Stanisław Matusik
Akademia Wychowania Fizycznego w Krakowie,
Zakład Statystyki i Informatyki

Michał Bernard Pietrzak
Uniwersytet Mikołaja Kopernika w Toruniu,
Katedra Ekonometrii i Statystyki

Justyna Wilk
Uniwersytet Ekonomiczny we Wrocławiu,
Katedra Ekonometrii i Informatyki

EKONOMICZNO-SPOŁECZNE UWARUNKOWANIA MIGRACJI WEWNĘTRZNYCH W POLSCE W ŚWIETLE METODY DRZEW KLASYFIKACYJNYCH

WSTĘP

W pracy zaproponowano statystyczne metody analizy migracji, które dotychczas są rzadko wykorzystywane w badaniach wewnętrznych wędrówek ludności. Jest to metoda drzew klasyfikacyjnych zastosowana do analizy wewnętrznego ruchu wędrówkowego w Polsce w latach 2008–2010. Rozpatrywane były: migracje międzyregionalne, a dokładnie: napływ i odpływ ludności oraz saldo migracji. Wyniki badań nad przyczynami migracji wskazują, że głównymi powodami wędrówek jest poszukiwanie lepszych warunków rozwoju, które najczęściej stwarzają ośrodki miejskie (zob. m.in. GUS 1928, Lowry 1966, GUS 1974, Bartel 1979, Gawryszewski 2005, Okólski 2005). Znalazło to swój wyraz w fundamentalnej teorii migracji sformułowanej przez Ravensteina (1885) oraz we współczesnych teoriach migracji. Przemiany ustrojowe w Polsce lat 90. doprowadziły do zmian struktury zatrudnienia, a także pojawienia się bezrobocia. Najdotkliwiej było ono odczuwalne w północnych i wschodnich województwach, przy czym jego poziom był mocno zróżnicowany[1].

[1] Zjawisko bezrobocia można prześledzić w zależności od wielkości jednostek terytorialnych: województw (Tokarski 2010), powiatów (Kostrzewska i Pawełek 2007, Kopczewska 2010, Müller-Frączek i Pietrzak 2011) i gmin (Matusik 2006, Śleszyński 2007, Matusik 2008, Matusik 2011).

Ówczesny rozwój gospodarczy w miastach nie wywołał jednak zwiększenia fali migracji wewnętrznych, które – przynajmniej w świetle oficjalnej rejestracji – obniżyły się o około 50% w odniesieniu do słabszych pod tym względem lat 80. (m.in. Słodczyk 2000, Gawryszewski 2005, Sasin 2011). Ujawniło się wówczas zjawisko suburbanizacji, polegające na dekoncentracji funkcji i rozszerzeniu terytoriów miejskich, przy równoczesnym przemieszczaniu się części ludności na ich obrzeża (Szymańska i Hołowiecka 2000, Okólski 2002, Matusik 2005b, Zborowski 2005, Zborowski i Soja 2009, Burdziak 2011, Korcelli 2011). Następnie ludność miejska zaczęła przenosić się na obszary wiejskie zlokalizowane w otoczeniu dużych miast. Przeprowadzone przez Główny Urząd Statystyczny badania oparte na dodatkowych informacjach zebranych w ramach Narodowego Spisu Powszechnego Ludności i Mieszkań z dnia 20 maja 2002 roku (GUS 2004: 35–48) wskazują, że głównymi powodami migracji były kwestie mieszkaniowe, praca zawodowa i sprawy rodzinne.

Przeprowadzone dotychczas analizy przyczyn migracji wewnętrznych wskazują na ich różnorodność, różnokierunkowość, zależności terytorialne i uwarunkowania wynikające z osobistych motywacji. W gospodarce rynkowej, w okresie wzrostu i rozwoju gospodarczego oraz względnej stabilności politycznej, większego znaczenia nabierają przesłanki o charakterze ekonomicznym (m.in. Krugman 1997, Okólski 2002, Korcelli 2011, Pietrzak i in. 2012). Wyjaśnienie w pełni mechanizmów migracji jest zagadnieniem złożonym i przekracza ramy niniejszego artykułu[2]. Wprawdzie ma on przede wszystkim charakter metodyczny, ale zastosowanie metody drzew klasyfikacyjnych pozwoliło także wykazać, że migracje pomiędzy regionami były silnie uwarunkowane ekonomicznie.

Celem pracy było zbadanie oddziaływania uwarunkowań ekonomiczno-społecznych na poziom migracji wewnętrznych w Polsce (napływu i odpływu ludności oraz salda migracji), oraz wskazanie obszarów podobnych pod względem tendencji obserwowanych w ruchach migracyjnych w 66 podregionach Polski. Podjęto próbę określenia przydatności i efektywności wybranej metody w modelowaniu zjawiska migracji wewnętrznych dla podregionów Polski ze wskazaniem głównych ośrodków napływu migracyjnego oraz obszarów odpływu ludności. Przyjęcie podziału na 66 podregionów było uzasadnione z jednej strony wymogami metody, a z drugiej dostępnością danych dla niższych poziomów agregacji przestrzennej.

Uwzględniając postawione cele sformułowano następujące hipotezy badawcze:

1. Zjawiska ekonomiczne w sposób istotny oddziałują na migracje ludności;
2. Podregiony charakteryzujące się wysokim poziomem rozwoju gospodarczego są obszarami równocześnie najsilniejszego napływu i odpływu ludności;

[2] Migracje zagraniczne są jednym z czynników oddziałujących na wielkość migracji wewnętrznych, który nie jest omawiamy w tej pracy; poświęcono jej wiele osobnych opracowań (przykładowo: Korcelli 2000, Kupiszewski 2001, Potrykowska 2009).

3. Obszary najsilniej rozwinięte gospodarczo charakteryzuje dodatnie saldo migracji;
4. Metoda drzew klasyfikacyjnych jest użytecznym narzędziem w wielowymiarowych analizach zjawiska migracji.

DANE STATYSTYCZNE I METODA ANALIZY

Migracje są procesem charakteryzującym się znaczną zmiennością i dlatego wskazane jest rozpatrywanie ich efektów w pewnym okresie (Holzer 2003: 283–297). Wśród problemów, które należy rozważyć podejmując analizę migracji na szczególną uwagę zasługują: sposoby ich pomiaru i jakość będących do dyspozycji danych, czas dokonywania pomiaru, porównywalność wyników, różnice w równomierności zasiedlenia, podział administracyjny i pomiar odległości (Rees i Kupiszewski 1999, Bell i in. 2002)[3].

Przeprowadzona w pracy analiza obejmuje trzyletni okres 2008–2010. Jednym z powodów takiego wyboru było wprowadzenie w 2008 roku w Polsce nowego podziału na 66 podregionów (NUTS 3) w miejsce istniejących wcześniej czterdziestu pięciu. Przedmiotem zainteresowań były migracje na pobyt stały między podregionami. Odpowiednie dane statystyczne zaczerpnięto z Banku Danych Lokalnych Głównego Urzędu Statystycznego. Należy mieć na uwadze, że dane statystyczne są obarczone pewnymi błędami systematycznymi i losowymi (m.in. Kordos 2004, Paradysz 2004, Paradysz 2009, Śleszyński 2004, Śleszyński 2011, Gołata 2012). W tym kontekście, z punktu widzenia tematyki pracy, istotny wpływ mogło mieć: nieuwzględnienie, w statystycznej sprawozdawczości dotyczącej migracji, części osób w wieku najbardziej mobilnym (20–29 lat) w związku z unikaniem przez nich obowiązku meldunkowego, emigracja lub braki w rejestracji zamieszkałych cudzoziemców. Błędy w danych rzutują na końcowe wyniki przeprowadzonych analiz.

Aby zilustrować ideę metody, autorzy ograniczyli się do analizy trzech współczynników dotyczących migracji. Posługując się danymi zagregowanymi dla trzech lat, dla każdego podregionu, w celu zapewnienia porównywalności (ze względu na liczbę ludności), obliczono współczynniki: odpływu ludności (*WO*), napływu ludności (*WN*) oraz salda migracji, czyli przyrostu wędrówkowego (*WS*) w przeliczeniu na 1000 mieszkańców. Współczynniki te potraktowano jako zmienne objaśniane zdefiniowane następująco:

[3] W pracach tych rozważana jest problematyka pomiaru migracji wewnętrznych oraz propozycje wskaźników. Omawiane są także kwestie porównywalności stosowanych wskaźników w ujęciu międzynarodowym (Bell i in. 2002).

$$WO = \frac{1000\,W}{L},\ WN = \frac{1000\,P}{L},\ WS = WN - WO, \tag{1}$$

gdzie: W jest liczbą wyjeżdżających (i równocześnie wymeldowanych) w latach 2008–2010, P jest liczbą przyjeżdżających (zameldowanych) w latach 2008–2010, a L – jest średnią liczbą ludności w analizowanym trzyletnim okresie (Holzer 2003: 274). O ile w analizach porównawczych dotyczących migracji międzynarodowych, na pierwszy plan wysuwają się kwestie sposobu i czasu zebrania informacji oraz podziału terytorialnego (Rees i Kupiszewski 1999), o tyle proponując współczynniki *WO*, *WN* i *WS*, można uznać, że spełniają one wymienione niżej warunki (Bell i in. 2002: 436:437):

1. Jednolitego sposobu pomiaru migracji dla badanych podregionów;
2. Czasowej zgodności pomiaru danych;
3. Niwelowania różnic wynikających z gęstości zaludnienia;
4. Omijania problemów wynikających z różnicy podziałów terytorialnych.

Ze względu na prawostronnie skośny rozkład prawdopodobieństwa tych współczynników (Matusik 2005b), podzielono je na cztery równoliczne podzbiory posługując się kwartylami. Wartości powyżej trzeciego kwartyla (Q3) opisano jako „bardzo duże" i oznaczono na mapach i diagramach numerem 1. Z kolei wartości zawarte między medianą (Q2) a kwartylem trzecim (Q3) określono jako „duże" (numer 2), między kwartylem pierwszym (Q1) a medianą – jako „małe" (numer 3) i poniżej pierwszego kwartyla (Q1) – jako „bardzo małe" (numer 4).

Jako zmienne objaśniające, uznane za potencjalne społeczno-ekonomiczne determinanty wewnętrznych migracji międzyregionalnych, przyjęto 6 wskaźników makroekonomicznych odzwierciedlających poziom rozwoju gospodarczego danego podregionu (Pietrzak i in. 2012, Pietrzak i in. 2013):

– Produkt Krajowy Brutto *per capita* [tysiące zł],
– liczba podmiotów gospodarki narodowej na 100 mieszkańców [szt.],
– nakłady inwestycyjne w przedsiębiorstwach *per capita* [tysiące zł],
– nakłady na środki trwałe *per capita* [tysiące zł],
– stopa bezrobocia rejestrowanego [%][4],
– przeciętne miesięczne wynagrodzenie brutto [setki zł].

