

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 328

Taksonomia 23

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

The Central and Eastern European Online Library www.ceeol.com,

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2014

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	11
Małgorzata Rószkiewicz , Wykorzystanie metaanalizy w budowaniu modelu pomiarowego w przypadku braku niezmienniczości zasad pomiaru na przykładzie pomiaru zadowolenia z życia.....	13
Elżbieta Sobczak , Harmonijność inteligentnego rozwoju regionów Unii Europejskiej	21
Ewa Roszkowska, Renata Karwowska , Analiza porównawcza województw Polski ze względu na poziom zrównoważonego rozwoju w roku 2010.....	30
Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel , Analiza porównawcza wybranych filtrów w analizie synchronizacji cyklu koniunkturalnego.....	41
Marcin Salamaga , Próba konstrukcji tablic „wymierania scenicznego” spektakli operowych na przykładzie Metropolitan Opera.....	51
Iwona Foryś , Wykorzystanie analizy dyskryminacyjnej do typowania rynków podobnych w procesie wyceny nieruchomości niemieszkalnych	59
Jerzy Korzeniewski , Selekcja zmiennych w klasyfikacji – propozycja algorytmu	69
Sabina Denkowska , Testowanie wielokrotne przy weryfikacji wieloczynnikowych modeli proporcjonalnego hazardu Coxa.....	76
Ewa Chodakowska , Teoria równań strukturalnych w klasyfikacji zmiennych jawnych i ukrytych według charakteru ich wzajemnych oddziaływań	85
Iwona Konarzewska , Model PCA dla rynku akcji – studium przypadku	94
Katarzyna Wójcik, Janusz Tuchowski , Dobór optymalnego zestawu słów istotnych w opiniach konsumentów na potrzeby ich automatycznej analizy	106
Aleksandra Łuczak , Zastosowanie metody AHP-LP do oceny ważności determinant rozwoju społeczno-gospodarczego w jednostkach administracyjnych	116
Aleksandra Witkowska, Marek Witkowski , Klasyfikacja pozycyjna banków spółdzielczych według stanu ich kondycji finansowej w ujęciu dynamicznym	126
Adam Depta , Zastosowanie analizy korespondencji do oceny jakości życia ludności na podstawie kwestionariusza SF-36v2	135
Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej , Indukcja reguł dla danych niekompletnych i niezbalansowanych: modele klasyfikatorów i próba ich zastosowania do predykcji ryzyka operacyjnego w torakochirurgii	146

Małgorzata Misztal , Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań.....	156
Anna M. Olszewska , Wykorzystanie wybranych metod taksonomicznych do oceny potencjału innowacyjnego województw	167
Iwona Bąk , Porównanie jakości grupowań powiatów województwa zachodniopomorskiego pod względem atrakcyjności turystycznej.....	177
Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn , Segmentacja gospodarstw domowych według wydatków na turystykę zorganizowaną.....	186
Agnieszka Wałęga , Podejście syntetyczne w analizie spójności ekonomicznej gospodarstw domowych.....	196
Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek , Zastosowanie analizy korespondencji do badania wpływu elektrowni wiatrowych na jakość życia ludności	205
Joanna Banaś, Krzysztof Małecki , Klasyfikacja punktów pomiarów ankietowych kierowców na granicy Szczecina z wykorzystaniem zmiennych symbolicznych.....	214
Aneta Becker , Wykorzystanie informacji granularnej w analizie wymagań rynku pracy.....	222
Katarzyna Cheba, Joanna Holub-Iwan , Wykorzystanie analizy korespondencji w segmentacji rynku usług medycznych.....	230
Adam Depta, Iwona Staniec , Identyfikacja czynników decydujących o jakości życia studentów łódzkich uczelni.....	238
Katarzyna Dębowska, Jarosław Kilon , Reguły asocjacyjne w analizie wyników badań metodą Delphi.....	247
Anna Domagała , O wykorzystaniu analizy głównych składowych w metodzie <i>Data Envelopment Analysis</i>	254
Alicja Grześkowiak , Analiza wykluczenia cyfrowego w Polsce w ujęciu indywidualnym i regionalnym.....	264
Anna M. Olszewska, Anna Gryko-Nikitin , Pomiar postrzegania jakości kształcenia uczelni wyższej na danych porządkowych z wykorzystaniem środowiska R.....	273
Karolina Paradysz , Hierarchiczna metoda grupowania powiatów jako podejście benchmarkowe w ocenie bezrobocia według BAEL-u w wybranych typach małych obszarów	282
Radosław Pietrzyk , Porównanie metod pomiaru efektywności zarządzania portfelami funduszy inwestycyjnych.....	290
Agnieszka Przedborska, Małgorzata Misztal , Wybrane metody statystyki wielowymiarowej w ocenie skuteczności terapeutycznej głębokiej stymulacji elektromagnetycznej u pacjentów z chorobą zwyrodnieniową stawów.....	299

