METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XIII/3, 2012, str. 7 – 17

WYKORZYSTANIE DRZEW KLASYFIKACYJNYCH I REGRESYJNYCH DO ANALIZY WYJAZDÓW TURYSTYCZNYCH GOSPODARSTW DOMOWYCH SENIORÓW W POLSCE

Iwona Bąk
Katedra Zastosowań Matematyki w Ekonomii
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
e-mail: iwona.bak@zut.edu.pl

Streszczenie: W artykule przedstawiono wyniki badań dotyczące klasyfikacji wyjazdów turystycznych seniorów ze względu na rodzaj wyjazdu oraz segmentacji gospodarstw domowych seniorów w Polsce ze względu na ich uczestnictwo w ruchu turystycznym. W badaniu uwzględniono indywidualne wyjazdy zrealizowane przez gospodarstwa domowe seniorów w okresie 1.10.2008-30.09.2009. Do klasyfikacji wyjazdów turystycznych ze względu na rodzaj wyjazdu wykorzystano drzewa klasyfikacyjne, natomiast do segmentacji gospodarstw domowych wykorzystano drzewo regresyjne.

Słowa kluczowe: gospodarstwa domowe seniorów, wyjazdy turystyczne, drzewa klasyfikacyjne i regresyjne

WSTĘP

W społeczeństwach wielu krajów świata, w tym Polski, coraz większą i znaczną część stanowi pokolenie ludzi starszych będących w wieku emerytalnym lub do niego się zbliżających. Zmiany zachodzące w strukturze demograficznej społeczeństw i stale rosnąca liczba osób starszych spowodowała, że producenci towarów i usług zaczęli dostosowywać ich rodzaje i asortymenty do potrzeb osób tej grupy wiekowej [Bąk 2011]. Współcześni seniorzy różnią się od swoich poprzedników tym, że żyją dłużej, są bogatsi, lepiej wykształceni i są zdrowsi. Także inaczej patrzą na wiek emerytalny – chcą realizować swoje marzenia i rozwijać pasje, a rynek towarów, a zwłaszcza usług w tym im pomaga. Rosnąca liczba osób starszych, a szczególnie przebywających na emeryturze, może stać się ważnym czynnikiem rozwoju szeroko pojętego przemysłu turystycznego. Profil demograficzny poszczególnych segmentów rynku turystycznego będzie

w przyszłości wyraźnie różnicować ofertę usługową przedsiębiorstw turystycznych. Zdaniem wielu autorów, nabycie przez osoby w wieku emerytalnym umiejętności zagospodarowania czasu wolnego i prowadzenie przez nie aktywnego (na miarę ich sił i potrzeb) stylu życia umożliwi zachowanie dobrej sprawności psychofizycznej w starości i czerpanie radości także i z tej fazy ich życia [Trafiałek 2006]. Pozytywny wpływ aktywności turystycznej na zdrowie i samopoczucie starszych osób jest niezaprzeczalny i potwierdzony wieloma wynikami badań.

Polscy seniorzy, na tle seniorów zachodnioeuropejskich czy amerykańskich, a także innych grup wiekowych własnego kraju wykazują stosunkowo niski poziom konsumpcji turystycznej. Główną przyczyną tej sytuacji jest nie tylko znacznie niższa siła nabywcza współczesnego, polskiego seniora, ale również wyznawany przez niego system wartości, w którym turystyka oraz inne formy aktywności spędzania czasu wolnego nie zajmują poczesnego miejsca. Biorąc pod uwagę rosnące zasoby czasu wolnego oraz zwiększającą się liczbę osób w tym wieku należy sądzić, że będą one stymulująco oddziaływać na kształtowanie koniunktury na rynku usług turystycznych. W tym celu niezbędne jest podejmowanie badań w sferze konsumpcji turystycznej, które dostarczą istotnych informacji o motywach i zachowaniach konsumpcyjnych turystów-seniorów oraz szacunkowych wielkościach środków, które mogą oni wydatkować na wypoczynek.

W artykule sformułowano dwa cele badawcze. Pierwszy z nich dotyczy klasyfikacji wyjazdów turystycznych seniorów[1] 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 seniorów (głowa gospodarstwa była w wieku 60 lat i więcej) 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 seniorów zaczerpnięto z badań ankietowych *„Turystyka i wypoczynek w gospodarstwach domowych”* przeprowadzonych przez Główny Urząd Statystyczny w 2009 roku. Dane mają charakter reprezentacyjny i pochodzą z badań cyklicznych przeprowadzanych co cztery lata. W badaniu uwzględniono indywidualne wyjazdy zrealizowane przez gospodarstwa domowe seniorów w okresie 1.10.2008-30.09.2009.

