PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Iwona Bąk, Katarzyna Wawrzyniak
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

DIAGNOZA WYJAZDÓW TURYSTYCZNYCH GOSPODARSTW DOMOWYCH EMERYTÓW I RENCISTÓW W POLSCE Z WYKORZYSTANIEM DRZEW KLASYFIKACYJNEGO I REGRESYJNEGO

Streszczenie: W artykule przedstawiono wyniki badań dotyczące klasyfikacji wyjazdów turystycznych emerytów i rencistów ze względu na rodzaj wyjazdu oraz segmentacji gospodarstw domowych emerytów i rencistów w Polsce ze względu na ich uczestnictwo w ruchu turystycznym. W badaniu uwzględniono indywidualne wyjazdy zrealizowane przez gospodarstwa domowe emerytów i rencistów w 2005 r. Do klasyfikacji wyjazdów turystycznych emerytów i rencistów ze względu na rodzaj wyjazdu wykorzystano drzewa klasyfikacyjne, natomiast do segmentacji gospodarstw domowych wykorzystano drzewa regresyjne.

1. Wstęp

W Polsce, tak jak w wielu krajach europejskich, ma miejsce silna tendencja wzrostowa udziału populacji osób w wieku starszym w ogólnej liczbie ludności. Zwiększająca się liczba osób w tym wieku oraz osób o ograniczonej sprawności (rencistów) będzie ważnym czynnikiem kształtującym koniunkturę na rynku usług turystycznych. W związku z tym niezbędne jest podejmowanie badań w sferze konsumpcji turystycznej, które dostarczą istotnych informacji o motywach i zachowaniach konsumpcyjnych emerytów i rencistów oraz szacunkowych wielkościach środków, które mogą być przez nich wydatkowane na wypoczynek.

W artykule sformułowano dwa cele badawcze. Pierwszy z nich dotyczy klasyfikacji wyjazdów turystycznych emerytów i rencistów ze względu na rodzaj wyjazdu (krajowy, zagraniczny), a tym samym wskazania tych zmiennych niezależnych (predyktorów), które dzielą próbę na najbardziej homogeniczne klasy pod względem wyjazdów. Natomiast cel drugi to segmentacja gospodarstw domowych emerytów i rencistów w Polsce ze względu na ich uczestnictwo w ruchu turystycznym. Jako narzędzia badawcze wykorzystano drzewa klasyfikacyjne (do klasyfikacji wyjazdów turystycznych) i regresyjne (do segmentacji gospodarstw domowych).

Dane statystyczne na temat turystyki wyjazdowej emerytów i rencistów zaczerpnięto z badań ankietowych „Turystyka i wypoczynek w gospodarstwach do-

mowych" przeprowadzonych przez Główny Urząd Statystyczny w 2005 r. W badaniu uwzględniono indywidualne wyjazdy zrealizowane przez gospodarstwa domowe emerytów i rencistów w 2005 r. Do klasyfikacji wyjazdów turystycznych wykorzystano 777 ankiet, natomiast do segmentacji gospodarstw domowych – 531 ankiet. Różnica w liczbie ankiet wynikała z tego, że dane gospodarstwo domowe mogło zrealizować w ciągu roku więcej niż jeden wyjazd turystyczny.

2. Istota drzew klasyfikacyjnych i regresyjnych

Drzewa klasyfikacyjne i regresyjne zaliczane są do metod statystycznej analizy wielowymiarowej. Znajdują zastosowanie do klasyfikacji obiektów wówczas, gdy [Gatnar, Walesiak 2004, s. 56-59]:

1) w zbiorze badanych zmiennych można wyróżnić zmienną zależną,
2) badane zmienne (zależna i niezależne) mogą być mierzone zarówno na skalach słabych (nominalna, porządkowa), jak i na skalach mocnych (przedziałowa, ilorazowa).