Wojciech Roszka, Marcin Szymkowiak , Podejście kalibracyjne w statystycznej integracji danych	308
Iwona Skrodzka , Zastosowanie wybranych metod klasyfikacji do analizy kapitału ludzkiego krajów Unii Europejskiej	316
Agnieszka Stanimir , Wielowymiarowa analiza czynników sprzyjających włączeniu społecznemu	326
Dorota Strózik, Tomasz Strózik , Przestrzenne zróżnicowanie poziomu życia w województwie wielkopolskim.....	334
Izabela Szamrej-Baran , Identyfikacja przyczyn ubóstwa energetycznego w Polsce przy wykorzystaniu modelowania miękkiego.....	343
Janusz Tuchowski, Katarzyna Wójcik , Klasyfikacja obiektów w systemie Krajowych Ram Kwalifikacji opisanych za pomocą ontologii	353
Aleksandra Matuszewska-Janica , Grupowanie krajów Unii Europejskiej ze względu na poziom feminizacji sektorów gospodarczych	361
Monika Rozkrut, Dominik Rozkrut , Identyfikacja strategii innowacyjnych przedsiębiorstw usługowych w Polsce	369

Summaries

Małgorzata Rószkiewicz , The use of meta-analysis in building the measurement model in case of the absence of measurement invariance on the example of measuring of life satisfaction.....	20
Elżbieta Sobczak , Harmonious smart growth of European Union regions.....	29
Ewa Roszkowska, Renata Karwowska , The comparative analysis of Polish voivodeships with respect to sustainable development in 2010.....	40
Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel , Comparative analysis of chosen filters in business cycles analysis	50
Marcin Salamaga , The attempt of construction of the life tables for opera works on the example of the Metropolitan Opera	58
Iwona Foryś , Using discriminant analysis to select similar markets in non-residential property valuation process.....	68
Jerzy Korzeniewski , Variable selection in classification – algorithm proposal	75
Sabina Denkowska , Multiple testing in the verification process of multifactorial Cox proportional hazards models	84
Ewa Chodakowska , The theory of structural equations modelling in the classification of observed variables and latent constructs according to the character of their relationship.....	93
Iwona Konarzewska , Modelling stock market by PCA factor model – case study	105

Katarzyna Wójcik, Janusz Tuchowski , Selection of the optimal set of relevant words in consumers opinions in the context of the opinion mining ..	115
Aleksandra Łuczak , Application of AHP-LP to the evaluation of importance of determinants of socio-economic development in the administrative units	125
Aleksandra Witkowska, Marek Witkowski , A dynamic approach to the ranking of cooperative banks by their financial condition	134
Adam Depta , Application of correspondence analysis for the measurement of quality of life – questionnaire SF-36v2 based research	145
Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej , Classification rules extraction for missing and imbalance data: models of classifiers and initial results in the rules-based thoracic surgery risk prediction.....	155
Małgorzata Misztal , Selected methods for assessing the performance of classifiers – an overview and examples of applications.....	166
Anna M. Olszewska , The application of selected quantitative methods to the evaluation of voivodeship innovation level potential.....	176
Iwona Bąk , The comparison of the quality of groupings of poviats of West Pomeranian Voivodeship in terms of tourism attractiveness	185
Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn , Household segmentation with respect to the expenditure on organized tourism.....	195
Agnieszka Wałęga , Synthetic approach in the analysis of economic coherence of households	204
Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek , Using the correspondence analysis to examine the impact of wind turbines on the quality of life.....	213
Joanna Banaś, Krzysztof Małecki , Classification of measurement survey points of drivers on the boundary of Szczecin using symbolic variables...	221
Aneta Becker , The use granular information in the analysis of the requirements of the labor market.....	229
Katarzyna Cheba, Joanna Hołub-Iwan , The application of the correspondence analysis of patients segmentation on the medical service market	237
Adam Depta, Iwona Staniec , Identification of the factors that determine the quality of students life at universities in Lodz.....	246
Katarzyna Dębkowska, Jarosław Kilon , Association rules in the analysis of research results the Delphi method	253
Anna Domagała , About using Principal Component Analysis in Data Envelopment Analysis	263
Alicja Grześkowiak , Analysis of the digital divide in Poland at the individual and regional level	272

Anna M. Olszewska, Anna Gryko-Nikitin , Assessment of perception of quality of teaching at an institution of higher learning based on the ordinal data with the utilization of R environment.....	281
Karolina Paradysz , The hierarchical method of grouping poviats as a benchmark approach in the assessment of unemployment by BAEL in selected types of small areas	289
Radosław Pietrzyk , Comparison of methods of measuring the performance of investment funds portfolios.....	298
Agnieszka Przedborska, Małgorzata Misztal , Selected multivariate statistical analysis methods in the evaluation of efficacy of deep electromagnetic stimulation in patients with degenerative joint disease	307
Wojciech Roszka, Marcin Szymkowiak , A calibration approach in statistical data integration	315
Iwona Skrodzka , Application of some methods of classification to the analysis of human capital in the European Union.....	325
Agnieszka Stanimir , Multivariate analysis of social inclusion factors.....	333
Dorota Strózik, Tomasz Strózik , Spatial differentiation of the standard of living in Great Poland Voivodeship	342
Izabela Szamrej-Baran , Identification of fuel poverty causes in Poland using soft modelling	352
Janusz Tuchowski, Katarzyna Wójcik , Classification of objects in the National Classification Framework described by the ontology.....	360
Aleksandra Matuszewska-Janica , Clustering of European Union states taking into consideration the levels of feminization of economic sectors..	368
Monika Rozkrut, Dominik Rozkrut , Identification of service sector innovation strategies in Poland.....	379

