

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

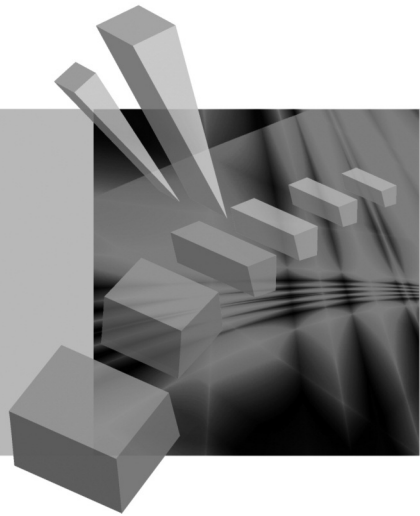
RESEARCH PAPERS

of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>
oraz w 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](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

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka, Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska, Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński, Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz, Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel, Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień, Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska, Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan, Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska, Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka, Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański, Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk, Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk, Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura, Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk, Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski, Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka, Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk, Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarz , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

Iwona Bąk

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

SEGMENTACJA GOSPODARSTW DOMOWYCH EMERYTÓW I RENCISTÓW POD WZGLĘDEM WYDATKÓW NA REKREACJĘ I KULTURĘ

Streszczenie: W artykule przedstawiono wyniki badań dotyczące segmentacji gospodarstw domowych emerytów i rencistów pod względem ich przeciętnych miesięcznych wydatków na rekreację i kulturę. Do klasyfikacji gospodarstw domowych wykorzystano drzewo regresyjne. Podstawę informacyjną badań stanowiły nieidentyfikowalne jednostkowe dane o dochodach i wydatkach indywidualnych gospodarstw domowych, pochodzące z badań budżetów gospodarstw domowych przeprowadzonych w 2009 r. przez GUS. Segmentacja została przeprowadzona dla reprezentatywnej próby 1308 gospodarstw domowych emerytów i 327 gospodarstw domowych rencistów.

Słowa kluczowe: gospodarstwa domowe emerytów i rencistów, segmentacja, drzewa regresyjne.

1. Wstęp

Procesowi starzenia się osób wchodzących w skład gospodarstwa domowego towarzyszy ograniczenie niektórych potrzeb i pojawienie się nowych (np. w zakresie ochrony zdrowia). Może także mieć miejsce ponowne rozbudzenie istniejących wcześniej potrzeb, np. w zakresie rekreacji i kultury, których zaspokojenie było utrudnione ze względu na wykonywanie pracy zawodowej. Uczestnictwo w rekreacji oraz kulturze można pokazać przez pryzmat wielkości wydatków przeznaczonych przez gospodarstwa domowe na dobra i usługi z tej dziedziny. Gospodarstwa, które przeznaczają więcej pieniędzy na różne typy dóbr i usług rekreacyjno-kulturalnych, intensywniej uczestniczą w kulturze i rekreacji niż te, w których wydaje się ich mniej. Dane statystyczne o przeciętnych miesięcznych wydatkach przypadających na 1 osobę w gospodarstwie domowym można m.in. uzyskać z badań Budżetów Gospodarstw Domowych prowadzonych przez GUS. Wydatki na rekreację i kulturę zbudowane są z 46 rodzajów wydatków. Wśród nich znajdują się m.in. wydatki związane ze sprzętem służącym rekreacji i kulturze, sprzętem turystycznym oraz turystyką zorganizowaną [*Budżety gospodarstw...* 2010].

Celem artykułu jest segmentacja gospodarstw domowych emerytów i rencistów w Polsce ze względu na poziom ich przeciętnych miesięcznych wydatków na rekreację i kulturę. Do klasyfikacji gospodarstw domowych wykorzystane zostanie drzewo regresyjne. Podstawę informacyjną badań będą stanowiły nieidentyfikowalne jednostkowe dane o dochodach i wydatkach indywidualnych gospodarstw domowych, pochodzące z badań budżetów gospodarstw domowych przeprowadzonych w 2009 r. przez GUS. Segmentacja zostanie przeprowadzona dla reprezentatywnej próby 1308 gospodarstw domowych emerytów i 327 gospodarstw domowych rencistów, którzy uczestniczyli w badaniu „Turystyka i wypoczynek w gospodarstwach domowych” przeprowadzonym na podpróbce badania budżetów gospodarstw domowych.

