 do 0,83 i NPV od 0,82 do 1,00) jak i miar zbalansowanych takich jak dokładność (ACC od 0,81 do 0,99), F1score (od 0,63 do 0,84) oraz współczynnika korelacji Matthewsa (MCC od 0,30 do 0,41) dowodzą poprawie jakości klasyfikacji (podobnie jak krzywe ROC i Lift skumulowany nie zamieszczone z uwagi na ograniczoną objętość pracy) względem wyniku otrzymanego z analizy skupień. Jedynie stosunkowo niskie wartości MCC dowodzą na tylko nieco lepsze wyniki klasyfikacji niż w ujęciu losowym.

Zarówno dla 2004 r., jak i 2016 r. w wyniku zastosowania metody lasów losowych zmieniła się liczebność wszystkich klas w stosunku do wyników analizy skupień (tabela 4). Wyraźnie wzrosła liczba (i udział) gmin w klasie bardzo wysokiego i wysokiego poziomu rozwoju (odpowiednio z 22 i 27 do 38 i 55 oraz z 220 i 235 do 260 i 285). Z kolei w klasie niskiego poziomu rozwoju odnotowano spadek liczebności (z 967 i 921 do 862 i 871). W pozostałych przypadkach zmiany były niewielkie. Największy udział niezbieżnych klasyfikacji (ok. 30%) występował w klasach przeciętnego i niskiego poziomu rozwoju, gdzie ok. 12-15% przypadków zostało zaklasyfikowanych do klas sąsiednich. Najmniej niezbieżnych klasyfikacji wystąpiło w klasie bardzo wysokiego (4-9%), bardzo niskiego (po ok. 15%) i wysokiego poziomu rozwoju (17-18%).

Tabela 4. Macierz zbieżności klasyfikacji obserwowanych i przewidywanych

			2004 r.						2016 r.					
			klasa przewidywana (lasy losowe)					Σ	klasa przewidywana (lasy losowe)					Σ
			1	2	3	4	5		1	2	3	4	5	
klasa obserwowana (analiza skupień)	1	n	20	2				22	26	1				27
		%w	90,9	9,1				0,9	96,3	3,7				1,1
		%k	52,6	0,8					47,3	0,4				
	2	n	14	181	22	3		220	26	195	12	2		235
		%w	6,4	82,3	10,0	1,4		8,9	11,1	83,0	5,1	0,9		9,5
		%k	36,9	69,6	3,6	0,4			47,3	68,4	2,2	0,2		
	3	n	4	75	431	93		603	2	89	409	92	2	594
		%w	0,7	12,4	71,5	15,4		24,3	0,3	15,0	68,9	15,5	0,3	24,0
		%k	10,5	28,8	69,6	10,8			3,6	31,2	75,5	10,6	0,3	
	4	n		2	158	675	132	967	1		121	676	123	921
		%w		0,2	16,3	69,8	13,7	39,0	0,1		13,1	73,4	13,4	37,2
		%k		0,8	25,5	78,3	18,9		1,8		22,3	77,6	17,0	
	5	n			8	91	567	666				101	600	701
		%w			1,2	13,7	85,1	26,9				14,4	85,6	28,3
		%k			1,3	10,5	81,1					11,6	82,7	
	Σ	n	38	260	619	862	699	2478	55	285	542	871	725	2478
		%	1,5	10,5	25,0	34,8	28,2	100,0	2,2	11,5	21,9	35,1	29,3	100,0

Objaśnienia do tabeli: jak w tabeli 1 oraz *n* - liczba gmin, %w - udział w wierszu (klasa obserwowana), %k - udział w kolumnie (klasa przewidywana)

