PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Barbara Batóg, Magdalena Mojsiewicz
Uniwersytet Szczeciński

Katarzyna Wawrzyniak
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

SEGMENTACJA GOSPODARSTW DOMOWYCH ZE WZGLĘDU NA POPYT POTENCJALNY I ZREALIZOWANY NA RYNKU UBEZPIECZEŃ ŻYCIOWYCH W POLSCE

Streszczenie: W artykule dokonano segmentacji gospodarstw domowych ze względu na popyt potencjalny i zrealizowany na rynku ubezpieczeń życiowych w Polsce z wykorzystaniem drzew regresyjnych i klasyfikacyjnych. Popyt potencjalny zdefiniowano jako zadeklarowaną miesięczną składkę na ubezpieczenia życiowe, natomiast popyt zrealizowany jako łączną liczbę polis na życie posiadanych przez gospodarstwo domowe. Dla każdego rodzaju popytu wydzielono segmenty gospodarstw domowych na podstawie cech demograficzno-ekonomicznych oraz preferencji dotyczących ubezpieczeń na życie. Chociaż zbiór zmiennych objaśniających był taki sam dla obu rodzajów popytu, to ich ranking ważności był inny w drzewach klasyfikacyjnych niż w drzewach regresyjnych.

Słowa kluczowe: popyt na rynku ubezpieczeń na życie, segmentacja, drzewa klasyfikacyjne i regresyjne.

1. Wstęp

Artykuł jest kontynuacją badań autorek, dotyczących zastosowania metod statystycznych w segmentacji rynku ubezpieczeń w Polsce. We wcześniejszych pracach [Mojsiewicz, Wawrzyniak 2005; Batóg, Wawrzyniak 2006; Batóg i in. 2007] jako narzędzie badawcze w segmentacji polskiego rynku ubezpieczeń autorki wykorzystywały skalowanie wielowymiarowe. Badani respondenci byli charakteryzowani głównie poprzez mierzone na skali porządkowej preferencje dotyczące produktów i usług ubezpieczeniowych.

Celem tego artykułu jest segmentacja gospodarstw domowych ze względu na popyt potencjalny i zrealizowany na rynku ubezpieczeń życiowych w Polsce. Jako

narzędzie segmentacji zaproponowano drzewa regresyjne i klasyfikacyjne, w których zmienne zależne zostały zdefiniowane jako[1]:

– dla popytu potencjalnego – deklarowana miesięczna składka na ubezpieczenia życiowe,
– dla popytu zrealizowanego – łączna liczba polis na życie posiadanych przez gospodarstwo domowe.

Dla obu rodzajów drzew wykorzystano taki sam zbiór zmiennych objaśniających, który zostanie omówiony w części metodologicznej artykułu. Segmentację przeprowadzono dla reprezentatywnej próby 500 gospodarstw domowych[2].

2. Metodologia drzew regresyjnych i klasyfikacyjnych

W badaniach marketingowych drzewa klasyfikacyjne i regresyjne są najczęściej stosowane jako narzędzie segmentacji oraz do określania cech produktów, które dla konsumentów są najistotniejsze. Wykorzystanie drzew klasyfikacyjnych i regresyjnych jest zasadne wówczas, gdy w zbiorze badanych zmiennych można wskazać zmienną zależną, a badane zmienne (zależna i objaśniające) mogą być mierzone zarówno na skalach słabych (nominalna, porządkowa), jak i na skalach mocnych (przedziałowa, ilorazowa) [Gatnar, Walesiak 2004].

Drzewa klasyfikacyjne i regresyjne są graficzną reprezentacją modelu postaci [Gatnar 2008, s. 37-39][3]:

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

gdzie: Y – zmienna zależna,
R_k (k = 1,…, K, K – liczba segmentów) to podprzestrzenie (segmenty) przestrzeni zmiennych objaśniających $\mathbf{X}^L$ (X_1, X_2, ..., 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,…, X_L). Jeżeli te zmienne mają charakter metryczny, to

[1] Zmienne zależne i objaśniające wybrane do budowy drzew regresyjnych i klasyfikacyjnych zostały przyjęte na podstawie badań autorek, których wyniki zamieszczono w pracy [Batóg i in. 2010].