2. Wydatki gospodarstw domowych emerytów

O sytuacji materialnej gospodarstwa domowego w znacznej mierze decyduje dochód rozporządzalny. Przeciętny miesięczny dochód rozporządzalny przypadający na jedną osobę w gospodarstwie domowym w Polsce wynosił w 2009 r. 1114,49 zł. Najbliższe przeciętnemu poziomowi były dochody w gospodarstwach pracowników (1123,30 zł). Najwyższy poziom dochodu osiągały gospodarstwa pracujących na własny rachunek (1396,47 zł), najniższy zaś gospodarstwa rencistów (870,55 zł). Stosunkowo wysokim dochodem charakteryzują się gospodarstwa emerytów (1180,65 zł). Analizując dochody gospodarstw emerytów, należy jednak pamiętać, że przeciętna liczba osób w tych gospodarstwach jest niższa od średniej liczby osób w pozostałych grupach społeczno-ekonomicznych¹. Nie pozostaje to bez wpływu na strukturę wydatków na poszczególne produkty i usługi [*Budżety gospodarstw...* 2010].

Od dochodu, jakim rozporządza gospodarstwo domowe, zależy ściśle poziom i struktura wydatków. Struktura wydatków jest źródłem dość dobrych informacji o społeczno-ekonomicznym położeniu gospodarstw [Podolec i in. 2008, s. 86-88]. Często stosowanym miernikiem jest udział wydatków na żywność w wydatkach ogółem. Wydatki związane z zakupem żywności stanowią największą część wydatków we wszystkich grupach społeczno-ekonomicznych, niezależnie od poziomu wydatków konsumpcyjnych ogółem. Przeznaczono na nie od 20,74% wydatków w gospodarstwach pracujących na własny rachunek do 33,16% w gospodarstwach rolników. Udział wydatków na użytkowanie mieszkania i nośniki energii w wydatkach konsumpcyjnych jest także zróżnicowany. Najwyższy udział wydatków przypada na gospodarstwa rencistów i emerytów². Te grupy gospodarstw przeznaczają także najwięcej na ochronę zdrowia, przy czym prym wiodą tutaj gospodarstwa emerytów.

¹ Przeciętna liczba osób w gospodarstwach domowych emerytów wynosiła 2,07 osoby, natomiast w gospodarstwach domowych ogółem było przeciętnie 2,95 osoby [*Emerytura...* 2010, s. 30].

² Większy udział wydatków na eksploatację mieszkania wiąże się z generalnie mniejszym zagęszczeniem w lokalach zamieszkiwanych przez emerytów i rencistów.

Gospodarstwa emerytów charakteryzują się także najniższym spośród pozostałych gospodarstw udziałem średnich wydatków na łączność i edukację.

Wydatki na rekreację i kulturę stanowią istotną pozycję w wydatkach gospodarstw domowych. W 2009 r. stanowiły one czwartą pozycję pod względem wielkości udziału (po wydatkach na żywność i napoje bezalkoholowe, użytkowanie mieszkania i nośniki energii, transport) we wszystkich grupach społeczno-ekonomicznych, niezależnie od poziomu wydatków konsumpcyjnych ogółem. Przeznaczano na nie od 5,31% w gospodarstwach rencistów do 10,02% wydatków w gospodarstwach pracujących na własny rachunek. Gospodarstwa domowe emerytów przeznaczają na rekreację i kulturę coraz więcej środków finansowych, udział tych wydatków w wydatkach ogółem systematycznie wzrastał od 4,9 w 2000 r. do 6,1% w roku 2009 [Budżety gospodarstw... 2010].