Mierniki te stanowią podstawę do zaprezentowania zarówno metody drzew klasyfikacyjnych (Shlien 1990, Loh i Shih 1997, Gatnar 1998, Gatnar 2001), jak też do zilustrowania jej funkcjonalnych i użytecznych właściwości do określenia relacji

[4] Bezrobocie ma znaczące, szczególnie w polskich warunkach, oddziaływanie społeczne, a nie wyłącznie ekonomiczne. Sytuacja taka dotyka bowiem nie tylko samych bezrobotnych, ale oddziałuje również na ich rodziny i na najbliższe otoczenie. Stąd wielkość bezrobocia rejestrowanego można potraktować jako pewnego rodzaju wskaźnik dobrostanu społecznego (Reszke, 1999).

pomiędzy zmiennymi objaśnianymi (współczynnikami napływu i odpływu ludności oraz salda migracji) a zmiennymi społeczno-ekonomicznymi. W analizie przyjęto wartości zmiennych objaśniających z 2008 roku, który jest uznawany za początek odczuwania przez Polskę skutków światowego kryzysu finansowego i gospodarczego.

Metoda drzew klasyfikacyjnych, jako statystyczna metoda eksploracyjna (wykorzystywana w technikach *Data Mining*), używana m.in. w systemach ekspertowych i bazach wiedzy; jest zaliczana do metod sztucznej inteligencji. Polega ona na rekurencyjnym podziale obiektów w *N*-wymiarowej przestrzeni zmiennych objaśniających na rozłączne podzbiory, aż do osiągnięcia ich jednorodności ze względu na ustalony poziom wyróżnionej zmiennej objaśnianej. Następnie dla każdego podzbioru buduje się lokalny model w oparciu o relacje bazujące na niezależnych zmiennych wejściowych[5]. Drzewa klasyfikacyjne, w odróżnieniu od analizy dyskryminacyjnej, nie wymagają spełnienia stosunkowo mocnych założeń, m.in. o wielonormalności rozkładów zmiennych i jednorodności macierzy kowariancji w poszczególnych podgrupach. W przeprowadzonych analizach, jako miarę jakości uzyskanych podziałów, przyjęto współczynnik zróżnicowania Giniego[6]. Miara Giniego, oceniająca zmianę poziomu niejednorodności, przyjmuje wartość zero, gdy dany węzeł jest jednorodny, zatem preferuje zmienne dzielące analizowany zbiór na podzbiory wyraźnie różniące się pod względem badanej zmiennej objaśnianej.

Rodzina metod drzew klasyfikacyjnych należy do procedur nieparametrycznych, tzn. nie zakłada się w nich znajomości postaci rozkładów, ani rodzajów związków pomiędzy zmiennymi. Istotną korzyścią stosowania drzew klasyfikacyjnych jest ich hierarchiczność i elastyczność. W jednowymiarowych drzewach typu binarnego, z każdego węzła wychodzą dwie gałęzie – liście stanowią klasy (zbiory obiektów), zaś gałęzie opisane są prostymi funkcjami reprezentującymi cechy, na podstawie których przeprowadzono podział (relacja: $x < C$, gdzie C jest wyliczoną, na podstawie użytej miary niejednorodności, wartością dyskryminującej cechy diagnostycznej). Spełnienie warunków w określonej kolejności, od korzenia drzewa (znajdującego się u góry grafu) do liścia, jest drogą umożliwiającą prześledzenie i analizę zależności między przynależnością obiektu do danego skupienia, a wartościami objaśniających zmiennych wejściowych.

Jako pozytywną cechę drzew klasyfikacyjnych należy wskazać graficzny, intuicyjny w interpretacji sposób przedstawiania reguł klasyfikacji, nawet dla skomplikowanych modeli (m.in. Matusik 2005a, Matusik 2005b, Matusik 2007). Kolejną zaletą jest uzyskanie rankingu cech oceniającego zdolność dyskryminacyjną w umownej

---

[5] Zmienna objaśniająca (diagnostyczna) może być zarówno zmienną mierzoną na skali porządkowej lub nominalnej, co znacząco zwiększa użyteczność metody drzew klasyfikacyjnych.

[6] Najczęściej używane miary jednorodności zbiorów to współczynnik zróżnicowania Giniego, G-kwadrat albo wartość opata na statystyce χ^2. Można także wykorzystać inne miary. E. Gatnar przedstawia 15 miar jakości jednorodności uzyskanych podzbiorów (Gatnar 2001: 31–49).

skali od 0 do 100, który jest pomocny w ocenie wpływu niezależnych zmiennych klasyfikujących na badane zjawisko.

L. Breiman, J. Friedman, R. Olshen i C. Stone (Breiman i in. 1984)[7] zaproponowali jeden z najbardziej efektywnych algorytmów CART (*Classification and Regression Trees*). Polega on na rozważeniu wszystkich kombinacji poziomów niezależnych zmiennych diagnostycznych w celu znalezienia najlepszego podziału. Podział ten jest wykonywany rekurencyjnie w *N*-wymiarowej przestrzeni obiektów ($N = 66$, liczba badanych podregionów), aż do utworzenia rozłącznych podzbiorów, jednorodnych pod kątem wyróżnionego poziomu zmiennej zależnej.

W niniejszej pracy, metodę CART[8], wyczerpującego poszukiwania jednowymiarowych podziałów z miarą heterogeniczności Giniego, wykorzystano do budowy trzech odrębnych modeli w postaci drzew klasyfikacyjnych. Zmiennymi objaśnianymi były odpowiednio trzy mierniki: współczynnik napływu ludności, współczynnik odpływu ludności oraz saldo migracji. Wartości tych współczynników (będących zmiennymi ciągłymi) zostały zredukowane do czterech kategorii porządkowych (1. bardzo duży, 2. duży, 3. mały, 4. bardzo mały odpowiednio napływ/odpływ) określonych przez kwartyle. W celu oceny skuteczności klasyfikacji, dla kategorii znacząco różniących się liczebnością, dodatkowej analizie poddano współczynniki odpływu podzielone na trzy kategorie wyodrębnione metodą Warda z metryką euklidesową.

WYNIKI BADAŃ

W pierwszej kolejności analizie poddano kształtowanie się współczynników napływów, odpływów i salda migracji. Rozkład gęstości prawdopodobieństwa współczynnika odpływu (*WO)* charakteryzuje silna prawostronna asymetria. Podobny kształt mają rozkłady współczynników *WN* i *WS*, przy czym w małym stopniu zależy on od stopnia agregacji danych (patrz Matusik 2005b: 212). Prawostronny rozkład wskazuje na występowanie stosunkowo nielicznych wysokich wartości, znacząco odbiegających od przeciętnych, co podwyższa wielkość odchylenia standardowego. W tablicy 1 zestawiono podstawowe parametry opisowe rozkładów trzech analizowanych zmiennych objaśnianych.

---

[7] Do innych efektywnych algorytmów można zliczyć QUEST (*Quick Unbiased Efficient Statistical Tree*) pomysłu (Loh i Shih 1997) oraz FACT (*Fast Algorithm for Classification Trees*) autorstwa (Loh i Vanichsetakul 1998).

[8] Algorytm CART jest zaimplementowany m.in. w module *Data Mining* programu *Statistica*, w *SPSS Decission Teres*, w pakiecie procedur *Lumenaut* do MS Excel i w bezpłatnym pakiecie `tree` programu R na licencji GNU GPL.

Tablica 1. Charakterystyki opisowe rozkładu współczynników odpływu ludności *WO* [‰], napływu *WN* [‰] i salda migracji *WS* [‰]

Table 1. Descriptive characteristics of the rates of out-migration (WO), in-migration (WN) and net migration (WS) [‰]

Parametr *Characteristic*	*WO* [‰] *out-migration rate*	*WN* [‰] *in-migration rate*	*WS* [‰] *net migration rate*
Średnia arytmetyczna *Arithmetic mean*	16,04	15,34	–0,51
Odchylenie standardowe *Standard deviation*	5,63	9,84	9,10
Minimum *Minimum*	8,20	6,02	–14,77
Maksimum *Maximum*	42,50	47,33	31,96
Mediana(*Median*	15,33	11,45	–3,46
Współczynnik zmienności *Coefficient of variation*	0,351	0,634	

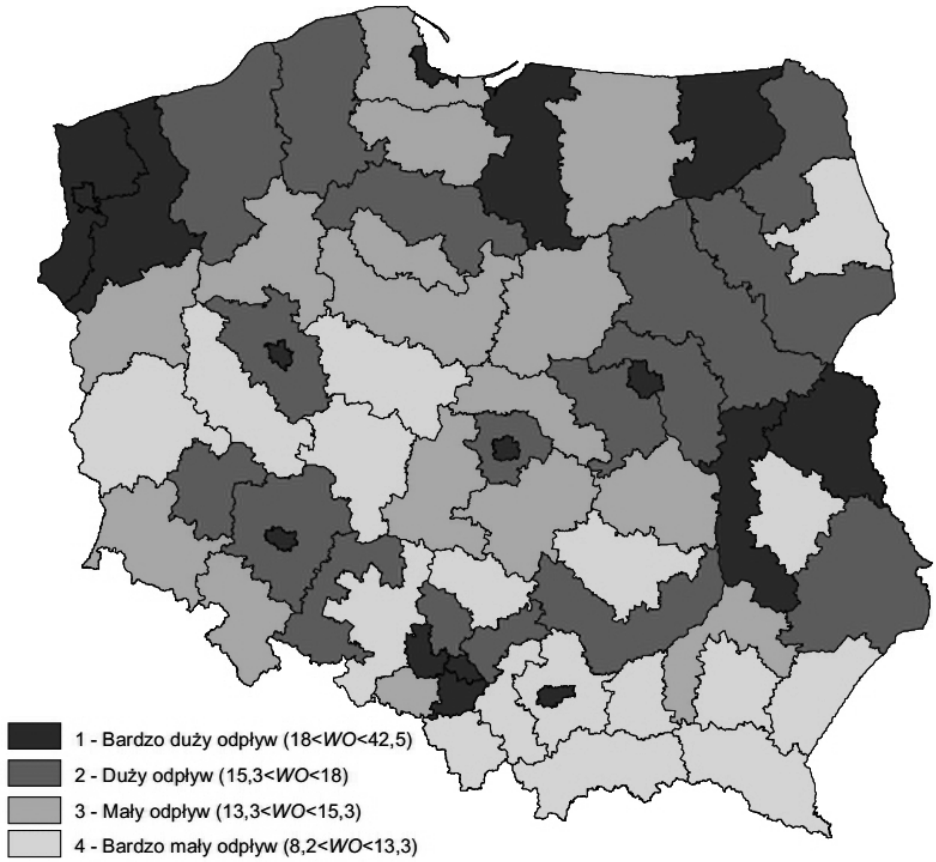
Brak wartości współczynnika zmienności dla WS wynika z bliskiej zeru wartości średniej arytmetycznej.
Missing value of the coefficient of variation for WS results from closed zero value of the arithmetic mean.