Źródło: opracowanie własne

WYNIKI I DYSKUSJA

Przeprowadzona klasyfikacja gmin na skali poziomu rozwoju społeczno-gospodarczego w obu badanych latach (tabela 5, rysunek 1) prowadzi do następujących wniosków: (1) nastąpił wzrost udziału gmin z bardzo wysokim i wysokim poziomem rozwoju z 12,0 do 13,7% przy jednoczesnym wzroście udziału gmin z niskim i bardzo niskim poziomem rozwoju z 63,0 do 64,4% – co

może świadczyć o postępującej polaryzacji i procesie dywergencji rozwoju społeczno-gospodarczego w skali kraju, (2) nastąpił wzrost udziału gmin z wysokim i bardzo wysokim poziomem rozwoju jedynie w woj. wielkopolskim, małopolskim, mazowieckim oraz dolnośląskim, pomorskim i łódzkim, a więc regionach z dużymi aglomeracjami miejskimi (z wyjątkiem woj. śląskiego), przyrost tego typu gmin następuje głównie w najbliższym sąsiedztwie ośrodków wojewódzkich, co może świadczyć o dyfuzji przestrzennej pozytywnych efektów rozwojowych na najbliższe ich otoczenie, (3) utrzymał się bardzo wysoki udział (ok. 75-80%) gmin z niskim i bardzo niskim poziomem rozwoju w woj. lubelskim, podlaskim, świętokrzyskim, podkarpackim i warmińsko-mazurskim (w tym ostatnim znacznie wzrósł udział gmin z bardzo niskim poziomem rozwoju), (4) pogorszyła się sytuacja woj. lubuskiego, dolnośląskiego i zachodniopomorskiego, w których w skali kraju przybyło relatywnie najwięcej gmin z niskim i bardzo niskim poziomem rozwoju – odpowiednio 13,4 pp, 8,3 pp i 3,5 pp.

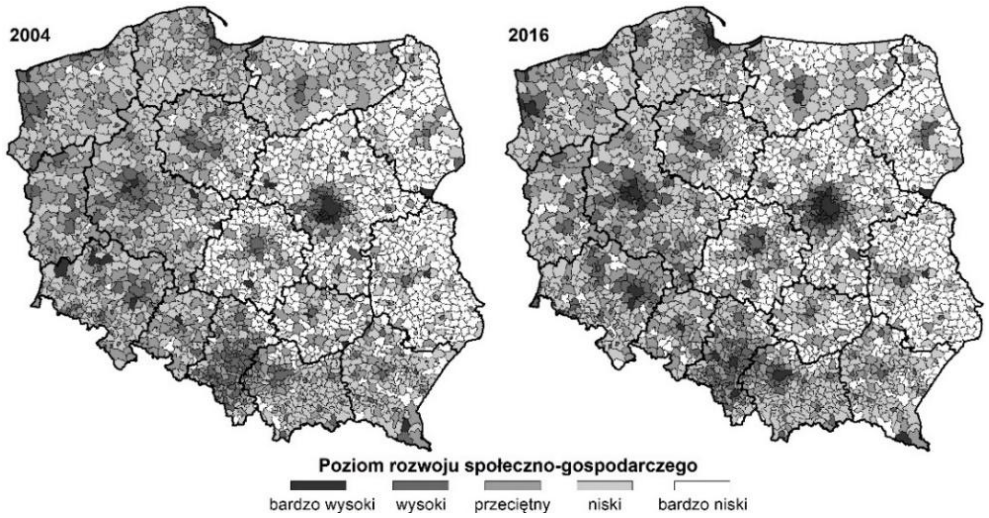
Tabela 5. Poziom rozwoju społeczno-gospodarczego gmin w Polsce w układzie województw (odsetek gmin z określonym poziomem rozwoju)