3. Istota drzew regresyjnych

Drzewa 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, która jest mierzona na skalach mocnych (przedziałowa, ilorazowa),

2) zmienne niezależne mogą być mierzone zarówno na skalach słabych (nominalna, porządkowa), jak i na skalach mocnych.

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

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

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

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

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

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, \dots, X_L mają charakter niemetryczny, to podprzestrzeń R_k można zdefiniować jako

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

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

Punkty podziału $v_{kl}^{(d)}$, $v_{kl}^{(g)}$ lub zbiory łączonych wartości zmiennych nominalnych (B_{kl}) są znajdowane w ten sposób, by dokonany podział poprawił jakość modelu.

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:

$$\alpha_k = \frac{1}{N(k)} \sum_{\mathbf{x}_i \in R_k} y_i, \quad (4)$$

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 wariancję zmiennej zależnej³.

4. Segmentacja gospodarstw domowych z wykorzystaniem drzew regresyjnych

Do segmentacji gospodarstw domowych wykorzystano drzewa regresyjne. Zmienną zależną zdefiniowano jako łączne miesięczne wydatki poniesione przez gospodarstwo domowe na rekreację i kulturę (wartości od 0 zł do 1087 zł dla gospodarstw domowych emerytów i od 0 zł do 1233 zł dla gospodarstw domowych rencistów), natomiast zbiór zmiennych niezależnych tworzyły:

- predyktory jakościowe: płeć (kobieta, mężczyzna) i wykształcenie głowy gospodarstwa domowego (co najwyżej 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),
- predyktory ilościowe: wiek głowy gospodarstwa domowego, przeciętny miesięczny dochód rozporządzalny gospodarstwa, liczba osób w gospodarstwie domowym, liczba dzieci w wieku do 14 lat włącznie, liczba osób pracujących, liczba osób bezrobotnych.

³ 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 wyznaczenia drzew regresyjnych wykorzystano dwie procedury: CART i CHAID oprogramowane w pakiecie *Statistica 8.0*⁴. Przystępując do budowy drzew, przyjęto założenia zaprezentowane w tab. 1.