[2] Badania terenowe zostały przeprowadzone w lipcu i sierpniu 2005 roku przez TNS OBOP w ramach grantu KBN *Metody statystyczne w segmentacji rynku ubezpieczeń w Polsce*, nr 1H02B02827.

[3] Metodologia drzew klasyfikacyjnych i regresyjnych jest wyczerpująco omówiona w pracach [Gatnar 2001, 2008].

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

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

Jeżeli te zmienne mają charakter niemetryczny, to

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

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

Parametry α_k modelu (1) w zależności od charakteru zmiennej zależnej *Y* wyznacza się według następujących wzorów:

– gdy *Y* jest zmienną nominalną

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

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,

– gdy *Y* jest mierzone na skalach mocnych

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

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

W pierwszym przypadku graficzną postacią modelu (1) jest drzewo klasyfikacyjne, a w drugim – drzewo regresyjne.

3. Empiryczne modele drzew regresyjnych i klasyfikacyjnych

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

Dla popytu potencjalnego (zmienna zależna – deklarowana miesięczna składka na ubezpieczenia życiowe, przyjmująca wartości z przedziału od 0 do 1000 zł) zbudowano drzewa regresyjne, natomiast dla popytu zrealizowanego (zmienna zależna – łączna liczba polis na życie posiadanych przez gospodarstwo domowe, przyjmująca wartości od 0 do 7 i więcej polis[4]) – drzewa klasyfikacyjne.

[4] Łączna liczba polis posiadanych przez gospodarstwo domowe została potraktowana jako zmienna porządkowa. Do ostatniej kategorii należały gospodarstwa, które posiadały przynajmniej

Tabela 1. Założenia przyjęte w procedurze CART

Wyszczególnienie	Modele ogólne drzew	
	regresyjnych	klasyfikacyjnych
Koszty błędnej klasyfikacji	–	równe
Miary dopasowania (reguła podziału)	–	wskaźnik Giniego
Kryterium stopu	przytnij według wariancji	przy błędnej klasyfikacji
Minimalna liczność	50	50
Maksymalna liczba węzłów	1000	1000

Źródło: opracowanie własne.

Dla obu modeli przyjęto taki sam zbiór zmiennych objaśniających, który tworzyły:
- predyktory jakościowe: płeć i wykształcenie głowy gospodarstwa domowego (co najwyżej gimnazjalne, zasadnicze zawodowe, średnie, wyższe), miejsce zamieszkania (wieś, miasto o liczbie ludności 50 tys. i mniej, miasto o liczbie ludności 51-200 tys., miasto o liczbie ludności 201-500 tys., miasto o liczbie ludności powyżej 500 tys.),
- predyktory ilościowe: wiek głowy gospodarstwa domowego, liczba osób w gospodarstwie domowym, miesięczny dochód rozporządzalny gospodarstwa domowego.

Dodatkowo do zbioru zmiennych objaśniających wprowadzono poziom posiadanych oszczędności w dwóch wariantach: jako predyktor jakościowy (brak oszczędności, w wysokości nie wyższej niż miesięczne dochody rodziny, w wysokości pomiędzy miesięcznymi a rocznymi dochodami rodziny, w wysokości co najmniej rocznych dochodów rodziny) i jako predyktor ilościowy.

W tabeli 2 przedstawiono procedurę wyboru drzewa regresyjnego oraz klasyfikacyjnego, które następnie wykorzystano do segmentacji gospodarstw domowych ze względu na popyt potencjalny i zrealizowany. W wyniku zastosowanej procedury CART otrzymano sekwencję 19 drzew regresyjnych i 10 drzew klasyfikacyjnych. Następnie, na podstawie analizy wykresu przedstawiającego poziom kosztów sprawdzianu krzyżowego oraz kosztów resubstytucji na tle złożoności drzewa, wybrano dla każdego rodzaju drzew po cztery drzewa optymalne (kryterium – najmniejsza różnica między kosztem sprawdzianu krzyżowego a kosztem resubstytucji)[5]. W kolejnym etapie drzewa optymalne oceniono pod względem złożoności oraz liczby i ważności wykorzystanych przy podziale predyktorów. Za najlepsze uznano drzewo regresyjne nr 13 (rys. 1) oraz drzewo klasyfikacyjne nr 6 (rys. 2).