województwo	liczba gmin	2004					2016				
		BW	W	P	N	BN	BW	W	P	N	BN
dolnośląskie	169	3,0	14,8	42,0	39,6	0,6	3,0	18,3	30,2	42,6	5,9
kujawsko-pomorskie	144	0,0	8,3	22,9	47,2	21,5	0,7	8,3	20,8	39,6	30,6
lubelskie	213	0,5	1,9	13,1	13,6	70,9	0,5	3,3	11,3	16,0	69,0
lubuskie	82	1,2	11,0	45,1	41,5	1,2	0,0	9,8	34,1	46,3	9,8
łódzkie	177	0,6	7,3	16,4	18,1	57,6	1,7	7,3	16,9	31,6	42,4
małopolskie	182	0,0	8,8	28,0	51,1	12,1	0,5	13,2	26,9	48,9	10,4
mazowieckie	314	3,8	12,1	13,7	17,2	53,2	5,4	14,6	10,8	24,2	44,9
opolskie	71	1,4	2,8	42,3	47,9	5,6	1,4	4,2	38,0	40,8	15,5
podkarpackie	160	0,6	8,8	16,9	55,0	18,8	0,6	6,3	16,3	53,8	23,1
podlaskie	118	0,8	5,1	13,6	13,6	66,9	0,8	5,9	10,2	15,3	67,8
pomorskie	123	2,4	13,8	26,8	54,5	2,4	4,9	13,8	24,4	48,8	8,1
śląskie	167	0,6	28,7	41,3	19,2	10,2	1,8	25,1	42,5	24,6	6,0
świętokrzyskie	102	2,0	2,9	14,7	30,4	50,0	0,0	2,9	15,7	26,5	54,9
warmińsko-mazurskie	116	0,0	7,8	25,0	52,6	14,7	0,9	3,4	25,0	29,3	41,4
wielkopolskie	226	1,8	12,8	31,4	44,7	9,3	3,1	19,5	23,5	46,0	8,0
zachodniopomorskie	114	4,4	13,2	32,5	48,2	1,8	6,1	12,3	28,1	43,9	9,6
Polska	2478	1,5	10,5	25,0	34,8	28,2	2,2	11,5	21,9	35,1	29,3

Objaśnienia do tabeli: poziom rozwoju społeczno-gospodarczego: BW – bardzo wysoki, W – wysoki, P – przeciętny, N – niski, BN – bardzo niski

Źródło: opracowanie własne

Rysunek 1. Klasyfikacja gmin na skali poziomu rozwoju społeczno-gospodarczego



Źródło: opracowanie własne

PODSUMOWANIE

Zastosowany algorytm klasyfikacji gmin na skali poziomu rozwoju społeczno-gospodarczego bazujący na wynikach analizy skupień i lasów losowych jak dotąd nie był stosowany na gruncie badań przestrzenno-ekonomicznych i stanowi pewną nową propozycję w tym zakresie. Zwłaszcza zastosowanie metody lasów losowych wykorzystującej wielopoziomową i wieloaspektową losowość i mnogość drzew sprawia, że uzyskana klasyfikacja może być przyjęta z wysokim prawdopodobieństwem i pozwala uzyskać wysoką jej jakość. Przyszłe badania powinny zmierzać do doskonalenia procedur klasyfikacyjnych i poszukiwania jeszcze mocniejszych klasyfikatorów, poprawy oceny możliwości ich zastosowania dla zróżnicowanych wielkościowo i ilościowo jednostek przestrzennych oraz możliwości zastosowania wybranych wskaźników społeczno-gospodarczych w tych procedurach. Uzyskany przestrzenny rozkład klas poziomu rozwoju społeczno-gospodarczego wskazuje na utrzymujące się różnice w poziomie rozwoju między gminami województw Polski wschodniej i zachodniej w postaci tzw. granic reliktowych [Sobczyński 2008; Jańczak 2015]. Niemniej jednak w związku z postępującą polaryzacją wewnątrzregionalną, także w regionach Polski zachodniej coraz liczniej pojawiają się gminy o niskim i bardzo niskim poziomie rozwoju. Wyraźniej w przestrzeni Polski zarysowują się układy gmin o wysokim i bardzo wysokim poziomie rozwoju w obrębie największych aglomeracji miejskich oraz skupienia gmin o niskim i bardzo niskim poziomie rozwoju w obszarach peryferyjnych województw. Sytuacja ta jest niekorzystna zwłaszcza z punktu widzenia oceny efektywności prowadzonej od 15 lat polityki

spójności. Postępująca polaryzacja przestrzenna poziomu rozwoju społeczno-gospodarczego, wbrew głoszonym poglądom politycznym [Mój region, moja Europa… 2017] i pomimo wykorzystania znacznych sum środków unijnych powinna skłaniać do zastanowienia się nad przyszłym modelem polityki rozwoju i stanowić przyczynek do dalszych badań w tym zakresie.