**Agnieszka Kozera, Joanna Stanisławska,
Romana Głowicka-Wołoszyn**

Uniwersytet Przyrodniczy w Poznaniu

SEGMENTACJA GOSPODARSTW DOMOWYCH WEDŁUG WYDATKÓW NA TURYSTYKĘ ZORGANIZOWANĄ

Streszczenie: W artykule przedstawiono wyniki badań dotyczące segmentacji gospodarstw domowych ze względu na posiadanie w budżecie rodziny wydatków na turystykę zorganizowaną oraz ich poziom. Jako narzędzie segmentacji zaproponowano drzewa klasyfikacyjne typu CART. Podstawę informacyjną badań stanowiły nieidentyfikowalne dane jednostkowe o dochodach i wydatkach gospodarstw domowych, pochodzące z badań budżetów gospodarstw domowych przeprowadzonych w 2010 r. przez Główny Urząd Statystyczny.

Słowa kluczowe: turystyka zorganizowana, gospodarstwa domowe, drzewa klasyfikacyjne typu CART.

1. Wstęp

Rola i społeczno-ekonomiczne znaczenie turystyki, przejawiające się między innymi w zdolności do tworzenia nowych miejsc pracy, podnoszenia jakości życia ludności, sprawiają, że wzrasta zapotrzebowanie na informacje, które pomogłyby podejmować optymalne decyzje dotyczące jej rozwoju [Dziedzic 2010, s. 8]. Zwłaszcza że pomimo występującego w kraju od 2009 r. spowolnienia gospodarczego i trudności, z jakimi zmagają się branża turystyczna (m.in. niestabilna sytuacja polityczna w Afryce Północnej, wybuch wulkanu na Islandii), pomiędzy latami 2005 i 2010 zaobserwowano w Polsce wzrost zainteresowania turystyką zorganizowaną (głównie zagraniczną), który przejawia się zarówno wzrostem uczestnictwa, jak również wydatków przeznaczanych na ten cel przez gospodarstwa domowe [Głowicka-Wołoszyn i in. 2013, s. 253, 262] i prognozuje się dalszy rozwój tego typu usług. Identyfikacja gospodarstw domowych, które zrealizowały popyt na usługi w zakresie turystyki zorganizowanej¹ i poznanie czynników społeczno-

¹ Jako zrealizowany popyt na turystykę zorganizowaną przyjmowana jest w pracy suma deklarowanych przez gospodarstwo domowe opłat za wycieczki zorganizowane, wczasy, obozy, zimowiska, kolonie dla dzieci i młodzieży (w kraju i za granicą) [*Budżety gospodarstw domowych w 2010 r.*, GUS, s. 258].

-ekonomicznych determinujących rzeczywisty popyt turystyczny stanowić może cenne źródło informacji dla podmiotów sfery turystyki i rekreacji.

Celem pracy było przeprowadzenie segmentacji gospodarstw domowych ze względu na fakt posiadania w budżecie rodziny² wydatków na usługi w zakresie turystyki zorganizowanej oraz ich poziom. Do wskazania głównych determinant społeczno-ekonomicznych tego typu popytu, które dzielą gospodarstwa domowe na najbardziej jednorodne klasy, zastosowano metodę drzew klasyfikacyjnych. Zastosowana metoda pozwoliła na określenie profili (segmentów) gospodarstw domowych uczestniczących w turystyce zorganizowanej poprzez wielowymiarową analizę zjawiska.

Podstawę informacyjną badań stanowiły nieidentyfikowalne dane jednostkowe pochodzące z badań budżetów gospodarstw domowych przeprowadzonych w 2010 r. przez GUS.

2. Metodyka badań

Do identyfikacji gospodarstw domowych, które odnotowały popyt na usługi w zakresie turystyki zorganizowanej zastosowano metodę drzew klasyfikacyjnych. Umożliwia ona hierarchiczny podział wielowymiarowej przestrzeni cech na maksymalne jednorodne podzbiory z punktu widzenia przyjętej zmiennej objaśnianej. Znajduje ona zastosowanie do klasyfikacji obiektów wówczas, gdy w zbiorze zmiennych można wyróżnić zmienną objaśnianą mierzoną na skali nominalnej lub porządkowej, a zmienne objaśniające mierzone są zarówno na słabych, jak i na mocnych skalach [Gatnar, Walesiak 2004]. Metoda drzew klasyfikacyjnych jest również odporna na występowanie wartości nietypowych [Gatnar 2008, s. 45] i jest elastycznym narzędziem do analizy zbiorów obserwacji, w których charakter zmiennych ogranicza możliwość zastosowania wielu narzędzi statystycznych.

Proces budowy drzewa klasyfikacyjnego jest wieloetapowy, bowiem na każdym etapie analizuje się wszystkie predyktory, tj. zmienne objaśniające i wybiera ten, który zapewnia najlepszy podział węzła, czyli wydziela najbardziej homogeniczne podzbiory. Początek budowy drzewa typu CART stanowi podział zbioru wszystkich obserwacji na dwa podzbiory. Dzielony zbiór nosi nazwę węzła macierzystego, natomiast wydzielone podzbiory – węzłów potomnych [Łapczyński 2010, s. 20]. Drzewa klasyfikacyjne i regresyjne są graficzną prezentacją modelu postaci [Gatnar 2008]:

$$Y = f(x_i) = \sum_{k=1}^K \alpha_k I(x_i \in R_k),$$

gdzie: Y – zmienna objaśniana,
 K – liczba segmentów,

² W pracy pojęcia „gospodarstwo domowe” i „rodzina” stosowane będą zamiennie.