Źródło: opracowanie własne na podstawie danych Głównego Urzędu Statystycznego.
Source: Own elaboration, based on the data of the Polish Central Statistical Office.

Przy średniej wartości odpływu ludności (16 na 1000 mieszkańców), największą wartość odnotowano dla miasta Poznania (42,5). Wysokie wartości *WO* miały także podregiony: miasto Wrocław (28,7), trójmiejski (27,7), miasta Warszawa, Kraków, Łódź, Szczecin, podregiony śląskie (gliwicki, katowicki, tyski, bialski), a także podregion puławski. Wymienić można także podregiony położone na północy Polski: elbląski, ełcki, stargardzki i szczeciński. Niewielkimi współczynnikami odpływu ludności ($WO < 13,3$) charakteryzowały się, ogólnie ujmując, rolnicze podregiony Małopolski, Podkarpacia i Wielkopolski (kaliski, koniński, leszczyński) oraz podregion białostocki, lubelski, opolski i zielonogórski. Przestrzenny rokład wielkości międzyregionalnych odpływów wewnętrznych w podregionach przedstawiono na rysunku 1. Gradacja odcieni odpowiada natężeniu odpływów na 1000 mieszkańców: kolor najciemniejszy – największym, a najjaśniejszy – najmniejszym wartościom.

Rysunek 1. Natężenie międzyregionalnych odpływów migracji wewnętrznych w kategoriach współczynnika *WO* [‰]

Figure 1. The intensity of migration outflows from sub-regions by four categories of the out-migration rate WO [‰] (Explanations: 1 – very high outflow, 2 – high outflow, 3 – low outflow, 4 – very low outflow)

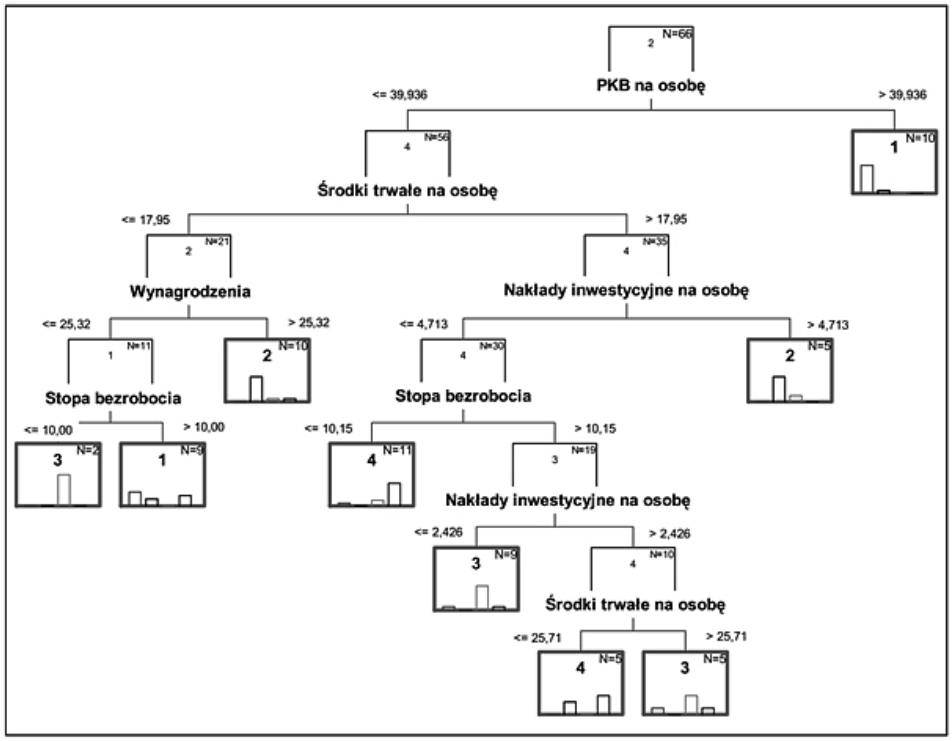


Źródło: opracowanie własne.
Source: Own elaboration.

Modele uzyskane metodą drzew klasyfikacyjnych pozwalają na graficzną prezentację zależności miedzy wartościami zmiennych społeczno-ekonomicznych, a poziomami współczynnika odpływu ludności. Rysunek 2 przedstawia schemat relacji binarnych. Pogrubione kwadraty (liście drzewa) zawierają numery kategorii *WO* oraz liczbę elementów (w prawym górnym rogu), a także histogram ukazujący liczebność elementów analizowanych kategorii. Nad poziomymi liniami zobrazo-

Rysunek 2. Drzewo klasyfikacyjne dla kategorii współczynnika odpływu ludności *WO* (1 – bardzo duży, 2 – duży, 3 – mały, 4 – bardzo mały)

Figure 2. Classification tree for the four categories of out-migration rate WO (Explanations: 1 – very high outflow, 2 – high outflow, 3 – low outflow, 4 – very low outflow)



Variable names: PKB na osobę – GDP *per capita* [1000 PLN], Podmioty – Number of economic entities [*per* 100 inhabitants], Nakłady inwestycyjne na osobę – Investment assets *per capita* [1000 PLN], Środki trwałe na osobę – tangible assets *per capita* [1000 PLN], Stopa bezrobocia – Unemployment rate [%], Wynagrodzenia – Average monthly salary [100 PLN]

Źródło: opracowanie własne.
Source: Own elaboration.

wane są wartości relacji. Pozwalają one prześledzić wpływ wartości diagnostycznych zmiennych społeczno-ekonomicznych na poziom wielkości współczynnika odpływu ludności.

Przykładowo, jak przedstawiono na rysunku 2, PKB na osobę wyższy od 39 936 zł wyodrębnia dziesięć podregionów o najwyższym odpływie ludności (oznaczenie 1). Mniejsze od tej liczby wartości PKB i inwestycje w środki trwałe mniejsze od 17 950 zł, jak też wynagrodzenia powyżej 2532 zł charakteryzują dziesięć podregionów o dużym współczynniku odpływu ludności (*WO* = 2). Anali-

zując dalej lewą gałąź, gdy wynagrodzenia są niższe od 2532 zł, a stopa bezrobocia rejestrowanego większa od 10%, to w tym przypadku model wskazuje na wysokie odpływy ludności w dziewięciu podregionach ($WO = 1$). Przy niższej od 10% stopie bezrobocia, otrzymujemy dwa podregiony o niskim odpływie ludności ($WO = 3$). Zatem wartości czterech zmiennych charakteryzują regiony o najwyższym poziomie odpływów ($WO = 1$): albo wysoki poziom PKB, albo PKB niższe od 39 936 zł, przy równocześnie niskim poziomie inwestycji w środki trwałe, niskim wynagrodzeniu i podwyższonej stopie bezrobocia. Oznacza to, że podregiony te należą zarówno do obszarów gospodarczo wysoko rozwiniętych, jak i relatywnie słabo rozwiniętych, co potwierdza mapa przedstawiona na rysunku 1. Można sformułować przypuszczenie, że z podregionów gospodarczo rozwiniętych, odpływ ludności można w znaczącym stopniu przypisać zjawisku migracji w strefy podmiejskie, a w przypadku migracji z podregionów charakteryzujących się niekorzystną sytuacją gospodarczą odpływ ludności jest wymuszony koniecznością poprawy sytuacji ekonomicznej.

Analizując graf drzewa klasyfikacyjnego od korzenia (u góry ryciny), aż do jego liści (pogrubione kwadraty), możemy dla zmiennych objaśniających ustalić ciągi relacji mniejsze/większe, które prowadzą do określonych poziomów badanej zmiennej objaśnianej. Należy przy tym zauważyć, że do określonego poziomu współczynnika *WO* na ogół prowadzi więcej niż jedna droga poprzez kilka gałęzi opartych na reacjach zmiennych objaśniających. Zmienne te algorytm CART dobiera automatycznie i na przykład dla tego modelu nie uwzględniono przeciętnego wynagrodzenia w podregionach.

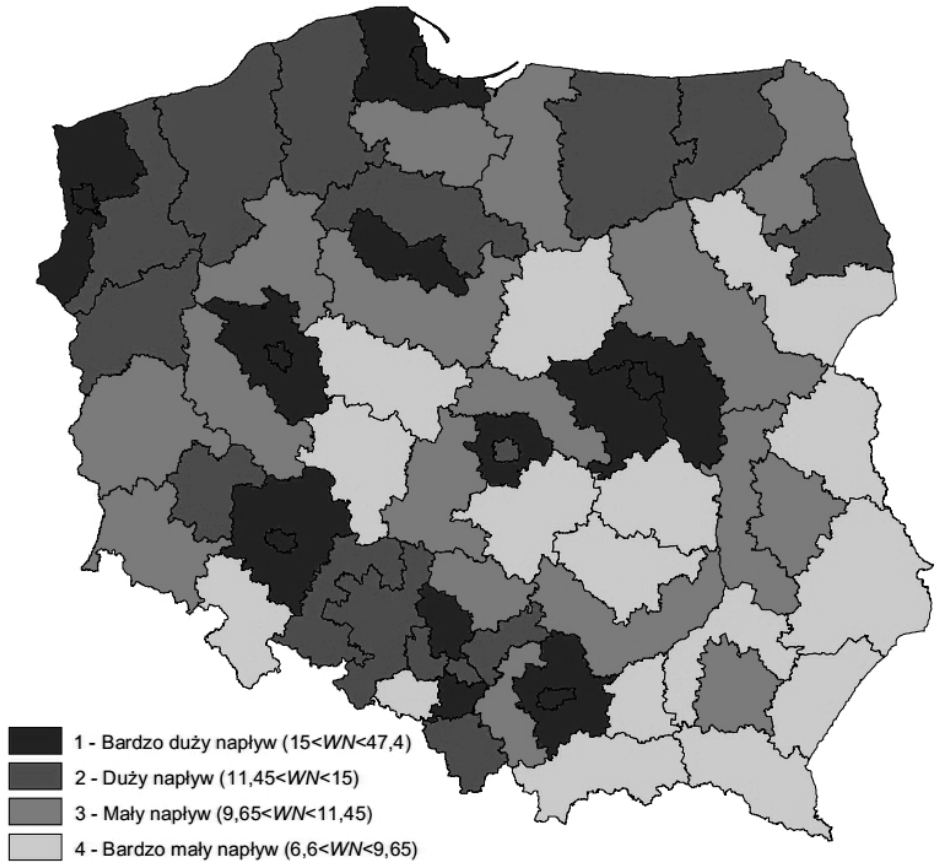
Powstaje zatem pytanie o trafność predykcji poziomu odpływów ludności na podstwie pięciu zmiennych o charakterze społeczno-gospodarczym. Dla najwyższego poziomu $WO = 1$ trafność osiągnęła 81,3%, dla $WO = 2$ była równa 70,6%, dla $WO = 3$ wyniosła 75%, a dla najniższych odpływów międzyregionalnych $WO = 4$ – 64,7%. Ogólna efektywność wskaznia właściwego poziomu *WO* wynosiła 72,7%.