[1] W pracy zamiennie używane są pojęcia takie jak: „osoby starsze” i „seniorzy” w odniesieniu do osób powyżej 60 roku życia.

ISTOTA DRZEW KLASYFIKACYJNYCH I REGRESYJNYCH

Drzewa klasyfikacyjne i regresyjne zaliczane są do metod statystycznej analizy wielowymiarowej. Znajdują zastosowanie do klasyfikacji obiektów wówczas, gdy w zbiorze badanych zmiennych można wyróżnić zmienną zależną, a badane zmienne (zależna i niezależne) mogą być mierzone zarówno na skalach słabych (nominalna, porządkowa), jak i na skalach mocnych (przedziałowa, ilorazowa) [Gatnar, Walesiak 2004, s. 56-59].

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.

Sposób definiowania funkcji wskaźnikowej $\boldsymbol{I}$ zależy od charakteru zmiennych objaśniających ($X_1,\ldots,X_L$). Gdy zmienne te 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.

Jeżeli 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:[2]

1. dla zmiennej zależnej niemetrycznej: błąd klasyfikacji, wskaźnik Giniego, miarę entropii, statystykę χ^2,
2. dla zmiennej zależnej metrycznej – wariancję zmiennej zależnej.

EMPIRYCZNE MODELE DRZEW KLASYFIKACYJNYCH I REGRESYJNYCH

W artykule do wyznaczenia drzew klasyfikacyjnych i regresyjnych wykorzystano procedurę CART. Obliczeń dokonano w programie Statistica 9.0 przy założeniach przedstawionych w tab. 1.

Tabela 1. Założenia przyjęte w procedurze CART

Wyszczególnienie	Modele ogólne drzew	
	klasyfikacyjnych	regresyjnych
Koszty błędnej klasyfikacji	równe	-
Miary dopasowania (reguła podziału)	wskaźnik Giniego	-
Kryterium stopu	przy błędnej klasyfikacji	przytnij według wariancji
Minimalna liczność	30	30
Maksymalna liczba węzłów	1000	1000

Źródło: opracowanie własne

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

Do klasyfikacji wyjazdów turystycznych seniorów ze względu na rodzaj wyjazdu wykorzystano drzewa klasyfikacyjne. Jako zmienną zależną przyjęto rodzaj wyjazdu (krajowy, zagraniczny), natomiast w zbiorze zmiennych niezależnych uwzględniono[3]:

1. predyktory jakościowe:
- 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;
- cel wyjazdu: wypoczynek (rekreacja, wakacje), zwiedzanie (architektura,, kultura, przyroda), odwiedziny u krewnych lub znajomych oraz uroczystości rodzinne, zdrowotny, inny;

2. predyktor ilościowy: roczne wydatki ogółem poniesione w związku z wyjazdem.

Do segmentacji gospodarstw domowych wykorzystano drzewa regresyjne. Zmienną zależną zdefiniowano jako łączne roczne wydatki poniesione przez gospodarstwo domowe na wyjazdy turystyczne (wartości od 37 zł do 14860 zł), natomiast zbiór zmiennych niezależnych tworzyły:
- predyktory jakościowe: płeć (kobieta, mężczyzna), wykształcenie (bez wykształcenia, podstawowe, zasadnicze zawodowe, średnie, wyższe), miejsce zamieszkania (wieś, miasto poniżej 20 tys. mieszkańców, miasto od 20 do 99 tys. mieszkańców, miasto od 100 do 199 tys. mieszkańców, miasto od 200 do 499 tys. mieszkańców, miasto 500 tys. mieszkańców i więcej), typ biologiczny gospodarstwa domowego (małżeństwo bez dzieci, małżeństwo z dziećmi, gospodarstwo jednoosobowe, pozostałe), rodzaj wyjazdu (krajowy, zagraniczny);
- predyktory ilościowe: przeciętny miesięczny dochód gospodarstwa, wiek, liczba osób pracujących, liczba wyjazdów w ciągu roku.

W tabeli 2 przedstawiono procedurę wyboru drzewa regresyjnego oraz klasyfikacyjnego, które następnie wykorzystano do interpretacji wyników.