Tabela 1. Założenia przyjęte przy wyznaczaniu drzew regresyjnych

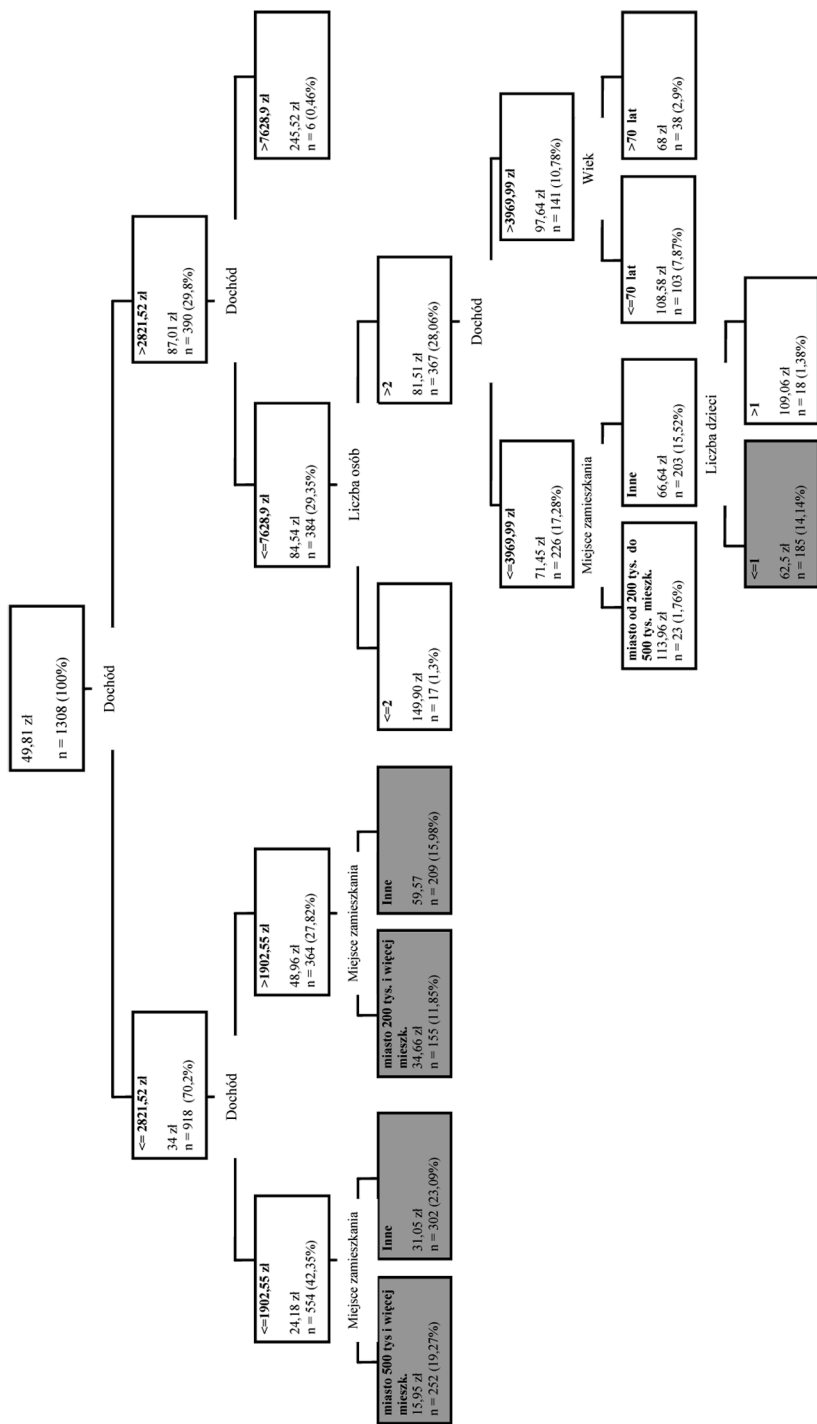
Założenia przyjęte w procedurze	Procedura CART		Procedura CHAID	
	model standardowy	drzewo interakcyjne*	model standardowy	drzewo interakcyjne
Kryterium stopu	przytnij według wariancji		–	
Minimalna liczność	30	30	30	30
Minimalna liczność potomków	–	30	–	30
Maksymalna liczba poziomów	–	10	10	10
Maksymalna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

* Korzystanie z drzew interakcyjnych pozwala na ręczne sterowanie wielkością i głębokością drzewa.

Źródło: opracowanie własne.

Wykorzystując wspomniane procedury, przy uwzględnieniu założeń z tab. 1 otrzymano drzewa regresyjne o różnej strukturze. Wyniki struktur drzew regresyjnych dla gospodarstw domowych emerytów zaprezentowano w tab. 2. Za najlepsze uznano drzewo regresyjne otrzymane z wykorzystaniem procedury CART model standardowy, które przedstawiono na rys. 1. Było to drzewo o numerze 57 wybrane z sekwencji 67 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 scharakteryzowanie segmen-

⁴ W metodzie CHAID wybierany jest taki podział, który daje największą istotność sprawdzianu testu χ^2 liczonego dla tablicy kontyngencji, opisującej podział przestrzeni na segmenty zawierające podzbiory zbioru uczącego. Na każdym etapie podziału drzewa tworzy się tablicę kontyngencji, w której zestawia się zmienną zależną i predyktor, a następnie dąży się do redukcji tabeli kontyngencji przez łączenie w dozwolony sposób kategorii predyktora. Do oceny istotności statystycznej podziału wykorzystywana jest nierówność Bonferroniego [Gatnar, Walesiak 2004]. Cechą charakterystyczną metody CART jest nadmierny rozrost drzewa i przycinanie poszczególnych gałęzi w celu redukcji opisu liści (przy nieznacznym wzroście błędu klasyfikacji). Pozwala to na porównanie modelu rozbudowanego i modelu ze zredukowaną liczbą węzłów. Sprawdza się, jaka jest różnica między błędem klasyfikacji całego drzewa a błędem klasyfikacji drzewa z usuniętą gałęzią i wybiera się najmniejszą różnicę. Drugą ważną zaletą tego algorytmu jest jednoczesne zestawienie kosztu resubstytucji (współczynnika błędu obliczonego ze zbioru uczącego) ze współczynnikiem błędu obliczonym na zbiorze testowym.