Analogiczną analizę przeprowadzono dla współczynników napływu *WN* [na 1000 mieszkańców]. Charakteryzuje się on blisko dwukrotnie większym rozproszeniem, niż współczynnik odpływu ludności, co świadczy o znacznym zróżnicowaniu.

Rysunek 3 przedstawia poziom współczynnika napływu ludności *WN* w podregionach Polski. Największy napływ ($WN > 15‰$) obserwujemy w podregionach: warszawskim wschodnim, zachodnim i mieście stołecznym Warszawa, mieście Poznań i poznańskim, w gdańskim i trójmiejskim, w podregionach: mieście Kraków i krakowskim, Wrocław i wrocławskim, Szczecin i szczecińskim, bydgosko-toruńskim i w śląskich (bytomskim, tyskim).

Rysunek 3. Natężenie międzyregionalnych wewnętrznych napływów ludności w kategoriach współczynnika *WN* [‰]

Figure 3. The intensity of migration inflows into sub-regions by four categories of the in-migration rate WN [‰] (Explanations: 1 – very high inflow, 2 – high inflow, 3 – low inflow, 4 – very low inflow)



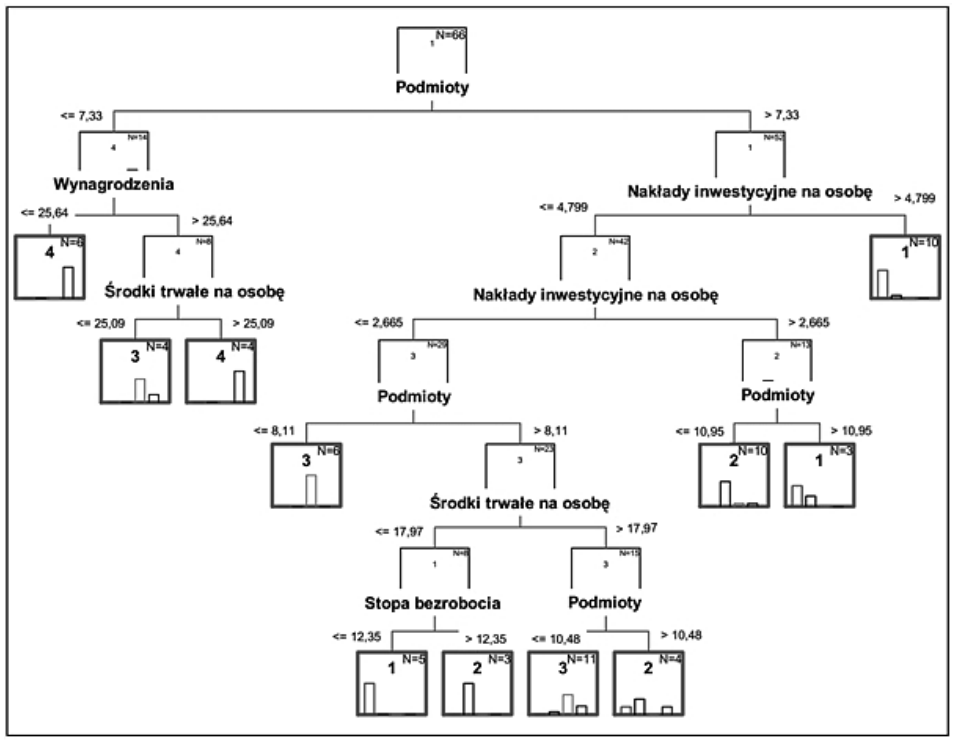
Źródło: opracowanie własne.
Source: Own elaboration.

Najniższy napływ (kategoria 4) odnotowano w podregionach ciechanowsko-płockim, wałbrzyskim, bialskim i rybnickim, podregionach Podkarpacia i Sądecczyzny oraz położonych na Ziemi Świętokrzyskiej, Radomskiej, Piotrkowskiej, Kaliskiej i Konińskiej, a także w Lubelskiem. Współczynnik napływu ludności w tych podregionach nie przekraczał 9,65‰. Z wyjątkiem kilku wymienionych, można przyjąć, że są to podregiony rolnicze o relatywnie wysokiej stopie bezrobocia.

Analiza rysunku 1 oraz rysunku 3 uwidacznia podobieństwo obszarów najwyższych napływów i równocześnie najwyższych odpływów ludności. Potwierdza to wysoka, statystycznie istotna ($p < 0{,}001$) wartość współczynnika korelacji rang Spearmana $\rho = 0{,}514$.

Rysunek 4. Drzewo klasyfikacyjne dla kategorii współczynnika napływu ludności *WN* (1 – bardzo duży, 2 – duży, 3 – mały, 4 – bardzo mały)

Figure 4. Classification tree for the four categories of the in-migration rate WN (Explanations: 1 – very high inflow, 2 – high inflow, 3 – low inflow, 4 – very low inflow)



Variable names: as in Figure 2.

Źródło: opracowanie własne.
Source: Own elaboration.

Przedstawione na rysunku 4 drzewo klasyfikacyjne ma większą głębokość i bardziej skomplikowaną budowę. Niemniej analiza tylko dwóch zmiennych prowadzi do pierwszego liścia z najwyższym poziomem *WN* = 1 (po prawej stronie rysunku): liczba podmiotów gospodarczych przypadających na 100 mieszkańców większa niż 7,33 i równocześnie nakłady inwestycyjne większe od 4799 zł na osobę, wyodrębniają dziesięć spośród siedemnastu podregionów tej grupy. Na kolejne trzy podre-

giony, wskazują wartości nakładów inwestycyjnych mniejsze od kwoty 4799, ale większe od 2665 zł oraz liczba podmiotów gospodarczych większa od 10,95 na 100 osób. Na podstawie tych wyników można sformułować wniosek, że są to podregiony dobrze rozwinięte gospodarczo.

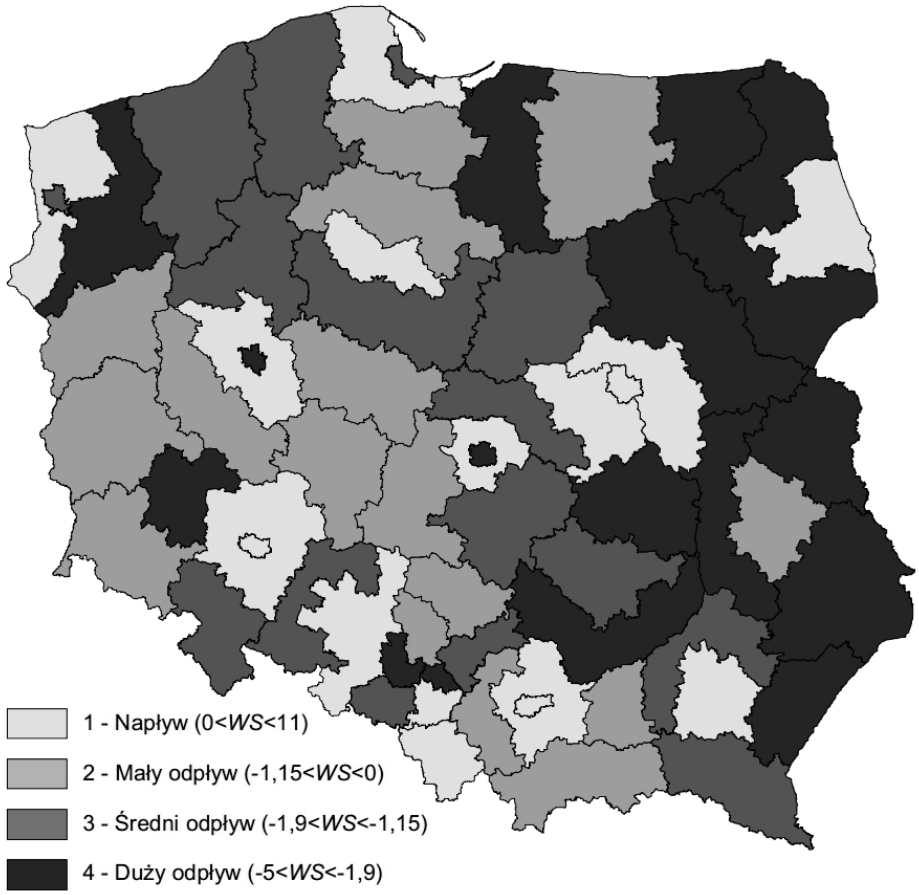
Dla poziomu $WN = 4$ (najniższe napływy), diagram wskazuje gałąź o mniejszym niż 7,33 wskaźniku przedsiębiorczości oraz wynagrodzeniach niższych niż 2564 zł – wskazuje wówczas sześć podregionów, a następne cztery – gdy nakłady na środki trwałe są większe od 25 090 zł. Trafność przewidywania jest zróżnicowana i wynosi: 94,1% dla poziomu $WN = 1$ oraz dla $WN = 3$, 81,3% dla dużych napływów ($WN = 2$) i jedynie 62,5% dla najniższych napływów ($WN = 4$). Ogólnie trafność przewidywania poziomu współczynnika *WN* metodą drzew klasyfikacyjnych była relatywnie wysoka (83,3%), przy wykorzystaniu tylko pięciu zmiennych objaśniających, gdyż algorytm nie uwzględnił zmiennej PKB[9].