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

W wyniku zastosowanej procedury CART otrzymano sekwencję 6 drzew klasyfikacyjnych i 19 drzew regresyjnych. Następnie na podstawie analizy wykresu przedstawiającego poziom kosztów sprawdzianu krzyżowego oraz kosztów resubstytucji na tle złożoności drzewa, wybrano dla każdego rodzaju drzew po trzy drzewa optymalne (kryterium – najmniejsza różnica między kosztem sprawdzianu krzyżowego a kosztem resubstytucji)[4]. W kolejnym etapie drzewa optymalne oceniono pod względem złożoności oraz liczby i ważności wykorzystanych przy podziale predyktorów [Batóg, Mojsiewicz, Wawrzyniak 2011, s. 169]. Za najlepsze uznano drzewo klasyfikacyjne nr 3 oraz drzewo regresyjne nr 14.

Interpretując wybrane drzewo klasyfikacyjne (rys. 1) sformułowano następujące wnioski wykorzystując w tym celu regułę zdań warunkowych typu *„jeżeli…, to…”*:

- jeżeli formą wyjazdu była wycieczka (impreza objazdowa, pielgrzymka) mająca na celu zwiedzanie (architektury, kultury, przyrody), to wyjazd był wyjazdem zagranicznym; natomiast, gdy wyjeżdżano w innym celu, to wyjazd był wyjazdem krajowym;
- jeżeli formą wyjazdu były wczasy lub inna forma (rodzina, działka), a środkiem transportu była kolej, PKS lub inna autobusowa linia przewozowa, to wyjazd był wyjazdem krajowym; natomiast gdy korzystano z innego środka transportu (samolot, prom), to wyjazd był wyjazdem zagranicznym.

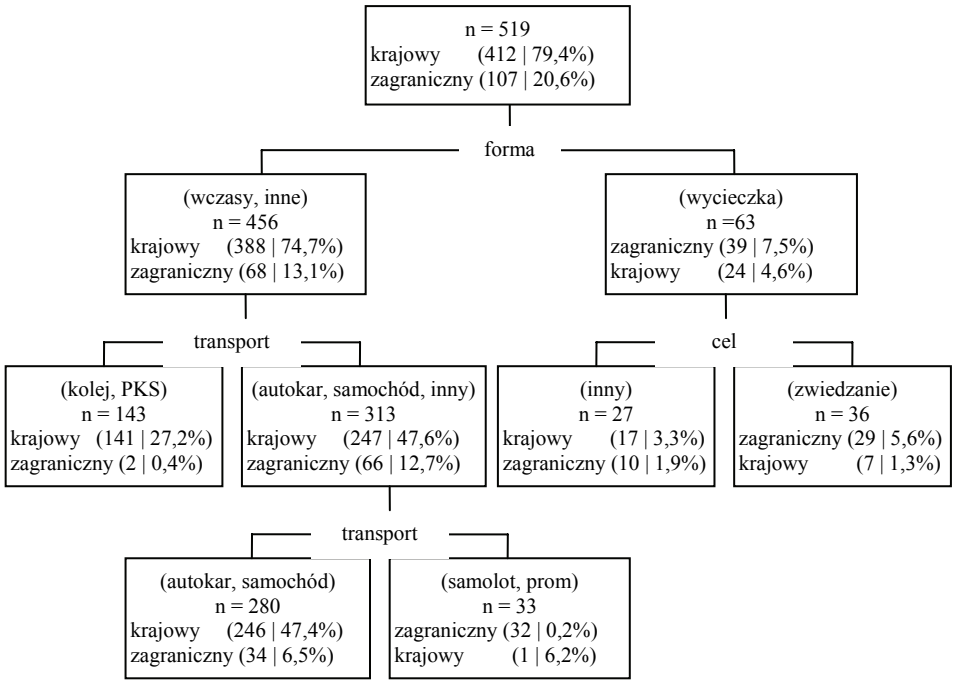
Tabela 2. Drzewa optymalne wybrane ze względu na zbieżność kosztów sprawdzianu krzyżowego oraz kosztów resubstytucji

Wyszczególnienie	Modele ogólne drzew	
	klasyfikacyjnych	regresyjnych
Liczba drzew w sekwencji drzew	6	19
Numery drzew optymalnych	5, 3,2	18, 16, 14
Numer drzewa wybranego	3	14
– liczba węzłów dzielonych		3
– liczba węzłów końcowych		6
– ważność predyktorów	1. forma 2. transport 3. cel 4. pośrednictwo 5. wydatki 6. charakter odwiedzanego obszaru	1. rodzaj wyjazdu 2. dochód 3. liczba wyjazdów 4. miejsce zamieszkania 5. wykształcenie 6. wiek 7. typ rodziny 8 liczba pracujących 9. płeć

Źródło: opracowanie własne

[4] Program *Statistica* za najlepsze drzewa uznał drzewo klasyfikacyjne o nr 5 i drzewo regresyjne o nr 18, czyli drzewa o jednym węźle dzielonym i o dwóch węzłach końcowych.