Drzewa klasyfikacyjne i regresyjne są graficzną reprezentacją modelu postaci [Gatnar 2008, s. 37-39]:

$$Y = f(\mathbf{x}_i) = \sum_{k=1}^{K} \alpha_k \boldsymbol{I}(\mathbf{x}_i \in R_k), \tag{1}$$

gdzie: Y – zmienna zależna, R_k ($k = 1, \ldots, K$, K – liczba segmentów) to podprzestrzenie (segmenty) przestrzeni zmiennych objaśniających $\mathbf{X}^L$ ($X_1, X_2, \ldots, X_L$, L – liczba zmiennych objaśniających), $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iL}]$ – obserwacje ze zbioru rozpoznawalnego, α_k – parametry modelu, $\boldsymbol{I}$ – funkcja wskaźnikowa.

Gdy zmienne $X_1, \ldots, X_L$ mają charakter metryczny, to każdy z segmentów R_k jest definiowany przez jego granice w przestrzeni $\mathbf{X}^L$ w następujący sposób:

$$\boldsymbol{I}(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} \boldsymbol{I}(v_{kl}^{(d)} \le x_{il} \le v_{kl}^{(g)}), \tag{2}$$

gdzie wartości $v_{kl}^{(d)}$ i $v_{kl}^{(g)}$ oznaczają odpowiednio górną oraz dolną granicę odcinka w l-tym wymiarze przestrzeni.

Gdy zmienne $X_1, \ldots, X_L$ mają charakter niemetryczny, to podprzestrzeń R_k można zdefiniować jako

$$\boldsymbol{I}(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} \boldsymbol{I}(x_{il} \in B_{kl}), \tag{3}$$

gdzie B_{kl} to podzbiór zbioru kategorii zmiennej X_l, tj. $B_{kl} \subseteq V_l$.

Jeżeli zmienna zależna Y w modelu (1) jest zmienną nominalną, to taki model nazywamy dyskryminacyjnym i reprezentuje go drzewo klasyfikacyjne. Parametry α_k dla tego modelu wyznaczamy jako

$$\alpha_k = \arg\max_j p(C_j / \mathbf{x}_i \in R_k), \tag{4}$$

gdzie $p(C_j/\mathbf{x}_i \in R_k)$ oznacza prawdopodobieństwo *a posteriori*, że obserwacja z segmentu R_k należy do klasy C_j.

Jeżeli zmienna zależna Y w modelu (1) jest mierzona na skalach mocnych, to ten model jest modelem regresji, a jego graficzną postacią jest drzewo regresyjne. Parametry modelu regresji obliczamy według wzoru:

$$\alpha_k = \frac{1}{N(k)} \sum_{\mathbf{x}_i \in R_k} y_i, \tag{5}$$

gdzie: $N(k)$ – liczba obserwacji znajdujących się w segmencie R_k, y_i – wartości przyjmowane przez zmienną zależną w segmencie R_k.

Do oceny jakości podziału przestrzeni zmiennych objaśniających $\mathbf{X}^L$ wykorzystuje się następujące miary[1]:
- dla zmiennej zależnej niemetrycznej: błąd klasyfikacji, wskaźnik Giniego, miarę entropii, statystykę χ^2,
- dla zmiennej zależnej metrycznej – wariancję zmiennej zależnej.

3. Drzewo klasyfikacyjne jako narzędzie klasyfikacji wyjazdów turystycznych

Do klasyfikacji wyjazdów turystycznych emerytów i rencistów ze względu na rodzaj wyjazdu wykorzystano drzewa klasyfikacyjne. Jako zmienną zależną przyjęto rodzaj wyjazdu (krajowy, zagraniczny), natomiast w zbiorze zmiennych niezależnych uwzględniono[2]:

1) predyktory jakościowe, takie jak:
- forma wyjazdu: wczasy, wycieczki (impreza objazdowa, pielgrzymka), inna (rodzina, działka);
- pośrednictwo w zakupie usług turystycznych: korzystał, nie korzystał;
- główny środek transportu wykorzystywany na dojazd: kolej, PKS lub inna autobusowa linia przewozowa, autokar, samochód osobowy, inny (samolot, prom);
- charakter odwiedzanego obszaru: obszar miejski (stolica, aglomeracje miejskie), miejscowość turystyczna, obszary górskie i wyżynne, obszary położone nad wodą (morze, akwen śródlądowy lub ciek wodny), uzdrowisko, obszar wiejski;

2) predyktor ilościowy: roczne wydatki ogółem poniesione w związku z wyjazdem: od 50 zł do 7200 zł.