R_k (dla $k = 1, 2, \dots, K$) – podprzestrzenie (segmenty) przestrzeni zmiennych objaśniających,

$x_i = [x_{i1}, x_{i2}, \dots, x_{iL}]$ – obserwacje ze zbioru rozpoznawalnego,

α_k – parametry modelu,

I – funkcja wskaźnikowa definiowana zależnie od charakteru zmiennych objaśniających.

Jeżeli zmienne mają charakter niemetryczny, jak w niniejszej pracy, to:

$$I(x_i \in R_k) = \prod_{l=1}^L I(x_{il} \in B_{kl}),$$

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

W opracowaniu zmienną poddaną analizie był poziom wydatków przeznaczonych na usługi związane z turystyką zorganizowaną w przeliczeniu na jednego członka rodziny przy zastosowaniu oryginalnej skali ekwiwalentnej OECD. Do wydatków tych zostały zaliczone, zgodnie z terminologią GUS, wydatki na turystykę zorganizowaną w kraju, za granicą i dla dzieci [*Metodologia...* 2011, s. 56-57, 106]. W przeprowadzonych badaniach za pomocą drzew klasyfikacyjnych określono profile gospodarstw domowych, które zrealizowały popyt na usługi w zakresie turystyki zorganizowanej (I część badania) oraz ze względu na poziom wydatków przeznaczanych na ten cel (II część badania):

– w I części badania zmienną Y zdefiniowano następująco:

$$y_{i(1)} = \begin{cases} \text{nie} & \text{– jeśli w budżecie } i\text{-tego gospodarstwa domowego nie odnotowano} \\ & \text{wydatków na turystykę zorganizowaną,} \\ \text{tak} & \text{– jeśli w budżecie } i\text{-tego gospodarstwa domowego odnotowano wy-} \\ & \text{datki na turystykę zorganizowaną.} \end{cases}$$

– w II części badania zmienną Y zdefiniowano następująco:

$$y_{i(2)} = \begin{cases} \text{niski} & \text{– jeśli wydatki na turystykę zorganizowaną przypadające na osobę} \\ & \text{w } i\text{-tym gospodarstwie domowym są niższe lub równe medianie,} \\ \text{wysoki} & \text{– jeśli wydatki na turystykę zorganizowaną przypadające na osobę} \\ & \text{w } i\text{-tym gospodarstwie domowym są wyższe od wartości mediany.} \end{cases}$$

W analizach wykorzystano sześć zmiennych objaśniających (predyktorów), tj. płeć, wiek i wykształcenie głowy odniesienia, wielkość gospodarstwa domowego, przynależność do grupy społeczno-ekonomicznej ludności oraz klasę miejscowości zamieszkania.

W artykule do budowy drzew klasyfikacyjnych wykorzystano algorytm CART. Obliczenia dokonano w programie SPSS 20. Wśród reguł podziału wybrano najczęściej stosowany – indeks Giniego [Łapczyński 2010, s. 58-59]. Jedną ze strategii

właściwej budowy modelu jest tzw. przycinanie drzewa, która przeciwdziała nadmiernemu rozrostowi drzewa i tym samym zmniejsza trudności w interpretacji wyników [Misztal 2000]. W przeprowadzonych analizach założono, że węzły końcowe nie mogą być mniejsze niż 5% badanej zbiorowości. W efekcie otrzymano względnie prosty model z niewielką liczbą węzłów końcowych o niezbyt dużej głębokości. Przy wyborze optymalnego drzewa brano pod uwagę wartości kosztów sprawdzianu krzyżowego oraz kosztów redystrybucji na tle złożoności drzewa. Kryterium wyboru stanowiła najmniejsza różnica między kosztem sprawdzianu krzyżowego a kosztem redystrybucji. Do oceny jakości podziału przestrzeni zmiennych objaśniających wykorzystano dla zmiennej objaśnianej błąd klasyfikacji.

3. Wyniki badań (I) – segmentacja gospodarstw domowych ze względu na odnotowany popyt na usługi w zakresie turystyki zorganizowanej

Ze względu na rzeczywisty popyt na usługi w zakresie turystyki zorganizowanej w 2010 r. gospodarstwa domowe podzielono na dwie grupy. Pierwszą grupę utworzyły te, które w budżetach nie odnotowały wydatków na turystykę zorganizowaną (89,5%), drugą grupę natomiast rodziny, które poniosły wydatki na ten cel (10,5%). Znaczna część gospodarstw domowych nie odnotowała popytu na turystykę zorganizowaną. Wydatki na ten cel zaliczane są do dóbr i usług wyższego rzędu, dlatego też ich brak w budżecie domowym mógł być związany ze słabą sytuacją materialną gospodarstw domowych, ale mogły one również nie należeć do ich koszyka dóbr preferowanych.

Spośród zmiennych objaśniających w ostatecznym modelu (po przycięciu drzewa) uwzględnione zostały cztery, tj. przynależność do grupy społeczno-ekonomicznej ludności, wiek i wykształcenie głowy odniesienia oraz liczba osób w rodzinie. Okazały się one najlepszymi predyktorami³ zrealizowanego przez gospodarstwa domowe popytu na turystykę zorganizowaną w 2010 roku (tab. 1). Zbudowany model (I) jest dobrej jakości, bowiem na jego podstawie poprawnie zakwalifikowano 65% gospodarstw domowych.