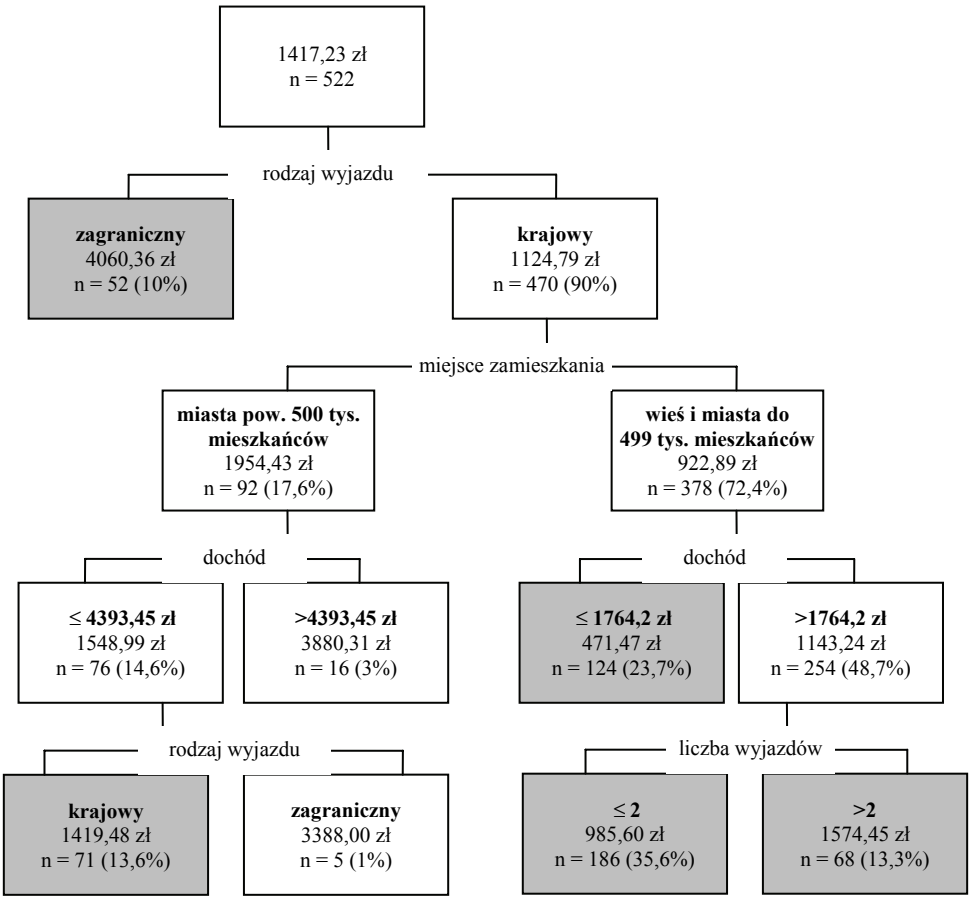
Rysunek 1. Drzewo klasyfikacyjne dla rodzaju wyjazdu (w nawiasach podano odsetki obliczone w stosunku do liczebności całej próby)



Źródło: opracowanie własne

Na podstawie drzewa regresyjnego nr 14 wydzielonych zostało 7 segmentów gospodarstw domowych ze względu na ich roczne wydatki na wyjazdy turystyczne. Przy wyborze segmentu do interpretacji przyjęto założenie, że jego liczebność powinna stanowić przynajmniej 10% liczebności próby. Warunek ten spełnia tylko pięć węzłów końcowych, które na rys. 2 zostały wyróżnione szarym kolorem. Wyniki segmentacji zamieszczono w tab. 3.

Rysunek 2. Drzewo regresyjne dla rocznych wydatków ogółem poniesionych w związku z wyjazdem turystycznym w ciągu roku (w nawiasach podano odsetki obliczone w stosunku do liczebności całej próby)



Źródło: opracowanie własne

Tabela 3. Charakterystyka segmentów gospodarstw domowych seniorów ze względu na ich wydatki na wyjazdy turystyczne