[1] Sposoby wyznaczania i własności miar wykorzystywanych do oceny jakości podziału przestrzeni zmiennych są szeroko omówione w pracach [Gatnar 2001; Gatnar, Walesiak 2004; Gatnar 2008].

[2] Wszystkie zmienne niezależne zaproponowane do budowy drzew klasyfikacyjnych wykazują statystycznie istotne zależności z rodzajem wyjazdu. Porównaj badanie [Bąk, Wawrzyniak 2009].

Do wyznaczenia drzew klasyfikacyjnych wykorzystano dwie procedury: CART i CHAID oprogramowane w pakiecie *Statistica 8.0*. Przystępując do budowy drzew, przyjęto założenia zaprezentowane w tab. 1.

Wykorzystując wspomniane procedury przy uwzględnieniu założeń z tab. 1, otrzymano drzewa klasyfikacyjne o różnej strukturze. Wyniki zaprezentowano w tab. 2.

Tabela 1. Założenia przyjęte przy wyznaczaniu drzew klasyfikacyjnych

Założenia przyjęte w procedurze	Procedura CART			Procedura CHAID		
	model standardowy	drzewo interakcyjne	drzewo interakcyjne przycięte	standardowy CHAID	interakcyjny CHAID	drzewo interakcyjne przycięte
Koszt błędnej klasyfikacji	równe					
Miary dopasowania (reguła podziału)	wskaźnik Giniego			statystyka χ^2		
Kryterium stopu	przy błędnej klasyfikacji					
Minimalna liczność	30	30	30	30	30	30
Minimalna liczność potomka	-	30	30	-	30	30
Maksymalna liczba poziomów	-	10	10	-	10	10
Maksymalna liczba węzłów	1000	1000	1000	1000	1000	1000

Źródło: opracowanie własne.

Tabela 2. Struktura drzew klasyfikacyjnych otrzymanych z wykorzystaniem procedur CART i CHAID

Struktura drzewa	Procedura CART			Procedura CHAID		
	model standardowy	drzewo interakcyjne	drzewo interakcyjne przycięte	standardowy CHAID	interakcyjny CHAID	drzewo interakcyjne przycięte
Liczba węzłów dzielonych	3	4	3	4	5	3
Kryterium podziału węzłów (od pierwszego do ostatniego węzła)	1) środek transportu 2) obszar 3) forma	1) środek transportu 2) obszar 3) forma 4) obszar	1) środek transportu 2) obszar 3) forma	1) forma 2) obszar 3) pośrednictwo 4) środek transportu	1) forma 2) środek transportu 3) pośrednictwo 4) obszar 5) środek transportu	1) forma 2) środek transportu 3) pośrednictwo
Liczba węzłów końcowych	4	5	4	7	8	5

Źródło: opracowanie własne.

Rys. 1. Drzewo klasyfikacyjne dla rodzaju wyjazdu uzyskane procedurą standardowy model CHAID (w nawiasach podano odsetki obliczone w stosunku do liczebności całej próby)

Źródło: opracowanie własne.

Dla wszystkich wyznaczonych drzew klasyfikacyjnych jakość podziału była podobna (błąd standardowy szacowany za pomocą 10-krotnej walidacji krzyżowej wahał się w granicach od 0,0127 do 0,0134, natomiast ocena ryzyka mieściła się w granicach od 0,132 do 0,151). Dokonując wyboru najlepszego drzewa, zwrócono uwagę na trzy aspekty: wybrane drzewo nie powinno być zbyt mocno rozbudowane, kryteria podziału węzłów nie powinny się powtarzać w kolejnych węzłach dzielonych, powinna istnieć możliwość logicznej interpretacji wyników końcowych. Na tej podstawie za najlepsze uznano drzewo uzyskane procedurą standardowy CHAID (rys. 1, oznaczenia W1-W4 oznaczają węzły dzielone, natomiast L1-L7 – węzły końcowe). Interpretując wybrane drzewo klasyfikacyjne, sformułowano następujące wnioski, wykorzystując w tym celu regułę zdań warunkowych typu „jeżeli…, to…":