Rys. 1. Drzewo regresyjne dla miesięcznych wydatków poniesionych na rekreację i kulturę w gospodarstwach domowych emerytów wyznaczone procedurą standardowy model CART (w nawiasach podano odsetki obliczone w stosunku do liczebności całej próby)

Źródło: opracowanie własne.

Tabela 2. Struktura drzew regresyjnych otrzymanych z wykorzystaniem procedury CART i CHAID dla gospodarstw domowych emerytów

Struktura drzewa	Procedura CART		Procedura CHAID	
	model standardowy	drzewo interakcyjne	model standardowy	drzewo interakcyjne
Liczba węzłów dzielonych	10	14	14	5
Kryterium podziału węzłów (od pierwszego do ostatniego węzła)	1) dochód 2) dochód 3) miejsce zamieszkania 4) liczba osób 5) dochód 6) miejsce zamieszkania 7) wiek 8) liczba dzieci	1) dochód 2) dochód 3) miejsce zamieszkania 4) miejsce zamieszkania 5) wiek 6) dochód 7) wiek 8) dochód 9) wykształcenie 10) dochód	1) dochód 2) dochód 3) miejsce zamieszkania 4) miejsce zamieszkania 5) wiek 6) dochód 7) wiek 8) dochód 9) wykształcenie 10) dochód	1) wykształcenie 2) dochód 3) dochód 4) typ gospodarstwa 5) miejsce zamieszkania
Liczba węzłów końcowych	11	15	15	11

Źródło: opracowanie własne.

Tabela 3. Charakterystyka segmentów gospodarstw domowych emerytów ze względu na ich wydatki na rekreację i kulturę

Nr	Charakterystyka segmentu	Przeciętne miesięczne wydatki na rekreację i kulturę (zł)	Liczebność segmentu (% badanej próby)
1	Gospodarstwa domowe mieszkające w miastach powyżej 500 tys. mieszkańców o miesięcznych dochodach nie wyższych niż 1902,60 zł	15,95	252 (19,3%)
2	Gospodarstwa domowe mieszkające na wsi lub w miastach do 499 tys. mieszkańców o miesięcznych dochodach nie wyższych niż 1902,60 zł	31,05	302 (23,1%)
3	Gospodarstwa domowe mieszkające w miastach powyżej 200 tys. mieszkańców o miesięcznych dochodach powyżej 1902,60 zł	34,66	155 (11,9%)
4	Gospodarstwa domowe mieszkające na wsi lub w miastach do 199 tys. mieszkańców o miesięcznych dochodach powyżej 1902,60 zł	59,57	209 (16,0%)
5	Gospodarstwa domowe posiadające co najwyżej jedno dziecko, mieszkające na wsi lub w miastach, z wyjątkiem miast od 200 do 499 tys. mieszkańców, o miesięcznych dochodach nie wyższych niż 3970 zł	62,51	185 (14,1%)

Źródło: opracowanie własne.

tów gospodarstw domowych ze względu na ich wydatki na rekreację i kulturę⁵. Ponadto przy wyborze segmentu przyjęto założenie, że jego liczebność powinna stanowić przynajmniej 10% liczebności próby. Warunek ten spełnia tylko pięć węzłów końcowych, które na rys. 1 zostały wyróżnione szarym kolorem. Wyniki segmentacji zamieszczono w tab. 3.