Z analizy struktury drzewa klasyfikacyjnego (I) oraz rankingu ważności predyktorów wynika (tab.1, rys. 1), że największe znaczenie w podziale obserwacji miała przynależność gospodarstwa domowego do grupy społeczno-ekonomicznej ludności. Z przeprowadzonych badań wynika, że gospodarstwa pracowników na

³ Na każdym etapie podziału drzewa dla zmiennych objaśniających obliczane są wartości tzw. wskaźnika poprawy, na podstawie którego tworzone są ich rankingi lokalne. Po zbudowaniu modelu drzewa sumuje się wartości tego wskaźnika dla wszystkich zmiennych objaśniających z poszczególnych etapów podziału, co stanowi podstawę do stworzenia ostatecznego (globalnego) rankingu ważności predyktorów [Łapczyński 2010]. Jakość wyjaśnienia zmiennej objaśnianej przedstawiona jest dla każdego predyktora na 100 punktowej skali (znormalizowana ważność).

stanowiskach nierobotniczych oraz pracujących na własny rachunek charakteryzowały się wyższym prawdopodobieństwem posiadania w budżecie domowym wydatków na turystykę zorganizowaną w stosunku do pozostałych grup społeczno-ekonomicznych ludności, wyniosło ono 17,4% (I węzeł) (rys. 1). Struktura zawodowa społeczeństwa implikuje więc wielkość popytu na usługi turystyczne. Decyduje ona bowiem o dochodach i statusie społecznym, które wpływają bezpośrednio na aktywność turystyczną społeczeństwa.

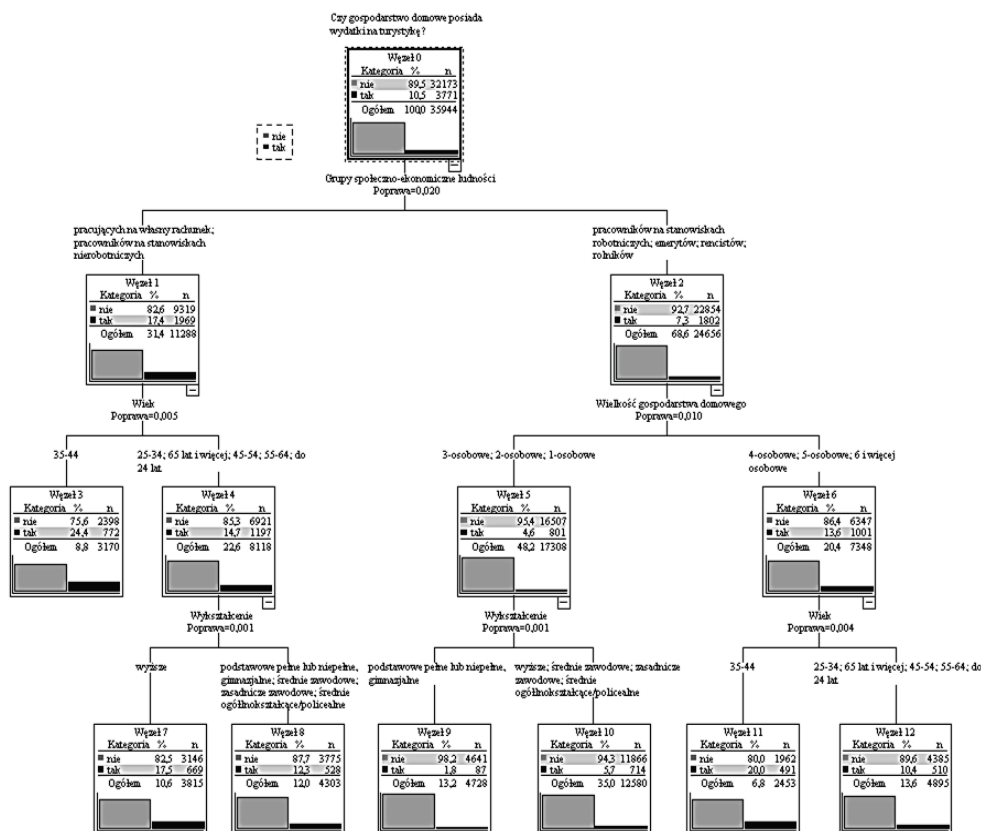
Tabela 1. Ranking ważności predyktorów występowania w budżetach gospodarstw domowych wydatków na turystykę zorganizowaną w Polsce w 2010 roku

Predyktory	Ważność	Znormalizowana wartość
Grupy społeczno-ekonomiczne ludności	0,025	100,0%
Wiek	0,016	63,6%
Wykształcenie	0,014	55,6%
Wielkość gospodarstwa domowego	0,011	42,0%
Klasa miejscowości zamieszkania	0,003	11,5%
Płeć	0,000	0,0%

Źródło: obliczenia własne na podstawie niepublikowanych danych jednostkowych, pochodzących z badań budżetów gospodarstw domowych, przeprowadzonych przez GUS w 2010 roku.