BIBLIOGRAFIA

Bray J. R., Curtis J. T. (1957) An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs, 27(4), 325-349.

Breiman L. (2001) Random Forests. Machine Learning, 45, 5-32.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1998) Classification and Regression Trees (CART). Chapman and Hall, CRC Press reprint.

Burton I. (1963) The Quantitative Revolution and Theoretical Geography. Canadian Geographer, 7(4), 151-162.

Chen C., Liaw A., Breiman L. (2004) Using Random Forest to Learn Imbalanced Data. Technical Report 666, Statistics Department, University of California at Berkeley.

Chojnicki Z. (1985) Orientacje filozoficzno-metodologiczne geografii – ich koncepcje i modele. Przegląd Geograficzny, 57(3), 255-281.

Chojnicki Z., Czyż T. (1973) Metody taksonomii numerycznej w regionalizacji geograficznej. PWN, Warszawa.

Duranton G., Henderson J. V., Strange W. C. (Eds.) (2015) Handbook of Regional and Urban Economics. 5A/5B, North-Holland.

Hastie T., Tibshirani R., Friedman J. (2013) The Elements of Statistical Learning Data Mining, Inference and Prediction. Second Edition. Springer.

Hartigan J. (1975) Clustering Algorithms. John Wiley & Sons.

Jańczak J. (2015) Phantom Borders and Electoral Behaviour in Poland. Historical Legacies, Political Culture and Their Influence on Contemporary Politics. Erdkunde, 69(2), 125-137.

Kaufman L., Rousseeuw P. J. (1990) Finding Groups in Data. An Introduction to Cluster Analysis. Wiley-Interscience.

Mój region, moja Europa, nasza przyszłość: siódme sprawozdanie w sprawie spójności gospodarczej, społecznej i terytorialnej (2017) Bruksela, 9.10.2017, COM(2017) 583.

Nijkamp P. (Ed.) (1986) Handbook of Regional and Urban Economics. Volume I. Regional Economics. North-Holland.

Powers D. M. W. (2011) Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

Sobczyński M. (2008) Polskie doświadczenia w zakresie badania granic reliktowych i krajobrazu pogranicza. [w:] Kulesza M. (red.) Czas i przestrzeń w naukach geograficznych: wybrane problemy geografii historycznej. Wyd. UŁ, Łódź, 66-78.

Walesiak M. (2016) Wybór grup metod normalizacji wartości zmiennych w skalowaniu wielowymiarowym. Przegląd Statystyczny, 63(1), 7-18.

Zeliaś A. (2002) Uwagi na temat wyboru metody normowania zmiennych diagnostycznych. [w:] Kufel T., Piłatowska M. (red.) Analiza szeregów czasowych na początku XXI wieku. Wyd. UMK, Toruń.

USING CLUSTER ANALYSIS AND TECHNIQUE OF RANDOM FORESTS IN THE CLASSIFICATION OF COMMUNES IN POLAND ON THE SCALE OF SOCIO-ECONOMIC DEVELOPMENT

Abstract: The article presents the algorithm of classification of communes on the scale of socio-economic development level. The algorithm includes four steps: (1) selection and reduction of variables, (2) construction of a synthetic measure and linear ordering of communes on the scale of socio-economic development level, (3) grouping of communes by cluster analysis (k-means algorithm) based on the synthetic measure, (4) verification of classification using the random forests method. As a result of the classification procedure was identified the progressive divergence of socio-economic development in Poland.

Keywords: cluster analysis, random forests, classification, communes, socio-economic development