Wartości salda migracji (w przeliczeniu na 1000 mieszkańców) przedstawiono na rysunku 5. Obszary charakteryzujące się dodatnim saldem migracji to na ogół podregiony wokół obszarów metropolitalnych: poznański (32,0), gdański (27,9), warszawski zachodni (26,2), warszawski wschodni (21,9), wrocławski (18,7), krakowski (17,7), łódzki (14,8). Pozostałe podregiony w tej grupie to ($0 < WS < 7$): miasto Warszawa, Kraków, Wrocław, podregiony białostocki, bielski, bydgosko-toruński, opolski, rzeszowski, tyski i szczeciński. Należy zauważyć, że duże ujemne saldo migracji (poziom $WS = 4$) odnotowano dla Poznania (-14,8) i Łodzi (-6,0), a średnie ujemne saldo migracji m.in. dla Szczecina, Gdańska oraz w regionie aglomeracji śląskiej. Wyniki tych analiz dowodzą występowaniu suburbanizacji, tzn. przemieszczania się na pobyt stały ludności z aglomeracji miejskich do sąsiadujących z nimi podregionów.

Duży i średni ujemny poziom salda migracji jest charakterystyczny także dla większości podregionów położonych przy wschodniej i północno-wschodniej granicy Polski, w rejonach centralnych oraz w podregionie legnicko-głogowskim i stargardzkim. Z kolei bliskie zeru saldo migracji występuje w trzech podregionach woj. małopolskiego i w trzech podregionach woj. wielkopolskiego oraz przy południowo-zachodniej granicy, w dwóch podregionach woj. śląskiego, a także w podregionach: sieradzkim, lubelskim, grudziądzkim, starogardzkim i olsztyńskim („2 – Mały odpływ" na rysunku 5).

[9] Przy korzystaniu z algorytmu CART wprowadzenie tej zmiennej nie poprawia jakości modelu.

Rysunek 5. Natężenie salda migracji ludności w kategoriach współczynnika *WS* [‰]
Figure 5. The intensity of net migration for sub-regions by four categories of the net migration rate WS [‰] (Explanations: 1 – net inflow; 2 – low net outflow, 3 – medium net outflow, 4 – high net outflow)

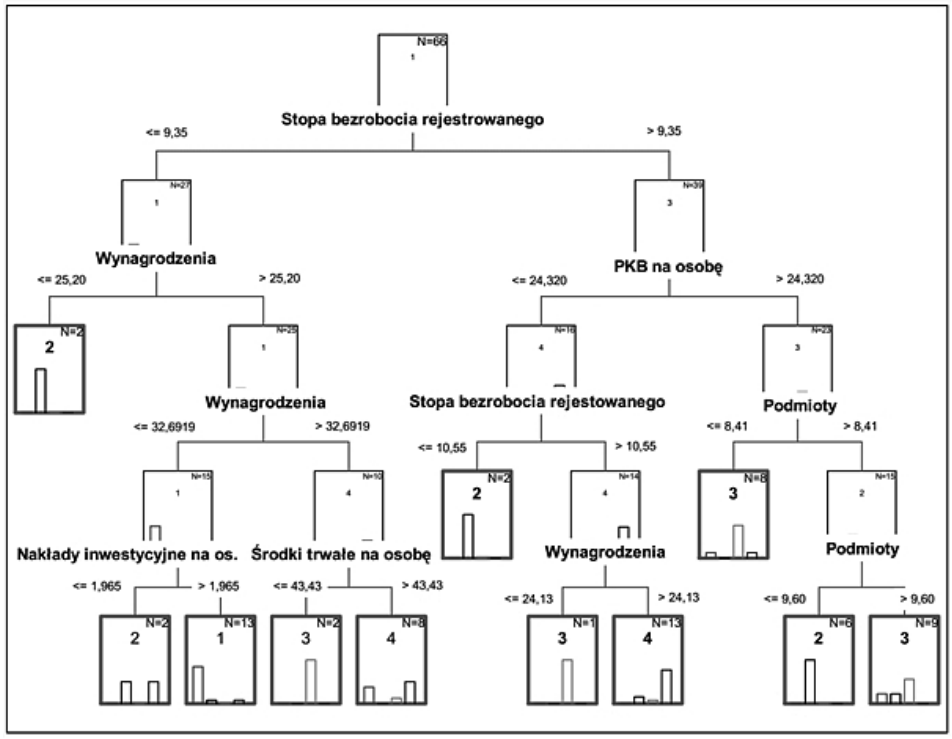


Źródło: opracowanie własne.
Source: Own elaboration.

Tę skomplikowaną sytuację poziomów salda migracji odzwierciedla model drzewa klasyfikacyjnego. Do zbudowania relacji zostały wykorzystane wszystkie zmienne i w przeważającej większości przypadków liście drzewa są zlokalizowane na najniższym poziomie, a nie jak we wcześniejszych modelach, stosunkowo blisko korzenia (rysunek 6). Dotyczy to zarówno dodatniego salda migracji ($WS = 1$), jak i dużego ujemnego salda ($WS = 4$). Interpretacja grafu jest analogiczna, jak dla modeli klasyfikacyjnych współczynnika napływu i współczynnika odpływu ludności.

Rysunek 6. Drzewo klasyfikacyjne dla współczynnika salda migracji *WS* (saldo dodatnie: 1 – napływ, saldo ujemne: 2 – mały odpływ, 3 – średni odpływ, 4 – duży odpływ ludności)

Figure 6. Classification tree for the four categories of the net migration rate WS (Explanations: 1 – net inflow; 2 – low net outflow, 3 – medium net outflow, 4 – high net outflow)



Variable names: as in Figure 2.

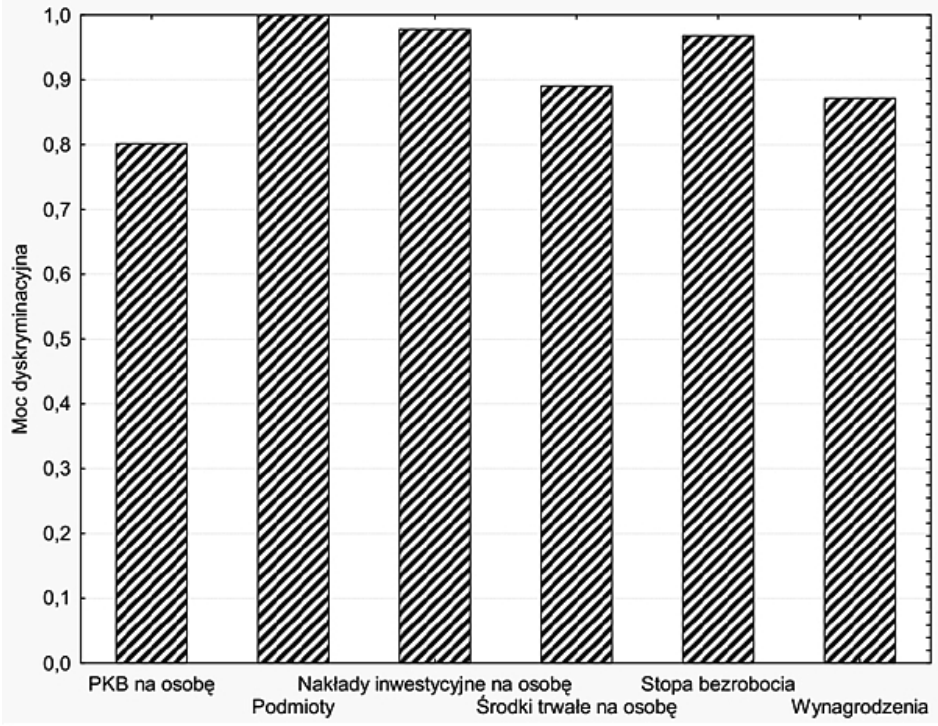
Źródło: opracowanie własne.
Source: Own elaboration.

Mimo relatywnie złożonych zależności, skuteczność przewidywania dodatniego salda migracji (*WS* = 1) wyniosła 64,7%, dla poziomu *WS* = 2 – 68,8%, dla *WS* = 3 – 87,5% i dla dużego salda ujemnego 82,4% (*WS* = 4). Ogólna skuteczność predykcji dla współczynnika salda migracji nieco przekraczała ¾ (75,8%).

Jak wspomniano we wstępie, metoda drzew klasyfikacyjnych dostarcza oceny „mocy dyskryminacyjnej" użytych zmiennych objaśniających. Zilustrowano ją na rysunku 7, na którym wyższe wartości oznaczają większą moc dyskryminacyjną salda migracji.

Rysunek 7. Moc dyskryminacyjna zmiennych objaśniających w modelu drzewa klasyfikacyjnego skonstruowanego dla salda migracji metodą CART

Figure 7. Discriminatory power of predictors in the classification tree model for net migration obtained by using the CART method



Variable names: as in Figure 2.

Źródło: opracowanie własne.
Source: Own elaboration.

Największą mocą dyskryminacyjną w umownej skali 0 do 1 wykazał się wskaźnik przedsiębiorczości, a kolejno nakłady inwestycyjne przypadające na osobę i stopa bezrobocia rejestrowanego. Warto zauważyć, że pozostałe zmienne objaśniające mają również wysokie zdolności dyskryminacyjne, co świadczy o celowości uwzględnienia ich w tego typu analizach. Wartości tych zmiennych trafnie odzwierciedlają poziom rozwoju społeczno-ekonomicznego regionu i są nierozerwalnie związane z poziomem życia jego mieszkańców.

W celu zilustrowania funkcjonowania metody drzew klasyfikacyjnych, w przypadku zbioru danych znacząco różniącego się liczebnością poszczególnych kategorii, przeprowadzono analizę dla współczynnika odpływu migracyjnego *WO*, podzielonego na trzy grupy metodą taksonomiczną Warda, z wykorzystaniem odległości euklidesowej (tablica 2).