- jeżeli formą wyjazdu była wycieczka (impreza objazdowa, pielgrzymka) organizowana za pomocą pośrednika, to wyjazd był wyjazdem zagranicznym; natomiast, gdy nie korzystano z usług pośrednika, to wyjazd był wyjazdem krajowym;
- jeżeli formą wyjazdu były wczasy na obszarze miejskim, a środkiem transportu był autokar, to wyjazd był zagraniczny; natomiast, gdy formą wyjazdu były wczasy organizowane na dowolnym obszarze lub był to wyjazd do rodziny albo na działkę, ale środkiem transportu była kolej lub PKS, to wyjazd był wyjazdem krajowym;
- jeżeli formą wyjazdu były wczasy (sanatorium) organizowane w miejscowości turystycznej lub uzdrowisku, to był to wyjazd krajowy.

Reasumując powyższe wnioski, stwierdzono, że wyjazd krajowy realizowany był głównie z wykorzystaniem kolei, PKS-u lub samochodu, bez korzystania z usług pośrednika, odwiedzanym terenem były głównie miejscowości turystyczne i uzdrowiska, natomiast wyjazd zagraniczny realizowany był głównie z wykorzystaniem autokaru lub samolotu, usług pośrednika, głównie była to wycieczka, a odwiedzanym obszarem był obszar miejski.

4. Segmentacja gospodarstw domowych z wykorzystaniem drzew regresyjnych

Do segmentacji gospodarstw domowych wykorzystano drzewa regresyjne. Zmienną zależną zdefiniowano jako łączne roczne wydatki poniesione przez gospodarstwo domowe na wyjazdy turystyczne (wartości od 50 zł do 7200 zł), natomiast zbiór zmiennych niezależnych tworzyły:

1) predyktory jakościowe: płeć (kobieta, mężczyzna), wykształcenie (bez wykształcenia, podstawowe, zasadnicze zawodowe, średnie, wyższe), rodzaj wyjazdu (krajowy, zagraniczny);

2) predyktory ilościowe: przeciętny miesięczny dochód gospodarstwa (od 350,60 zł do 4334, 39 zł), wiek (od 17 do 91 lat), liczba wyjazdów w ciągu roku (od 1 do 10), liczba osób wyjeżdżających w ciągu roku (od 1 do 19).

Drzewa regresyjne wyznaczono, stosując te same procedury jak w przypadku drzew klasyfikacyjnych. Przyjęte przy obliczeniach założenia prezentuje tab. 3, a strukturę otrzymanych drzew zamieszczono w tab. 4.

Tabela 3. Założenia przyjęte przy wyznaczaniu drzew regresyjnych

Założenia przyjęte w procedurze	Procedura CART		Procedura CHAID	
	model standardowy	drzewo interakcyjne	model standardowy	drzewo interakcyjne
Kryterium stopu	przytnij według wariancji		–	
Minimalna liczność	30	30	30	30
Minimalna liczność potomka	–	30	–	30
Maksymalna liczba poziomów	–	10	10	10
Maksymalna liczba węzłów	1000	1000	–	1000