Z analizy drzewa klasyfikacyjnego (I) (rys. 1) wynika ponadto, że gospodarstwa pracowników na stanowiskach nierobotniczych i pracujących na własny rachunek, w których głowa odniesienia była w wieku 35-44 lata, odnotowały popyt na usługi w zakresie turystyki zorganizowanej z najwyższym prawdopodobieństwem, które wyniosło 24,4% (XIII węzeł końcowy). Najwięcej podróżują rodziny prowadzone przez osoby młode, które mają ustabilizowaną sytuację na rynku pracy. Gospodarstwa domowe, w których osoba odniesienia była w wieku 35-44 lata, posiadają zazwyczaj niepełnoletnie dzieci, co także może być dodatkową przesłanką do zorganizowanych wyjazdów turystycznych. Turystyka zorganizowana jest bowiem postrzegana jako wygodna i bezpieczna forma spędzania wolnego czasu.

Gospodarstwa pracowników na stanowiskach robotniczych, emerytów, rencistów i rolników w 2010 roku charakteryzowały się słabszą sytuacją materialną, która mogła mieć wpływ na zrealizowany przez nie popyt na usługi w zakresie turystyki zorganizowanej. Gospodarstwa te cechowały się bowiem niższym prawdopodobieństwem posiadania w budżecie rodziny wydatków na tę grupę potrzeb w stosunku do rodzin pracowników na stanowiskach nierobotniczych i pracujących na własny rachunek, wyniosło ono 7,3% (II węzeł) (rys. 1). Sytuacja ta była jednak zróżnicowana w zależności od liczby osób w rodzinie oraz wykształcenia i wieku głowy odniesienia. Najniższym prawdopodobieństwem posiadania w budżecie rodziny wydatków na turystykę zorganizowaną charakteryzowały się gospodarstwa małe (do trzech osób w rodzinie), pracowników na stanowiskach robotniczych, rolników, emerytów i rencistów, w których głowa odniesienia miała niski poziom



Rys. 1. Schemat drzewa klasyfikacyjnego (I) – występowanie w budżetach gospodarstw domowych wydatków na turystykę zorganizowaną

Źródło: obliczenia własne na podstawie niepublikowanych danych jednostkowych, pochodzących z badań budżetów gospodarstw domowych, przeprowadzonych przez GUS w 2010 roku.

wykształcenia (1,8%, XIX węzeł końcowy). Wśród gospodarstw pracowników na stanowiskach robotniczych, emerytów, rencistów i rolników najwyższym popytem na usługi w zakresie turystyki zorganizowanej charakteryzowały się gospodarstwa większe, liczące cztery osoby i więcej, w których głowa rodziny była w wieku 35-44 lata (20%; XI węzeł końcowy). Może to być związane z występowaniem w tej grupie rodzin relatywnie wysokiego popytu na usługi w zakresie turystyki zorganizowanej dla dzieci (np. kolonie, obozy wypoczynkowe). Wielodzietne gospodarstwa domowe, które charakteryzują się słabszą sytuacją materialną, korzystają często z różnych form zorganizowanych wyjazdów turystycznych dla dzieci i młodzieży, które w znacznej części są finansowane między innymi przez instytucje państwowe i organizacje pozarządowe.

4. Wyniki badań (II) – segmentacja gospodarstw domowych ze względu na poziom wydatków na turystykę zorganizowaną

W celu identyfikacji grup gospodarstw domowych ze względu na poziom ich wydatków na turystykę zorganizowaną brano pod uwagę tylko rodziny, które w budżetach odnotowały daną grupę wydatków. Spośród zmiennych objaśniających w ostatecznym modelu uwzględnione zostały trzy zmienne, tj. wielkość gospodarstwa domowego, wykształcenie głowy odniesienia oraz klasa miejscowości zamieszkania. Zmienne te okazały się najlepszymi predyktorami stopnia zaspokojenia potrzeb rodzin w zakresie turystyki zorganizowanej (tab. 2). Zbudowany model (II) jest dobrej jakości, bowiem na jego podstawie (rys. 2) poprawnie zakwalifikowano blisko 70% badanych gospodarstw domowych.