Tablica 2. Charakterystyki opisowe rozkładu dla trzech kategorii współczynnika odpływu ludności *WO* [‰]
Table 2. Descriptive characteristics for the three categories of the out-migration rate WO [‰]

Parametr *Parameter*	1. Bardzo wysoki odpływ *1. Very high outflow*	2. Przeciętny odpływ *2. Medium outflow*	3. Bardzo mały odpływ *3. Very low outflow*
N *Number of territorial units*	7	43	16
Średnia arytmetyczna *Arithmetical mean*	28,6	16,0	10,6
Odchylenie standardowe *Standard deviation*	6,46	2,17	1,43
Minimum *Minimum*	23,3	13,2	8,2
Maksimum *Maximum*	42,5	20,9	12,8
Współczynnik zmienności *Coefficient of variation*	22,6%	13,6%	13,5%

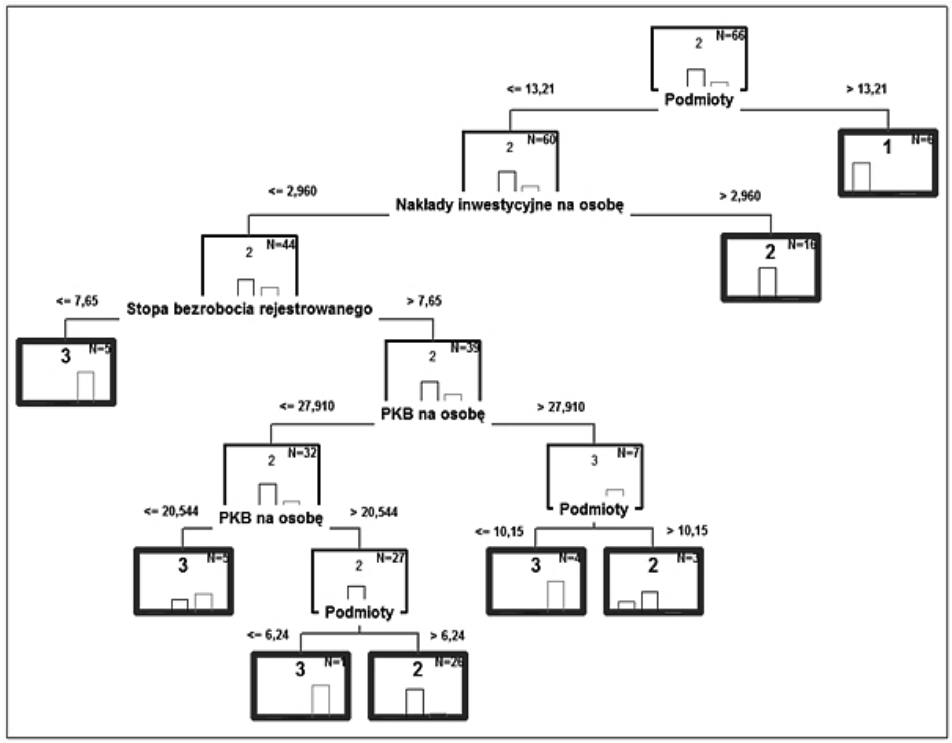
Źródło: opracowanie własne na podstawie danych Głównego Urzędu Statystycznego.
Source: Own elaboration, based on the data of the Polish Central Statistical Office.

Cechą charakterystyczną metody Warda, jest wyznaczanie podziałów taksonomicznych tak, by wartości w danej grupie były do siebie podobne a wydzielone grupy były wyraźnie między sobą zróżnicowane. W efekcie uzyskano możliwie jednorodne kategorie podobnych co do wartości *WO* podregionów, oznaczone jako „1. Bardzo duży odpływ" (7 podregionów; według malejących wartości: miasta Poznań, Wrocław, Szczecin, podregion trójmiejski, Kraków, Warszawa, podregion szczeciński), „2. Przeciętny odpływ" (43 podregiony) oraz „3. Bardzo mały odpływ" (16 podregionów; według malejących wartości: lubelski, przemyski, zielonogórski, oświęcimski, koniński, opolski, krakowski, leszczyński, krośnieński, białostocki, częstochowski, kaliski, nowosądecki, tarnowski, bielski, rzeszowski).

Otrzymane podgrupy regionów różnią się znacząco między sobą nie tylko pod względem przeciętnych wartości współczynników odpływu ludności *WO*, ale także liczebnością (7-16-43). Podobnie jak we wcześniejszych analizach, tak i w tym przypadku, do opisu analizowanych kategorii *WO* została wykorzystana mniejsza liczba zmiennych – tylko cztery spośród sześciu (bez nakładów na środki trwałe *per capita* i bez przeciętnego miesięcznego wynagrodzenia). Jak uwidoczniono na rysunku 8, tylko jedna zmienna – liczba podmiotów gospodarczych przypadających na 100 mieszkańców, której wartości są większe od 13,21 (prawa gałąź drzewa) – wystarcza, aby zaliczyć sześć podregionów do kategorii o najwyższym współczynniku odpływu ludności (*WO* = 1). Jeden podregion został nieprawidłowo zaliczony do innej grupy (tzn. model CART klasyfikuje do grupy „1" tylko sześć, a nie siedem zaliczonych do niej metodą Warda), co daje trafność klasyfikacji 85,7%.

Rysunek 8. Drzewo klasyfikacyjne dla trzech kategorii współczynnika odpływu ludności *WO* [‰] uzyskanych metodą Warda (1– bardzo duży odpływ, 2 – przeciętny odpływ, 3 – bardzo mały odpływ)

Figure 8. Classification tree for the three categories of the out-migration rate WO (Explanations: 1 – very high outflow, 2 – medium outflow, 3 – very low outflow)



Variable names: as in Figure 2.

Źródło: opracowanie własne.
Source: Own elaboration.

Niższa wartość liczby przedsiębiorstw przypadających na 100 mieszkańców, niższe od 2860 zł nakłady inwestycyjne *per capita* oraz relatywnie niska stopa bezrobocia rejestrowanego (nie większa od 7,65%) pozwalają zaliczyć pięć podregionów do kategorii o najniższym odpływie ludności – lewa gałąź drzewa (rysunek 8, *WO* = 3). Analiza następnych relacji, opartych na wielkości PKB *per capita* i na liczbie przedsiębiorstw na 100 osób, prowadzi do klasyfikacji kolejnych dziesięciu podregionów w kategorii *WO* = 3. Trzy podregiony z tej kategorii zostały zaklasyfikowane nieprawidłowo do drugiej kategorii, jak też dwa podregiony z kategorii *WO* = 2 nieprawidłowo sklasyfikowano jako podregiony o najniższym odpływie ludności. W konkluzji skuteczność klasyfikacji dla *WO* = 3 wyniosła 81,3% (13 pra-

widłowo zaklasyfikowanych podregionów), a dla kategorii *WO* = 2 prawie 95,3% (41 prawidłowo sklasyfikowanych podregionów spośród 43).

Wyniki te można ocenić jako bardziej zadowalające niż w przypadku analizy podgrup wydzielonych kwartylami. Zadecydowała o tym metoda podziału zapewniająca większą homogeniczność tworzonych kategorii podregionów. Należy zauważyć, że konkretny podział jest zadany „zewnętrznie" (w tym ostatnim przypadku jest on uzyskany metodą taksonomiczną Warda, zastosowaną dla jednej zmiennej *WO*).

Potwierdzają się w tym badaniu wcześniejsze wnioski, że najsilniejszy odpływ migracji wewnętrznych dotyczy wydzielonych podregionów-miast (z wyjątkiem Łodzi, gdzie *WO* = 20,8 jest najwyższy w kategorii „2" przeciętnych odpływów), natomiast regiony o dużym udziale rolnictwa w gospodarce notują najniższe odpływy migracyjne.

DYSKUSJA I WNIOSKI

We wstępie artykułu przedstawiono pokrótce uwarunkowania migracji wewnętrznych w Polsce, ze wskazaniem na ich różnorodność. Specyfika migracji polega m.in. na tym, że z jednej strony są one dobrowolne, tzn. wynikają z chęci zmiany miejsca zamieszkania, a z drugiej strony mogą być wymuszone panującą sytuacją w otoczeniu czy np. chęcią podniesienia standardu życia albo brakiem środków utrzymania. Istotnymi powodami jest także zwiększenie możliwości rozwoju, poszukiwanie „swojego miejsca na Ziemi", podnoszenie poziomu edukacji. Można więc zauważyć, że na migracje wpływają zarówno okoliczności oraz czynniki zewnętrzne, jak i wewnętrze potrzeby poszczególnych ludzi i całych grup. Dlatego problematyką tą zajmują się demografowie, geografowie, ekonomiści, antropolodzy, socjolodzy i inni.

Czynniki ekonomiczne i społeczno-demograficzne są najczęściej wskazywanymi powodami migracji wewnętrznych. Wymienić tutaj można zwłaszcza uczęszczanie do szkół i uczelni wyższych, małżeństwo, opieka nad rodziną (Pabiańska 2011), cechy socjodemograficzne, takie jak wiek, wykształcenie, czas zamieszkania na danym terenie i zdolności mobilne pracowników (Zdrojewski 2000, Sztanderska 2006). W literaturze wymienia się ponadto czynniki związane z infrastrukturą techniczną i społeczną (Matusik 2005b), co jest ściśle związane z poziomem życia i które mogą być rozpatrywane jako czynniki społeczno-ekonomiczne. Wpływa to na kierunki migracji w układzie miasto-wieś (lub obszary zurbanizowane-obszary peryferyjne), charakteryzujące się różnym poziomem rozwoju społeczno-gospodarczego.

Z przeprowadzonych badań wynika, że przyjęte zmienne objaśniające, mające ekonomiczno-społeczny charakter, pozwalają z dużą trafnością wyjaśniać poziom współczynników opisujących krajowe ruchy migracyjne. Dowodzi to wpływu sytuacji ekonomicznej na wielkość ruchu wędrówkowego i tym samym stanowi podstawę

do uznania, że jedna z postawionych hipotez została zweryfikowana. Podregiony o wysokim napływie ludności są ośrodkami najlepiej rozwiniętymi gospodarczo, o czym świadczą wysokie wartości PKB *per capita*, wysoki poziom inwestycji, wysokie przeciętne wynagrodzenia i niska stopa bezrobocia (Pietrzak i in. 2013). Są to podregiony o metropolitalnym chrakterze, jak Warszawa, Kraków, Wrocław, Poznań, Gdańsk, Szczecin i przylegające do nich podregiony: warszawski wschodni i zachodni, krakowski, wrocławski, poznański, trójmiejski, szczeciński, a także łódzki i bydgosko-toruński oraz bytomski i tyski w alglomeracji śląskiej. Wnioski te są zgodne zarówno z neoklasyczną teorią ekonomii, nową ekonomiczną teorią migracji, jak również z nową geografią ekonomiczną w odniesieniu do regionów (Krugman 1997, Jennissen 2007).

W znaczącej części, te same wysoko rozwinięte miejskie podregiony notują równocześnie wysokie odpływy ludności, co w części związane jest ze zjawiskiem suburbanizacji. Wymienić można pięć wspomnianych podregionów miejskich oraz Łódź i Katowice. Oprócz wysokiego poziomu gospodarczego i społecznego ośrodków wielkomiejskich, na duży napływ i odpływ ludności oddziałuje także fakt zlokalizowania w nich ośrodków akademickich i szkół, znaczące migracje młodzieży w okresie studiowania i nauki w szkołach średnich[10]. Przedstawione wyniki badań potwierdzają słuszność sformułowanej hipotezy, wskazującej na dobrze rozwinięte podregiony jako obszary najsilniejszego napływu i równocześnie najsilniejszego odpływu ludności, co szczególnie wyraźnie uwidacznia się w przypadku wielkich miast.