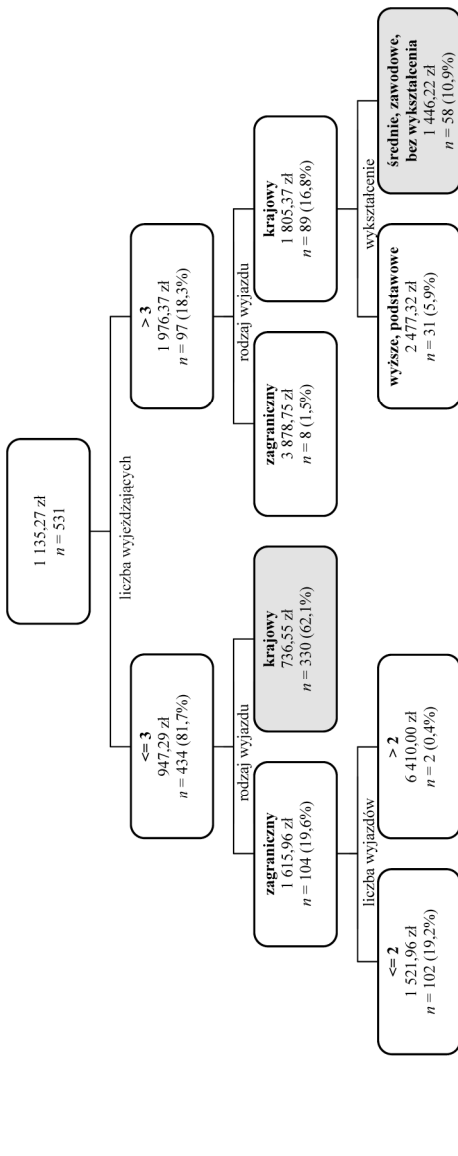
Źródło: opracowanie własne.

Tabela 4. Struktura drzew regresyjnych otrzymanych z wykorzystaniem procedury CART i CHAID

Struktura drzewa	Procedura CART		Procedura CHAID	
	model standardowy	drzewo interakcyjne	model standardowy	drzewo interakcyjne
Liczba węzłów dzielonych	6	5	5	3
Kryterium podziału węzłów (od pierwszego do ostatniego węzła)	1) liczba wyjeżdżających 2) rodzaj wyjazdu 3) rodzaj wyjazdu 4) liczba wyjazdów 5) wykształcenie 6) dochód	1) liczba wyjeżdżających 2) rodzaj wyjazdu 3) dochód 4) liczba wyjeżdżających 5) liczba wyjeżdżających	1) liczba wyjeżdżających 2) rodzaj wyjazdu 3) rodzaj wyjazdu 4) dochód 5) rodzaj wyjazdu	1) liczba wyjeżdżających 2) rodzaj wyjazdu 3) rodzaj wyjazdu
Liczba węzłów końcowych	7	6	8	5

Źródło: opracowanie własne.

Za najlepsze uznano drzewo regresyjne otrzymane z wykorzystaniem procedury CART model standardowy, które przedstawiono na rys. 2. Było to drzewo o numerze 23 wybrane z sekwencji 29 drzew. Drzewo to charakteryzuje się nieznacznie wyższymi wartościami parametrów oceniających ryzyko niż pozostałe drzewa zaprezentowane w tab. 2. Jednak uwzględnia ono istotne z punktu widzenia prowadzonych badań predyktory, dzięki którym możliwe było merytoryczne scha-



Rys. 2. Drzewo regresyjne dla rocznych wydatków ogółem poniesionych w związku z wyjazdem turystycznym w ciągu roku wyznaczone procedurą standardowy model CART (w nawiasach podano odsetki obliczone w stosunku do liczebności całej próby)

Źródło: opracowanie własne.

rakteryzowanie segmentów gospodarstw domowych ze względu na ich uczestnictwo w ruchu turystycznym[3]. Ponadto przy wyborze segmentu przyjęto założenie, że jego liczebność powinna stanowić przynajmniej 10% liczebności próby. Warunek ten spełniają tylko trzy węzły końcowe, które na rys. 2 zostały wyróżnione szarym kolorem. Wyniki segmentacji zamieszczono w tab. 5.