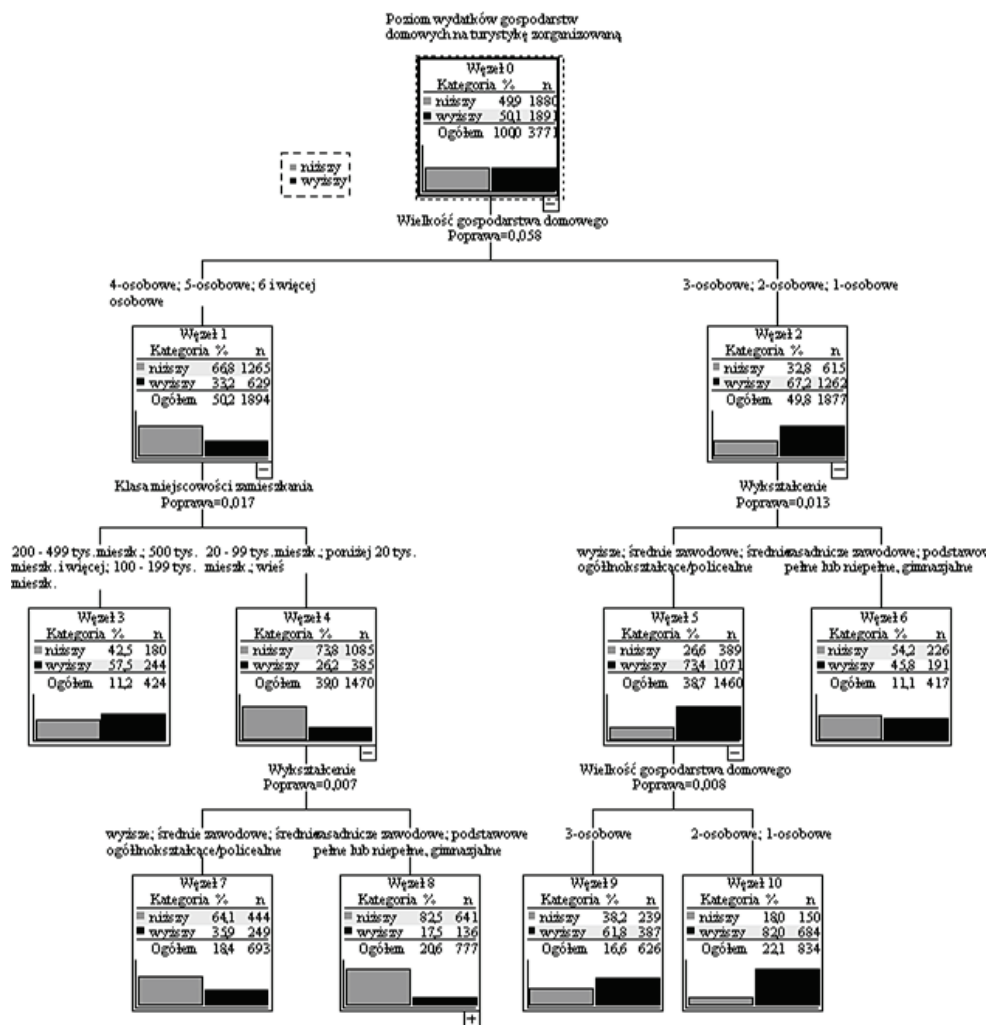
Tabela 2. Ranking ważności predyktorów poziomu wydatków gospodarstw domowych na turystykę zorganizowaną w 2010 roku

Predyktory	Ważność	Znormalizowana wartość
Wielkość gospodarstwa domowego	0,073	100,0%
Wykształcenie	0,071	98,1%
Klasa miejscowości zamieszkania	0,063	86,3%
Grupy społeczno-ekonomiczne ludności	0,040	54,4%
Wiek	0,012	15,9%
Płeć	0,002	2,2%

Źródło: obliczenia własne na podstawie niepublikowanych danych jednostkowych, pochodzących z badań budżetów gospodarstw domowych, przeprowadzonych przez GUS w 2010 r.

Zbudowany model drzewa klasyfikacyjnego (II) obejmował sześć węzłów końcowych, na podstawie których można było określić różne profile gospodarstw domowych, różniących się poziomem wydatków przeznaczanych na turystykę zorganizowaną (rys. 2). Z analizy struktury drzewa klasyfikacyjnego (II) oraz rankingu ważności predyktorów (tab. 2) wynika, że największe znaczenie w podziale obserwacji na grupy miała liczba osób w rodzinie. Wielkość gospodarstwa domowego implikuje w znacznym stopniu poziom i strukturę nabywanych przez nie dóbr i usług, wpływa na powstanie efektów skali gospodarowania, a z drugiej strony kombinacja cech demograficznych jego członków wyznacza hierarchię potrzeb konsumpcyjnych.

Z przeprowadzonych badań wynika, że wśród gospodarstw domowych, które odnotowały popyt na usługi w zakresie turystyki zorganizowanej, gospodarstwa małe, tj. jedno-, dwu- i trzyosobowe, charakteryzowały się wysokim poziomem wydatków przypadających na jednostkę ekwiwalentną, przeznaczanych na ten cel z prawdopodobieństwem wynoszącym 67,3% (II węzeł). Niskim poziomem wydatków na turystykę zorganizowaną cechowały się natomiast gospodarstwa wieloosobowe, tj. czteroosobowe i liczniejsze, z prawdopodobieństwem 66,8% (I wę-



Rys. 2. Schemat drzewa klasyfikacyjnego (II) – poziom wydatków gospodarstw domowych na turystykę zorganizowaną

Źródło: obliczenia własne na podstawie niepublikowanych danych jednostkowych, pochodzących z badań budżetów gospodarstw domowych, przeprowadzonych przez GUS w 2010 r.

zeń). Z analizy struktury drzewa klasyfikacyjnego (II) wynika, że wśród gospodarstw małych (jedno- i dwuosobowych) najwyższe prawdopodobieństwo wystąpienia w budżetach domowych wysokiego poziomu wydatków na turystykę zorganizowaną (82%) charakteryzowało te, które prowadzone były przez osoby reprezentujące średni lub wyższy poziom wykształcenia (X węzeł końcowy). W gospodarstwach małych, w których głowa rodziny posiadała wykształcenie podstawowe lub zasadnicze zawodowe prawdopodobieństwo to było niższe i wyniosło niespeł-

na 46% (XV węzeł końcowy). Można to tłumaczyć tym, że wzrost poziomu wykształcenia implikuje wzrost poziomu przeciętnego miesięcznego dochodu rozporządzalnego oraz przeciętnych wydatków przypadających na jedną osobę w gospodarstwie domowym. Wśród gospodarstw wieloosobowych najwyższe prawdopodobieństwo wystąpienia wysokiego poziomu wydatków na turystykę zorganizowaną charakteryzowało te rodziny, które zamieszkiwały w większych miastach, powyżej 100 tys. mieszkańców (57,5%; XIII węzeł końcowy). Miejsce zamieszkania jest bowiem czynnikiem silnie różnicującym aktywność turystyczną. Z badań przeprowadzonych przez GUS [*Turystyka...* 2010, s. 48] wynika, że mieszkańcy miast charakteryzują się większą aktywnością turystyczną niż mieszkańcy wsi.