Pomimo znaczącego ruchu wędrówkowego, dodatnie saldo migracji odnotowano w stosunkowo niewielu podregionach. Zwiększenie liczby ludności netto i dodatnie saldo migracji odnotowano w trzech podregionach warszawskich, dwóch krakowskich i dwóch wrocławskich, w podregionie bielskim, tyskim i opolskim oraz białostockim, bydgosko-toruńskim i rzeszowskim. Należą również do nich podregiony sąsiadujące z dużymi miastami (Gdańsk, Łódź, Poznań, Szczecin), będące w pewnym sensie ich zapleczem mieszkalnym: podregion trójmiejski, łódzki, poznański szczeciński. Wymienione miasta i podregiony są mocno oddziałującymi na otoczenie ośrodkami nauki, kultury, edukacji, przemysłu oraz siedzibami wielu urzędów administracyjnych[11]. Dodatnie saldo migracji jest zatem związane z wysokim poziomem gospodarczym podregionu bądź ze zjawiskiem suburbanizacji.

Przedstawione wyniki dowodzą także występowaniu na wielu obszarach wschodniej i centralnej Polski oraz woj. świętokrzyskiego i Pomorza Środkowego ujemnego salda migracji. Dotyczy to podregionów o stosunkowo niskich wartościach

[10] W mniejszym stopniu dotyczy to Białegostoku, Kielc, Lublina i Rzeszowa, będących także ośrodkami nauki i edukacji.

[11] W przypadku Bielska i Tych dodatnie saldo migracji można uzasadnić efektem mnożnikowym oddziaływania fabryk samochodów Fiata i Opla na rynek pracy (Domański 2001). Kryzys finansowy i zmniejszenie popytu na auta w Europie oraz w kraju może jednak mieć wpływ na zmianę sytuacji demograficznej w tych podregionach.

wskaźników makroekonomicznych uwzględnionych w przeprowadzonych analizach. Chociaż nie badano kierunków migracji można przypuszczać, że napływ ludności z tych regionów kieruje się (zgodnie z teorią Ravensteina) do wysoko rozwiniętych ośrodków gospodarczych, stosunkowo blisko położonych (Warszawa, Kraków, Wrocław, Poznań, Gdańsk, Szczecin). Należy również odnotować występowanie w Polsce obszarów o stosunkowo niskim ruchu ludności. Należą do nich podregiony Małopolski (tarnowski i nowosądecki) i Wielkopolski (kaliski i koniński).

Mimo złożoności badanego zjawiska, w tym m.in. występowania również pozaekonomicznych uwarunkowań migracji[12], zastosowana metoda drzew klasyfikacyjnych z algorytmem CART wykazała relatywnie wysoką skuteczność w opisie poziomów współczynników wewnętrznego ruchu wędrówkowego. Możliwe okazało się skonstruowanie dla badanego okresu modeli opartych na ciągach relacji binanych zmiennych objaśniających, które pozwoliły na trafne określenie w blisko 73% współczynnika odpływu ludności *WO* (dla trzech jednorodnych podzbiorów aż w 90,9%), w 83% współczynnika napływu *WN* oraz w blisko 76% współczynnika salda migracji. Wyniki te dowodzą również, że oddziaływania czynników ekonomiczno-społecznych są bardziej złożone w przypadku odpływu ludności, niż w przypadku napływu, który jest przeważnie uwarunkowany kwestiami ekonomicznymi.

Jak ukazują modele, określony poziom analizowanych wskaźników migracji jest osiągany przy różnych poziomach wskaźników gospodarczych. Może to stanowić wskazówkę do prowadzenia lokalnej polityki gospodarczej, gdyż wyniki badań dowodzą dywergencji demograficznej podregionów Polski. Obrazują także drenaż zasobów ludzkich na dużych obszarach, skutkujący napływem ludności do wielkich miast i ich otoczenia oraz do centrów gospodarczych.

Równocześnie należy podkreślić, że wybierając podział administracyjny na poziomie podregionu, w celu zilustrowania funkcjonowania metody drzew klasyfikacyjnych, autorzy mają świadomość, że nie wszystkie interesujące zjawiska mogły być uchwycone i przedstawione w pełni satysfakcjonujący sposób. Przykładowo zjawisko suburbanizacji, wskazane zostało tylko w przypadku wydzielonych podregionów-miast: Krakowa, Łodzi, Poznania, Szczecina, Warszawy, Wrocławia, a także Trójmiasta. Ponadto niemożliwe było uchwycenie w przypadku dużych miast (np. Białegostoku, Kielc, Lublina, Olsztyna czy Rzeszowa), napływu i odpływu migracyjnego, gdyż strefa miasta i jego otoczenia w dużej części wyrównuje (znosi) te efekty. W tym kontekście, w przyszłości optymalne i interesujące poznawczo byłoby przeprowadzenie badań na mniejszych geograficznie i bardziej heterogenicznych jednostkach przestrzennych, np. powiatach i gminach, choć poważnym problemem metodycznym jest tutaj możliwość zebrania odpowiednich danych społeczno-ekonomicznych. Równie ciekawa i ważna z badawczego punktu widzenia wydaje się

[12] Można tutaj wskazać np. psychologiczną teorię względnej deprywacji, wskazującą na motywacje wynikające z subiektywnego poczucia dysharmonii między rzeczywistym, a oczekiwanym poziomem życia.

klasyfikacja podregionów przy użyciu metody CART na podstawie wielu miar procesu migracji[13], dzięki której uwidoczniłaby się wielowymiarowość tego zjawiska, częściowo ukazana także w tej pracy.

Niedogodnością wykorzystania metody drzew klasyfikacyjnych jest ograniczona zdolność predykcji, a wielkość i stopień złożoności modelu (drzewa) są w części uzależnione od ustawienia parametrów zatrzymania algorytmu. Spore możliwości wyboru miar jednorodności tworzonych podzbiorów utrudniają niekiedy otrzymanie optymalnego drzewa opisującego reguły klasyfikacyjne. Brakuje także obiektywnych kryteriów oceny „dobroci modelu", który może efektywnie pracować na zbiorze uczącym, a otrzymane reguły klasyfikacyjne nie muszą już być tak skuteczne dla innych danych, w tym niekompletnych bądź umożliwiających zastosowanie wręcz sprzecznych reguł klasyfikacyjnych (Stefanowski i Wilk 2008). Należy również pamiętać, że otrzymane reguły pozwalają na klasyfikację obiektów tylko z pewnym prawdopodobieństwem, a kategoryzacja zmiennych, redukująca skalę przedziałową do skali nominalnej wiąże się z utratą informacji.

Zastosowana metoda analizy, jak większość wielowymiarowych metod statystycznych, opiera się na liczbowych asocjacjach. Biorąc pod uwagę współczesne teorie migracji (np. nową ekonomiczną teorię migracji), powiązania te mogą być traktowane jako związki przyczynowo-skutkowe.

Przedstawiona na przykładzie współczynników migracji metoda drzew klasyfikacyjnych pozwoliła na potwierdzenie sformułowanych w pracy hipotez badawczych. Łącząc zalety różnych metod wielowymiarowej analizy danych, w tym analizy dyskryminacyjnej i analizy skupień, pozwala równocześnie na intuicyjną interpretację wyników. Umożliwia w ten sposób analizę złożonych, wielowymiarowych procesów demograficznych i społeczno-gospodarczych. W niektórych przypadkach uproszczenie ciągłej skali pomiarowej[14] do kilkustopniowej skali porządkowej (w tym przypadku cztero- i trzystopniowej), pozwala na lepsze uchwycenie uwarunkowań analizowanego zjawiska. Równocześnie przedstawione przykłady wskazują na możliwość zastosowania tej metody także w przypadku posługiwania się skalami nominalnymi. Warto także zwrócić uwagę na prostotę interpretacji wyników i brak wymagań, co do postaci rozkładów rozpatrywanych zmiennych, jak też na możliwość analizowania związków nieliniowych bez znajomości ich funkcyjnej postaci. Walorem metody jest również automatyczny dobór cech i przedstawienie reguł, pozwalających na klasyfikację danego obiektu do zadanych przez badacza, uprzednio wydzielonych podzbiorów, określonych w oparciu o zobiektywizowane zasady statystyczne bądź inne, uzasadnione celem badania.

---

[13] Można tu wymienić m.in. mierniki natężenia migracji, odległości migracji, obszarów jej ogniskowania i nierównomierności czy uogólnionych indeksów salda migracji. Bell i in. (2002) proponuje 15 różnych mierników (Tab. 8: 461).

[14] Metoda *CHAID* należąca do kategorii algorytmów drzew klasyfikacyjnych pozwala na analizę zmiennych ciągłych (Gatnar 2001).

LITERATURA

Bartel A.P., 1979, *The Migration Decision: What Role Does Job Mobility Play?*, "The American Economic Review", vol. 69(5): 775–786.

Bell M., Blake M., Boyle P., Duke-Williams O., Rees P., Stillwell J., Hugo G., 2002, *Cross-national comparison of internal migration: issues and measures*, "Journal of The Royal Statistical Society", Series A – Statistics In Society, no. 165 (3), 435–464.

Breiman L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984, *Classification and Regression Trees*, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey CA.

Burdziak A., 2011, *Empiryczna weryfikacja efektów aglomeracji w polskich podregionach przy wykorzystaniu metod panelowych*, [w:] J. Suchecka (red. nauk.), *Ekonometria przestrzenna i regionalne analizy ekonomiczne*, Acta Universitatis Lodziensis, Folia Oeconomica, nr 253, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 41–54.

Domański B., 2001, *Przekształcenia terenów poprzemysłowych w województwach śląskim i małopolskim – prawidłowości i uwarunkowania*, [w:] Z. Zioło (red.), *Problemy przemian struktur przemysłowych w procesie wdrażania reguł gospodarki rynkowej*, „Prace Komisji Geografii i Przemysłu PTG", nr 3, 51–59.

Gatnar E., 1998, *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., 2001, *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Gawryszewski A., 2005, *Ludność Polski w XX wieku*, Instytut Gospodarki i Przestrzennego Zagospodarowania PAN, Warszawa.

Gołata E., 2012, *Spis ludności i prawda*, „Studia Demograficzne", nr 1 (161), 23–55.

GUS, 1928, *Pierwszy Powszechny Spis Rzeczpospolitej Polskiej z dnia 30 września 1921 roku: Miejsce urodzenia. Czas pobytu*, Statystyka Polski, t. XXXVI, z. 2, Główny Urząd Statystyczny, Warszawa.

GUS, 1974, *Przyczyny migracji wewnętrznych w 1974 r. Wyniki badania ankietowego. Tablice wynikowe*, Główny Urząd Statystyczny, Warszawa.

GUS, 2004, *Migracje długookresowe ludności w latach 1989–2002 na podstawie Ankiety Migracyjnej 2002*, Główny Urząd Statystyczny, Warszawa.