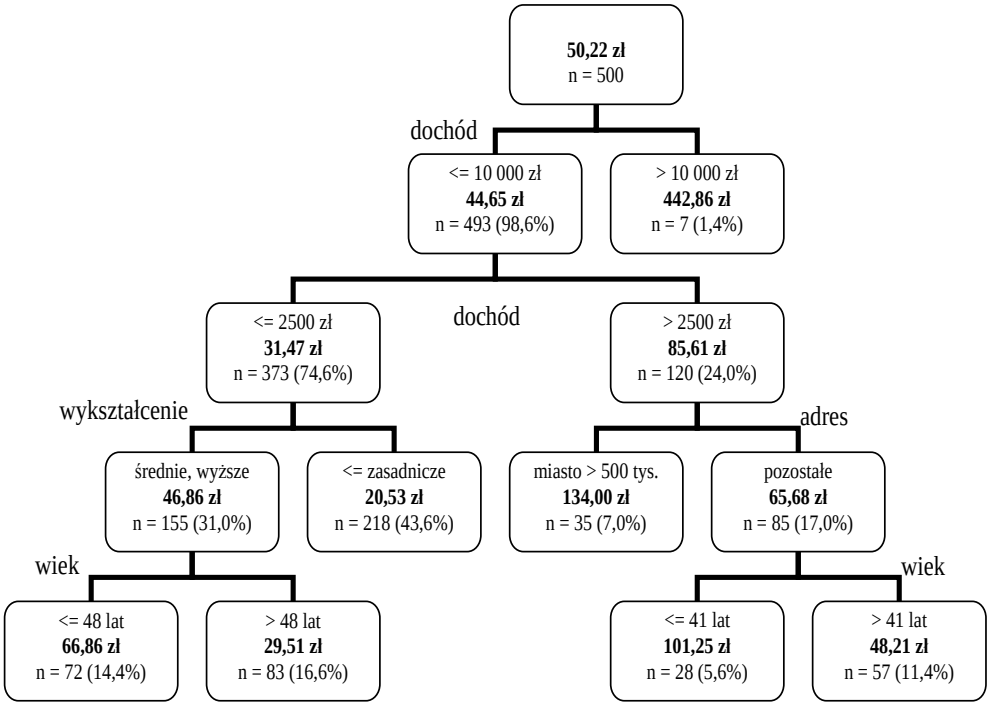
7 polis na życie. Liczba polis nie jest dokładnym pomiarem jakości zabezpieczeń gospodarstwa domowego, byłaby nim łączna suma, na którą zawarto ubezpieczenia.

[5] Program *Statistica* za najlepsze drzewa uznał drzewo regresyjne o nr. 18 i drzewo klasyfikacyjne o nr. 9, czyli drzewa o jednym węźle dzielonym i o dwóch węzłach końcowych.