Tabela 4. Struktura drzew regresyjnych otrzymanych z wykorzystaniem procedury CART i CHAID dla gospodarstw domowych rencistów

Struktura drzewa	Procedura CART		Procedura CHAID	
	model tandardowy	drzewo interakcyjne	model standardowy	drzewo interakcyjne
Liczba węzłów dzielonych	8	2	0	0
Kryterium podziału węzłów (od pierwszego do ostatniego węzła)	1) dochód 2) dochód 3) dochód 4) dochód 5) dochód 6) wiek 7) wiek 8) miejsce zamieszkania	1) dochód 2) miejsce zamieszkania	–	–
Liczba węzłów końcowych	9	3	1	1

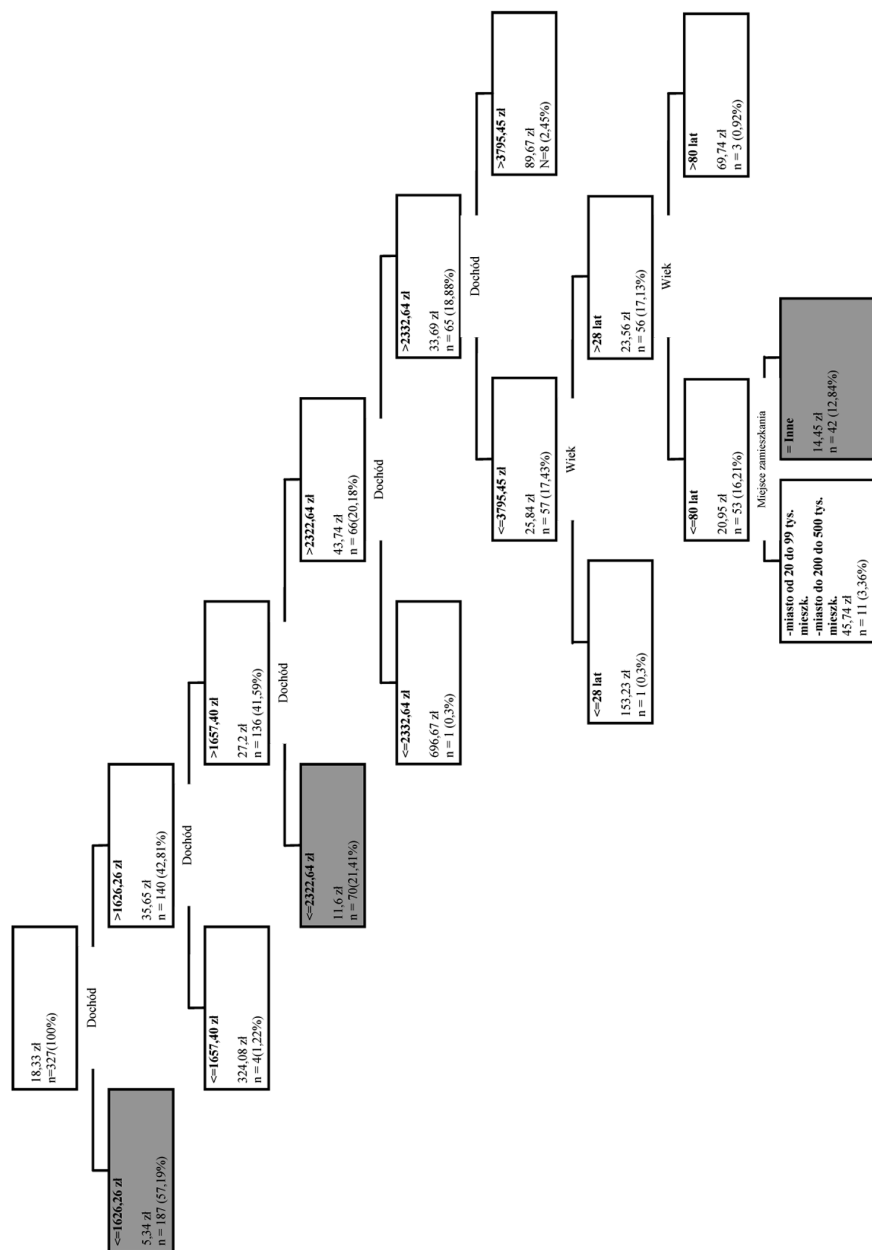
Źródło: opracowanie własne.