Holzer J.Z., 2003, *Demografia*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Jennissen, R., 2007, *Causality Chains in the International Migration Systems Approach*, "Population Research and Policy Review", no 26 (4), 411–436.

Kopczewska K., 2010, *Modele zmian stopy bezrobocia w ujęciu przestrzennym*, „Wiadomości Statystyczne", nr 5, 26–40.

Korcelli P., 2000, *Migration from Poland: recent trends and their evaluation*, "Der Donauraum", vol. 40, nr 1/2, 63–77 .

Korcelli P., 2011, *Obszary metropolitalne a funkcjonalne obszary wiejskie*, [w:] S. Kaczmarek (red. nauk.), *Miasto*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 43–50.

Kordos J., 2004, *Niektóre aspekty jakości w statystyce małych i średnich obszarów*, [w:] A. Zeliaś (red.), *Tradycje i obecne zadania statystyki w Polsce*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.

Kostrzewska J., Pawełek B., 2007, *Analiza rynku pracy w ujęciu terytorialnym*, „Wiadomości Statystyczne", nr 10, 53–65.

Kupiszewski M., 2001, *Demograficzne aspekty wybranych prognoz migracji zagranicznych*, [w:] A. Stępniak (red.), *Swobodny przepływ pracowników w kontekście wejścia Polski do Unii Europejskiej. Zrozumieć negocjacje*, Warszawa: UKIE, s. 73–98.

Krugman P.R., 1997, *Development, Geography, and Econonic Theory*, MIT Press.

Loh W.Y., Shih Y.S., 1997, *Split Selection Methods for Classification Trees*, „Statistica Sinica", nr 7, 815–840.

Loh W.Y., Vanichsetakul N., 1998, *Tree-Structured Classification via Generalized Discriminant Analysis*, „Journal of the American Statistical Association", nr 83, 715–728.

Lowry I.S., 1966, *Migration and Metropolitan Growth: Two Analytical Models*, Chandler Pub. Co., San Francisco.

Matusik S., 2005a, *Modeling a Level of Development in Malopolskie Using Decision Trees*, "International Advances in Economic Research", vol. 11(3), Springer Science+Business Media B.V., 343–344.

Matusik S., 2005b, *Migracje wewnętrzne i zagraniczne w gminach województwa małopolskiego w świetle społeczno-ekonomicznych modeli opartych na drzewach decyzyjnych*, [w:] A. Orłowski (red.), *Metody ilościowe w badaniach ekonomicznych – V*, Wydawnictwo SGGW, Warszawa, 208–222.

Matusik S., 2006, *Bezrobocie w gminach województwa małopolskiego w 2002 r.*, [w:] A. Orłowski (red.), *Metody ilościowe w badaniach ekonomicznych – VI*, Wydawnictwo SGGW, Warszawa, 205–213.

Matusik S., 2007, *Demographic Factors in Economic Development Models for the Malopolskie Communes in 2002*, [w:] W. Welfe, P. Wdowiński (red.), *Modelling Economies in Transition 2006*, AMFET Monographs, Łódź, 153–169.

Matusik S., 2008, *Kształtowanie się stopy bezrobocia w gminach woj. małopolskiego*, „Wiadomości Statystyczne", nr 1, 60–72.

Matusik S., 2011, *Ekonometryczne modele zmian stopy bezrobocia rejestrowanego w Polsce oraz w woj. małopolskim w latach 1997–2009*, [w:] J. Suchecka (red. nauk.), *Ekonometria przestrzenna i regionalne analizy ekonomiczne*, Acta Universitatis Lodziensis, Folia Oeconomica, nr 253, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 199–213.

Müller-Frączek I., Pietrzak M.B., 2011, *Analiza stopy bezrobocia w Polsce z wykorzystaniem przestrzennego modelu MESS*, [w:] J. Suchecka (red. nauk.), *Ekonometria przestrzenna i regionalne analizy ekonomiczne*, Acta Universitatis Lodziensis, Folia Oeconomica, nr 253, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 115–223.

Okólski, M., 2002, *Przemiany ludnościowe w Polsce w perspektywie minionego stulecia*, [w:] M. Marody (red.), *Wymiary życia społecznego. Polska na przełomie XX i XXI wieku*, Wydawnictwo Naukowe „Scholar" Sp. z o.o., Warszawa, 26–68.

Okólski M., 2005, *Demografia*, Wydawnictwo Naukowe „Scholar" Sp. z o.o., Warszawa.

Pabiańska P., 2011, *Czynniki ekonomiczne w analizie zjawiska migracji pomiędzy województwami w Polsce*, [w:] M. Jewczak, A. Żółtaszek (red. nauk.), *Ekonometria przestrzenna i regionalne analizy ekonometryczne*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 159–172.

Paradysz J., 2004, *Zasilanie publicznej statystyki regionalnej za pomocą estymacji dla małych obszarów w perspektywie wykorzystania rejestrów administracyjnych*, „Wiadomości Statystyczne", nr 3, 1–14.

Paradysz J., 2009, *Ocena dobroci estymacji dla małych obszarów*, [w:] E. Gołata (red.), *Metody i źródła pozyskiwania informacji w statystyce publicznej*, Zeszyt Naukowy WIGE nr 128, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań, 17–28.

Pietrzak M., Wilk J., Matusik S., 2013, *Gravity model as a tool for internal migration analysis in Poland in 2004–2010,* [w:] J. Pociecha (red.), *Quantitative Methods for Modelling and Forecasting Economic Processes*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków (w druku).

Pietrzak M., Żurek M., Matusik S., Wilk J., 2012, *Application of Structural Equation Modeling for analysing internal migration phenomena in Poland,* „Przegląd Statystyczny", tom 59, nr 4, 487–503.

Potrykowska A., 2009, *Migracje zagraniczne a polityka rodzinna w Polsce*, [w:] *Migracje zagraniczne a polityka rodzinna*, „Biuletyn RPO" – Materiały nr 66, 17–46.

Ravenstein E.G., 1885, *The laws of migration*, "Journal of the Royal Statistical Society", vol. 46, 167–235.

Rees P., Kupiszewski M., 1999, Internal Migration and Regional Population Dynamics in Europe: a Synthesis, Strasbourg: Council of Europe Publishing.

Reszke I., 1999, *Wobec bezrobocia: opinie i stereotypy*, Wydawnictwo Naukowe Śląsk, Katowice.

Sasin M., 2011, *Główne determinanty migracji stałych w Polsce w latach 2003–2008*, [w:] J. Suchecka (red. nauk.), *Ekonometria przestrzenna i regionalne analizy ekonomiczne*, Acta Universitatis Lodziensis, Folia Oeconomica, nr 253, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, 85–98.

Shlien S., 1990, *Multiple Binary Decision Tree Classifiers*, „The Journal of the Pattern Recognition Society", vol. 23, 757–763.

Słodczyk J., 2000, *Natężenie i kierunki przepływów migracyjnych na Śląsku opolskim*, [w:] D. Szymańska (red.), *Procesy i formy ruchliwości przestrzennej ludności w okresie przemian ustrojowych*, Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń, 135–144.

Stefanowski J., Wilk Sz., 2008, *Selective Pre-processing of Imbalanced Data for Improving Classification Performance*, [in:] I.Y. Song, J. Eder, T. Nguyen (eds.), *Data Warehousing and Knowledge Discovery*, "Lecture Notes in Computer Science", vol. 5182, 283–292.

Sztanderska U., 2006, *Strukturalne źródła lokalnego bezrobocia*, „Olympus", nr 2, 61–71.

Szymańska D., Hołowiecka B., 2000, *Ruch wędrówkowy ludności i jego zasięg oddziaływania na przykładzie miasta Bydgoszczy i Torunia*, [w:] D. Szymańska (red.), *Procesy i formy ruchliwości przestrzennej ludności w okresie przemian ustrojowych*, Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń, 217–226.

Śleszyński P., 2004, *Regionalne różnice pomiędzy liczbą ludności według Narodowego Spisu Powszechnego w 2002 roku i szacowaną na podstawie ewidencji bieżącej*, „Studia Demograficzne", nr 1 (145), 93–103.

Śleszyński P., 2007, *Zmiany liczby bezrobotnych w gminach*, „Wiadomości Statystyczne", nr 2, 55–67.

Śleszyński P., 2011, *Oszacowanie rzeczywistej liczby ludności gmin województwa mazowieckiego z wykorzystaniem danych ZUS*, „Studia Demograficzne", nr 2 (160), 35–57.

Tokarski T., 2010, *Regionalne zróżnicowanie bezrobocia w Polsce*, „Wiadomości Statystyczne", nr 5, 41–55.

Zborowski A., 2005, *Przemiany struktury społeczno-przestrzennej regionu miejskiego w okresie realnego socjalizmu i transformacji ustrojowej (na przykładzie Krakowa)*, Wydawnictwo Instytutu Geografii i Gospodarki Przestrzennej Uniwersytetu Jagiellońskiego, Kraków.

Zborowski A., Soja M., 2009, *Demograficzne uwarunkowania rewitalizacji w miastach polskich*, [w:] A. Zborowski (red.), *Demograficzne i społeczne uwarunkowania rewitalizacji miast w Polsce*, Rewitalizacja Miast Polskich, t. V, Wydawnictwo Instytut Rozwoju Miast, Kraków, 13–60.

Zdrojewski E., 2000, *Regiony o chrakterze napływowym i odpływowym w Polsce*, [w:] D. Szymańska (red.), *Procesy i formy ruchliwości przestrzennej ludności w okresie przemian ustrojowych*, Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń, 159–168.

ECONOMIC AND SOCIAL DETERMINANTS OF INTERNAL MIGRATION IN POLAND IN THE LIGHT OF THE CLASSIFICATION TREES METHOD

ABSTRACT

Abstract. This paper analyses socio-economic determinants of migration flows between Polish sub-regions in the years 2008–2010 by using the classification trees (CART) method. Six explanatory variables are selected to determine migration, including GDP *per capita*, number of economic entities (firms) per 100 inhabitants, investment assets and tangible assets *per capita*, registered unemployment rate, as well as *average monthly salary*. The CART method is then used to build models explaining the classification of migration flows into four quartile-based categories.

The results confirm that the classification of internal migration flows is strongly determined by socio-economic features, in particular the number of economic entities, investment assets *per capita* and the unemployment rate. The suburbanisation from cities to neighbouring sub-regions is clearly demonstrated. Better developed regions, especially the largest Polish cities, have the highest migration outflows as well as inflows, yet retain positive net migration. We argue that the proposed analytical approach enables to determine the multidimensional relationships between explanatory social-economic variables and the migration coefficients under study.

Key words: classification trees method, internal migration, economic and social factors