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

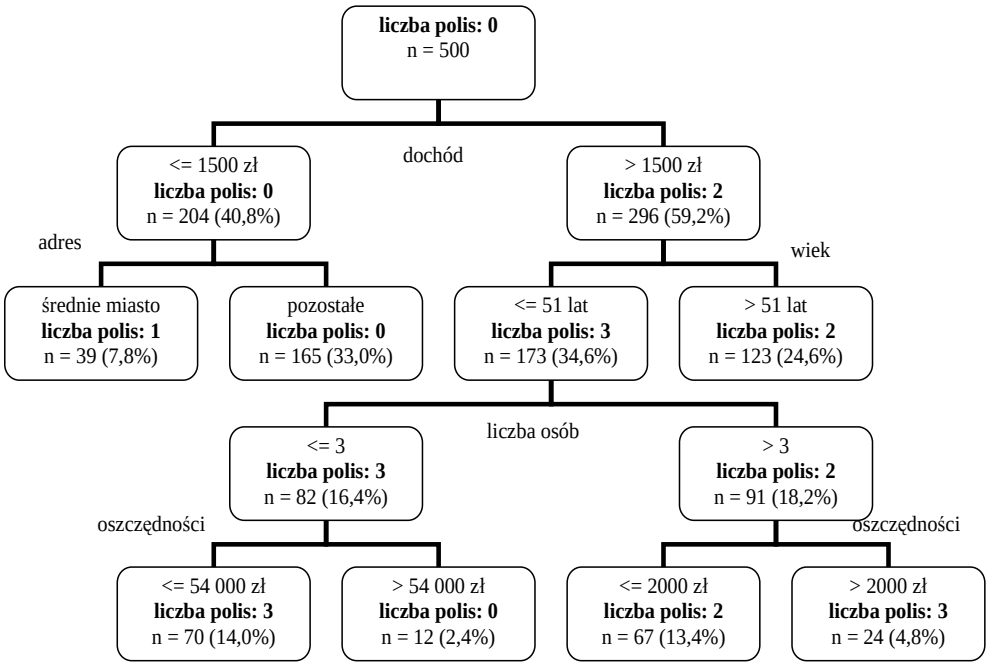
Wyszczególnienie	Modele ogólne drzew	
	regresyjnych	klasyfikacyjnych
Liczba drzew w sekwencji drzew	19	10
Numery drzew optymalnych	18, 16, 13, 11	9, 7, 6, 5
Numer drzewa wybranego	13	6
− liczba węzłów dzielonych	6	6
− liczba węzłów końcowych	7	7
− ważność predyktorów	1. dochód 2. adres 3. wykształcenie 4. wiek 5. oszczędności 6. liczba osób 7. płeć	1. oszczędności 2. dochód 3. wiek 4. adres 5. wykształcenie 6. liczba osób 7. płeć

Źródło: opracowanie własne.



Rys. 1. Drzewo regresyjne nr 13 dla popytu potencjalnego

Źródło: obliczenia własne.



Rys. 2. Drzewo klasyfikacyjne nr 6 dla popytu zrealizowanego

Źródło: obliczenia własne.

4. Segmentacja gospodarstw domowych ze względu na popyt potencjalny na podstawie drzewa regresyjnego

Na podstawie drzewa regresyjnego nr 13 wydzielonych zostało 5 segmentów gospodarstw domowych ze względu na popyt potencjalny na rynku ubezpieczeń życiowych. Cztery z nich spełniały warunek, że liczba gospodarstw domowych w segmencie stanowiła przynajmniej 10% liczby wszystkich badanych gospodarstw domowych. Segment nr 5 został wyróżniony ze względu na najwyższą zadeklarowaną składkę na ubezpieczenia życiowe. Szczegółowa charakterystyka wydzielonych segmentów na podstawie zmiennych społeczno-demograficznych (drzewo regresyjne) oraz na podstawie preferencji (dodatkowe obliczenia)[6] została zamieszczona w tab. 3.

[6] W ramach wydzielonych na podstawie drzewa segmentów przeprowadzono statystyczną analizę rozkładów zmiennych charakteryzujących gospodarstwa domowe ze względu na ich preferencje dotyczące produktów i usług ubezpieczeniowych na życie. W opisie uwzględniono tylko te preferencje, które istotnie różnicowały segmenty między sobą.

Tabela 3. Charakterystyka segmentów gospodarstw domowych ze względu na popyt potencjalny