Tabela 5. Charakterystyka segmentów gospodarstw domowych emerytów i rencistów ze względu na ich uczestnictwo w ruchu turystycznym

Nr	Charakterystyka segmentu na podstawie		Przeciętne roczne wydatki na wyjazdy (zł)	Liczebność segmentu (% badanej próby)
	drzewa regresyjnego	predyktorów nieuwzględnionych w drzewie regresyjnym		
1	Gospodarstwa domowe, w których w ciągu roku wyjeżdżają nie więcej niż trzy osoby i wyjazdy turystyczne realizowane są w kraju	Członkowie tych gospodarstw domowych wyjeżdżają głównie do uzdrowisk, rodziny, dominuje obszar nad wodą i górski, główną formą transportu jest kolej	736,55	330 (62,1%)
2	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 1613,89 zł, w których wyjeżdżają w ciągu roku nie więcej niż trzy osoby i w których w ciągu roku realizowane są co najwyżej dwa wyjazdy zagraniczne	Członkowie tych gospodarstw domowych wyjeżdżają głównie na wycieczki, dominuje obszar miejski, główną formą transportu jest autokar	1223,66	82 (15,4%)
3	Gospodarstwa domowe, w których głowa gospodarstwa domowego ma wykształcenie co najwyżej średnie, w których w ciągu roku wyjeżdżają więcej niż trzy osoby i wyjazdy turystyczne realizowane są w kraju	Członkowie tych gospodarstw domowych wyjeżdżają głównie na wczasy, dominuje obszar nad wodą i w górach, a główną formą transportu jest samochód	1446,22	58 (10,9%)

Źródło: opracowanie własne.

5. Podsumowanie

Zastosowane w artykule drzewa klasyfikacyjne umożliwiły wykrycie tych predyktorów, które w sposób istotny dzielą próbę na jednorodne klasy ze względu na rodzaj wyjazdu. Najistotniejszymi predyktorami w tym przypadku okazały się następujące zmienne niezależne: forma wyjazdu, charakter odwiedzanego obszaru, po-

[3] Rezygnując z pozostałych drzew, kierowano się tym, że węzły końcowe w tych drzewach były wyznaczone na podstawie małej liczby predyktorów (2-3), które powtarzały się przy różnych węzłach dzielonych.

średnictwo w zakupie usług turystycznych i środek transportu. Wśród wyjazdów turystycznych realizowanych przez gospodarstwa domowe emerytów i rencistów w 2005 r. dominowały wyjazdy krajowe organizowane bez pośrednika do miejscowości turystycznych lub uzdrowisk, a głównym środkiem transportu była kolej, PKS lub samochód. Natomiast wyjazd zagraniczny był organizowany w formie wycieczki z udziałem pośrednika, środkiem transportu był głównie autokar lub samolot, a odwiedzanym obszarem był obszar miejski.

Wykorzystanie drzew regresyjnych pozwoliło na wydzielenie segmentów gospodarstw domowych, które znacznie różniły się pod względem poziomu przeciętnych rocznych wydatków na wyjazdy turystyczne. Okazało się, że najliczniejszą grupę gospodarstw domowych emerytów i rencistów (ponad 60%) stanowiły gospodarstwa wydające na turystykę przeciętnie w ciągu roku ok. 740 zł i osoby z tych gospodarstw wybierały przede wszystkim wyjazdy krajowe (sanatoria, odwiedziny u krewnych, pielgrzymki), a głównym środkiem transportu była kolej.

Z przeprowadzonych badań wynika również, że istotnym uzupełnieniem wyników uzyskanych za pomocą drzew klasyfikacyjnych i regresyjnych mogą być charakterystyki wyznaczone dla tych predyktorów, które nie były uwzględnione w wybranym do interpretacji drzewie.

Literatura

Bąk I., Wawrzyniak K., *Zastosowanie analizy korespondencji w badaniach zawiązanych z motywami wyboru rodzajów wyjazdów turystycznych przez emerytów i rencistów w 2005 r.*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, UE, Wrocław 2009.

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

Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2008.

Gatnar E., Walesiak M., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław 2004.

THE DIAGNOSIS OF TOURISM OF HOUSEHOLDS OF RETIREES AND PENSIONERS IN POLAND BY MEANS OF CLASSIFICATION AND REGRESSION TREES

Summary: In the article, the authors present results of:

- classification of touristic travels of retirees and pensioners according to the kind of the travel,
- segmentation of households of retirees and pensioners according to their participation in tourist traffic.

In research the statistical data about individual trips of retirees and pensioners households in Poland in 2005 were used. The classification of trips was conducted by means of classification trees and the segmentation of households – by means of regression trees.