Nr	Charakterystyka segmentu	Przeciętne roczne wydatki na wyjazdy turystyczne (zł)	Liczebność segmentu (% badanej próby)
1	Gospodarstwa domowe biorące udział w wyjazdach zagranicznych	4060,37	52 (10,0%)
2	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 4393,45 zł., zlokalizowane w miastach powyżej 500 tys. mieszkańców biorące udział w wyjazdach krajowych	1419,48	71 (13,6%)
3	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 1764,47 zł., zlokalizowane na wsi lub w miastach do 499 tys. mieszkańców biorące udział w wyjazdach krajowych	471,47	124 (23,8%)
4	Gospodarstwa domowe o miesięcznych dochodach powyżej 1764,47 zł., zlokalizowane na wsi lub w miastach do 499 tys. mieszkańców biorące udział w co najwyżej dwóch wyjazdach krajowych	985,60	186 (35,63%)
5	Gospodarstwa domowe o miesięcznych dochodach powyżej 1764,47 zł., zlokalizowane na wsi lub w miastach do 499 tys. mieszkańców biorące udział w co najmniej dwóch wyjazdach krajowych	1574,46	68 (13,0%)

Źródło: opracowanie własne

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 predykatorami w tym przypadku okazały się następujące zmienne niezależne: forma wyjazdu, cel wyjazdu i środek transportu. Wśród wyjazdów turystycznych realizowanych przez gospodarstwa domowe seniorów w okresie 1.10.2008-30.09.2009 dominowały wyjazdy krajowe na wczasy, do rodziny lub działkę, a głównym środkiem transportu była kolej, PKS lub samochód. Natomiast wyjazd zagraniczny był organizowany w formie wycieczki mającej na celu zwiedzanie architektury, kultury, przyrody a środkiem transportu był samolot.

Wykorzystanie drzew regresyjnych pozwoliło na wydzielenie segmentów gospodarstw domowych, które znacząco różniły się pod względem poziomu przeciętnych rocznych wydatków na wyjazdy turystyczne. Okazało się, że najliczniejszą grupę gospodarstw domowych seniorów (ponad 35%) stanowiły gospodarstwa wydające na turystykę przeciętnie w ciągu roku około 986 zł i są to gospodarstwa, których członkowie biorą udział w co najwyżej dwóch wyjazdach krajowych. Dogłębna analiza tego segmentu wykazała, że znajdujące się w nim gospodarstwa najczęściej korzystają z wyjazdów krajowych krótkoterminowych (2-4 dni) mających na celu odwiedziny krewnych lub znajomych. Najwyższe wydatki na wyjazdy ponoszą gospodarstwa biorące udział w wyjazdach zagranicznych, z tym, że liczebność tego segmentu jest nieznaczna i wynosi tylko 52 gospodarstwa, co stanowi ok. 10% wszystkich gospodarstw aktywnych turystycznie.

Uzyskane wyniki badań pozwalają poznać preferencje gospodarstw domowych seniorów w zakresie rodzaju wyjazdu oraz przeciętnych miesięcznych wydatków na wyjazdy turystyczne. Powyższe informacje mogą pozwolić odpowiednim organizacjom (np. Klubom Seniora), stowarzyszeniom i biurom turystycznym na przygotowanie oferty odpowiedniej do oczekiwań badanych zbiorowości.

BIBLIOGRAFIA

Batóg B., Mojsiewicz M., Wawrzyniak K. (2011) Segmentacja gospodarstw domowych ze względu na popyt potencjalny i zrealizowany na rynku ubezpieczeń życiowych w Polsce, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Bąk I., Wawrzyniak K. (2009) Zastosowanie analizy korespondencji w badaniach związanych z motywami wyboru rodzajów wyjazdów turystycznych przez emerytów i rencistów w 2005 roku, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, Taksonomia 16. Klasyfikacja i analiza danych – teoria i zastosowania, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Bąk I. (2011) Analiza wyjazdów turystycznych emerytów, Wiadomości Statystyczne nr 12, GUS, Warszawa.

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

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

Gatnar E., Walesiak M. (2004) Metody statystycznej analizy wielowymiarowej w badaniach marketingowych, Wydawnictwo AE we Wrocławiu, Wrocław.

Trafiałek E. (2006) Starzenie się i starość. Wybór tekstów z gerontologii społecznej, Wszechnica Świętokrzyska, Kielce.

USING OF CLASSIFICATION AND REGRESSION TREES FOR ANALYSIS OF HOUSEHOLDS TOURIST TRIPS OF SENIORS IN POLAND

Summary: In the article the present results of the classification of touristic travels of seniors according to the kind of the travel and the segmentation of households of seniors according to their participation in tourist traffic. In the research were used the individual trips realized by seniors households in Poland in period 1.10.2008-30.09.2009. The classification of touristic travels of seniors was conducted by means of classification trees and the segmentation of households – by means of regression trees.

Keywords: households of seniors, tourist departures, classification and regression trees