Numer	Charakterystyka segmentu		Przeciętna miesięczna deklarowana składka na ubezpieczenia życiowe (zł)	Liczebność segmentu (odsetek badanej próby)
	na podstawie drzewa	na podstawie preferencji		
1	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 2500 zł, w których głowa gospodarstwa domowego ma wykształcenie co najwyżej zasadnicze zawodowe.	Nie sugerują się informacjami o wydolności publicznego systemu opieki zdrowotnej, nie planują tego wydatku w budżecie domowym, nie korzystają z ubezpieczeń grupowych.	20,53 zł	218 (43,6 %)
2	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 2500 zł, w których głowa gospodarstwa domowego ma wykształcenie co najmniej średnie i jest w wieku powyżej 48 lat.	Ważniejszy jest dla nich zakres świadczenia niż cena, nie korzystają z ubezpieczeń grupowych ani porad pośrednika, wyszukują dużą firmę z dobrą marką.	29,51 zł	83 (16,6%)
3	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 2500 zł, w których głowa gospodarstwa domowego ma wykształcenie co najmniej średnie i jest w wieku nie wyższym niż 48 lat.	Sugerują się informacjami o wydolności systemu ubezpieczeń społecznych, korzystają z ubezpieczeń grupowych, planują ten wydatek w budżecie, ważniejszy jest dla nich zakres świadczenia niż cena, wyszukują dużą firmę z dobrą marką.	66,86 zł	72 (14,4%)
4	Gospodarstwa domowe o miesięcznych dochodach powyżej 2500 zł (ale nie więcej niż 10 000 zł), znajdujące się w małych i średnich miastach oraz na wsi; głowa gospodarstwa domowego ma więcej niż 41 lat.	Sugerują się informacjami o wydolności systemu ubezpieczeń społecznych, korzystają z ubezpieczeń grupowych, wyszukują dużą firmę z dobrą marką.	48,21 zł	57 (11,4%)
5	Gospodarstwa domowe o miesięcznych dochodach powyżej 2500 zł (ale nie więcej niż 10 000 zł), znajdujące się w dużych miastach.	Sugerują się informacjami o wydolności systemu ubezpieczeń społecznych oraz publicznego systemu opieki zdrowotnej, planują ten wydatek w budżecie, ważniejszy jest dla nich zakres świadczenia niż cena, wyszukują dużą firmę z dobrą marką, korzystają z porad pośrednika.	134 zł	35 (7%)

Źródło: opracowanie własne.

5. Segmentacja gospodarstw domowych ze względu na popyt zrealizowany na podstawie drzewa klasyfikacyjnego

Analogicznie jak w przypadku segmentacji gospodarstw domowych ze względu na popyt potencjalny na podstawie drzewa klasyfikacyjnego nr 6 wyznaczono cztery segmenty gospodarstw domowych o liczebnościach przekraczających 10% liczby wszystkich badanych gospodarstw domowych. Segment nr 5 wydzielono, gdyż stanowi on uzupełnienie segmentu nr 1, a mianowicie ukazuje odmienne zachowania gospodarstw domowych o niskich dochodach, znajdujących się w średnich miastach, w odróżnieniu od zachowań mieszkańców wsi i pozostałych miast. Charakterystykę segmentów gospodarstw domowych ze względu na popyt zrealizowany, uwzględniającą zarówno zmienne społeczno-demograficzne, jak i preferencje, zamieszczono w tab. 4.

Tabela 4. Charakterystyka segmentów gospodarstw domowych ze względu na popyt zrealizowany

Numer	Charakterystyka segmentu		Dominująca w segmencie liczba polis	Liczebność segmentu (odsetek badanej próby)
	na podstawie drzewa	na podstawie preferencji		
1	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 1500 zł, znajdujące się w małych i dużych miastach oraz na wsi.	Nie planują tego wydatku w budżecie, nie korzystają z ubezpieczeń grupowych, niższa cena nie stanowi motywacji do zakupu, nie korzystają z porad pośrednika.	0	165 (33%)
2	Gospodarstwa domowe o miesięcznych dochodach powyżej 1500 zł, w których głowa gospodarstwa domowego jest w wieku powyżej 51 lat.	Nie korzystają z ubezpieczeń grupowych, ważniejszy jest dla nich zakres świadczenia niż cena, nie korzystają z porad pośrednika, wyszukują firmę dużą z dobrą marką.	2	123 (24,6%)
3	Gospodarstwa domowe liczące 3 osoby i mniej o miesięcznych dochodach powyżej 1500 zł oraz oszczędnościach nie wyższych niż 54 000 zł, w których głowa gospodarstwa domowego ma nie więcej niż 51 lat.	Sugerują się informacjami o wydolności systemu ubezpieczeń społecznych, planują ten wydatek w budżecie, korzystają z ubezpieczeń grupowych, ważniejszy jest dla nich zakres świadczenia niż cena, wyszukują dużą firmę z dobrą marką, korzystają z porad pośrednika.	3	70 (14%)
4	Gospodarstwa domowe powyżej 3 osób, o miesięcznych dochodach powyżej 1500 zł oraz oszczędnościach nie wyższych niż 2000 zł, w których głowa gospodarstwa domowego ma nie więcej niż 51 lat.	Ważniejszy jest dla nich zakres świadczenia niż cena, wyszukują dużą firmę z dobrą marką.	2	67 (13,4%)
5	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 1500 zł, znajdujące się w średnich miastach.	ważniejszy jest dla nich zakres świadczenia niż cena, korzystają z porad pośrednika.	1	39 (7,8%)