Wśród gospodarstw domowych, które odnotowały popyt na usługi w zakresie turystyki zorganizowanej, gospodarstwa wieloosobowe (liczące cztery osoby i więcej), w których głowa odniesienia miała niski poziom wykształcenia, zamieszkujące wsie i mniejsze ośrodki miejskie (poniżej 100 tys. mieszkańców) charakteryzowały się niskim poziomem wydatków przeznaczanych na ten cel z prawdopodobieństwem wynoszącym 82,5% (VIII węzeł końcowy). Wynikać to może z słabszej sytuacji finansowej tych gospodarstw domowych. Z tego względu mogą one preferować inne formy wypoczynku i spędzania wolnego czasu.

5. Wnioski

Wykorzystanie drzew klasyfikacyjnych typu CART pozwoliło na wydzielenie segmentów gospodarstw domowych, które różniły się między sobą pod względem rzeczywistego popytu na turystykę zorganizowaną oraz poziomu wydatków przeznaczanych na ten cel. Z przeprowadzonej analizy wynika, że w roku 2010 popyt na usługi w zakresie turystyki zorganizowanej w znacznym stopniu uzależniony był od przynależności gospodarstwa domowego do grupy społeczno-ekonomicznej ludności, liczby osób w rodzinie oraz wieku głowy odniesienia. Popyt na te usługi najczęściej występował w gospodarstwach pracowników na stanowiskach nierobotniczych oraz pracujących na własny rachunek, w których głowa rodziny była w wieku 35-44 lata, natomiast rzadko występował w gospodarstwach małych (do trzech osób w rodzinie) pracowników na stanowiskach robotniczych, rolników, emerytów i rencistów, w których głowa odniesienia posiadała niski poziom wykształcenia.

Wśród gospodarstw domowych, które odnotowały popyt na usługi w zakresie turystyki zorganizowanej w 2010 roku, poziom wydatków na ten cel uzależniony był głównie od liczby osób w rodzinie, wykształcenia głowy odniesienia oraz klasy miejscowości zamieszkania. Gospodarstwa małe (do trzech osób w rodzinie), w których głowa odniesienia posiadała wyższy poziom wykształcenia, charakteryzowały się wysokim poziomem wydatków przeznaczanych na turystykę zorganizowaną. Natomiast niskim poziomem tych wydatków charakteryzowały się gospo-

darstwa wieloosobowe, zamieszkujące wieś i mniejsze miasta, w których głowa odniesienia reprezentowała niski poziom wykształcenia.

Uzyskane profile gospodarstw domowych mogą stanowić cenne źródło informacji dla podmiotów branży turystycznej, zainteresowanych przygotowaniem działań marketingowych i mogą stanowić podstawę do pogłębionych badań preferencji konsumpcyjnych.

Literatura

- Budżety gospodarstw domowych w 2010 roku* (2011), Główny Urząd Statystyczny, Warszawa.
- Dziedzic E. (2010), *Potrzeby i luki informacyjne u podmiotów zarządzających turystyką*, [w:] Dziedzic E. (red.), *Regionalne badania konsumentów usług turystycznych*, Polska Organizacja Turystyczna, Warszawa, s. 8.
- 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 Akademii Ekonomicznej we Wrocławiu, Wrocław.
- Głowicka-Wołoszyn R., Stanisławska J., Wołoszyn A., Kozera A. (2013), *Wydatki na turystykę zorganizowaną w gospodarstwach domowych i ich determinanty w Polsce w latach 2005 i 2010*, „Handel Wewnętrzny” nr 4 (2), Warszawa, s. 251-264.
- Kruczek W. (2008), *Turystyka*, Wydawnictwo Naukowe PWN, Warszawa.
- Łapczyński M. (2010), *Drzewa klasyfikacyjne i regresyjne w badaniach marketingowych*, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.
- Metodologia badania budżetów gospodarstw domowych* (2011), Departament Warunków Życia, Główny Urząd Statystyczny, Warszawa.
- Misztal M. (2000), *Wykorzystanie drzew klasyfikacyjnych do wspomagania procesów podejmowania decyzji*, StatSoft, Kraków.
- Turystyka i wypoczynek w gospodarstwach domowych w 2009 roku* (2010), Główny Urząd Statystyczny, Warszawa.

HOUSEHOLD SEGMENTATION WITH RESPECT TO THE EXPENDITURES ON ORGANIZED TOURISM

Summary: The article presents the study results of household segmentation with respect to the presence and the level of budgetary items for organized tourism. The segmentation was carried out using CART classification trees on anonymized microdata databases of household income and expenditures compiled in the *Household Budget Survey in 2010* by the Polish Central Statistical Office.

Keywords: organized tourism, households, classification and regression trees.