Źródło: opracowanie własne.

6. Podsumowanie

Na podstawie przeprowadzonych badań można sformułować następujące wnioski:
- zastosowanie drzew regresyjnych i klasyfikacyjnych umożliwiło wydzielenie segmentów gospodarstw domowych na podstawie cech demograficzno-ekonomicznych, które istotnie różnicowały je pod względem popytu potencjalnego i zrealizowanego;
- rankingi predyktorów dla popytu potencjalnego i zrealizowanego potwierdzają, że inne zmienne decydują o przynależności do segmentu w przypadku pomiaru skłonności (deklaracji), a inne w przypadku pomiaru wysokości składki wydatkowanej w rzeczywistości;
- poziom posiadanych oszczędności różnicował wydzielone segmenty tylko w przypadku popytu zrealizowanego, natomiast w przypadku popytu potencjalnego deklarowana składka nie zależała od poziomu oszczędności;
- uzupełnienie wyników segmentacji demograficzno-ekonomicznej o analizę preferencji pozwoliło wyróżnić takie preferencje, które nie różnicują segmentów (np. wpływ reklamy), i takie, które pojawiają się w segmentach z wyższą tak zwaną świadomością ubezpieczeniową (np. korzystanie z porad pośrednika, planowanie wydatku w budżecie).

Literatura

Batóg B., Wawrzyniak K., *Identyfikacja osi w skalowaniu wielowymiarowym na przykładzie segmentacji rynku ubezpieczeniowego*, Prace Naukowe AE we Wrocławiu nr 1126, Wrocław 2006, s. 372-379.

Batóg B., Mojsiewicz M., Wawrzyniak K., *Efektywność metod statystycznej analizy wielowymiarowej jako narzędzia segmentacji rynku ubezpieczeniowego*, Prace Naukowe AE we Wrocławiu nr 1169, Wrocław 2007, s. 393-401.

Batóg B., Mojsiewicz M., Wawrzyniak K., *Klasyfikacja gospodarstw domowych pod względem popytu potencjalnego i zrealizowanego na rynku ubezpieczeń w Polsce*, Prace Naukowe UE we Wrocławiu nr 107, Wrocław 2010, s. 153-160.

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

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

Gatnar E., Walesiak M., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wyd. Akademii Ekonomicznej we Wrocławiu, Wrocław 2004.

Mojsiewicz M., Wawrzyniak K., *Metodologia segmentacji rynku ubezpieczeniowego*, Prace Naukowe AE we Wrocławiu nr 1076, Wrocław 2005, s. 416-422.

THE SEGMENTATION OF HOUSEHOLDS ACCORDING TO POTENTIAL AND ACTUAL DEMAND ON THE LIFE INSURANCE MARKET IN POLAND

Summary: In the paper the authors present the segmentation of households according to potential and actual demand on the life insurance market in Poland done by means of classification and regression trees. The potential demand was defined as the declared monthly premium and the actual demand was defined as the number of life insurance policies bought by a given household. The segments were distinguished on the base of socio-demographic attributes and life insurance preferences. Though the explanatory variables were the same for both demands, their ranking in case of classification trees was different than in case of regression trees.