Tabela 5. Charakterystyka segmentów gospodarstw domowych rencistów ze względu na ich wydatki na rekreację i kulturę

Nr	Charakterystyka segmentu	Przeciętne miesięczne wydatki na rekreację i kulturę (zł)	Liczebność segmentu (% badanej próby)
1	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 1626,27 zł	5,34	187 (57,2%)
2	Gospodarstwa domowe o miesięcznych dochodach nie wyższych niż 2322,50 zł	11,60	70 (21,4%)
3	Gospodarstwa domowe, w których głowa mieszkańca jest w wieku do 80 lat, mieszkające na wsi lub w miastach: od 20 do 99 tys. mieszkańców, od 200 do 499 tys. mieszkańców i powyżej 500 tys. mieszkańców, o miesięcznych dochodach powyżej 3795,45 zł	14,45	42 (12,8%)

Źródło: opracowanie własne.

⁵ Rezygnując z pozostałych drzew, kierowano się tym, że węzły końcowe w tych drzewach były wyznaczone na podstawie małej liczby predyktorów (1-2), które powtarzały się przy różnych węzłach dzielonych.



Rys. 2. Drzewo regresyjne dla miesięcznych wydatków poniesionych na rekreację i kulturę w gospodarstwach domowych rencistów wyznaczone procedurą standardowy model CART (w nawiasach podano odsetki obliczone w stosunku do liczebności całej próby)

Źródło: opracowanie własne.

Wyniki struktur drzew regresyjnych dla gospodarstw domowych rencistów zaprezentowano w tab. 4. Za najlepsze uznano drzewo regresyjne otrzymane z wykorzystaniem procedury CART model standardowy. Było to drzewo o numerze 15 wybrane z sekwencji 19 drzew. Zaprezentowano je na rys. 2. W tabeli 5 zamieszczono wyniki segmentacji gospodarstw domowych rencistów pod względem ich miesięcznych wydatków na rekreację i kulturę.

5. Podsumowanie

Wykorzystanie drzew regresyjnych pozwoliło na wydzielenie segmentów gospodarstw domowych, które znacznie różniły się pod względem poziomu przeciętnych miesięcznych wydatków na rekreację i kulturę. Wysokość tych wydatków została najsilniej zdeterminowana przez dochody gospodarstw. Istotnymi predyktorami okazały się również: miejsce zamieszkania i wiek. Okazało się, że najliczniejszą grupę gospodarstw domowych emerytów (ponad 23%) stanowiły gospodarstwa wydające na rekreację i kulturę przeciętnie w ciągu miesiąca ok. 31 zł i są to gospodarstwa mieszkające na wsi lub w miastach do 499 tys. mieszkańców o miesięcznych dochodach nie wyższych niż 1902,60 zł. Natomiast większość gospodarstw domowych rencistów (ponad 57%) wydaje miesięcznie na rekreację i kulturę tylko 5,34 zł. Są to gospodarstwa o dochodach nieprzekraczających 1626,27 zł. Niska przeciętna kwota wydatków wynika z faktu, że większość gospodarstw z tej grupy w ogóle nie przeznaczają żadnych środków na rekreację i kulturę.

Uzyskane wyniki badań pozwalają poznać preferencje gospodarstw domowych emerytów i rencistów w zakresie przeciętnych miesięcznych wydatków na rekreację i kulturę. Informacje o wysokości środków finansowych kierowanych na różne formy wypoczynku mogą pozwolić odpowiednim organizacjom (np. klubom seniora), stowarzyszeniom i biurom turystycznym na przygotowanie oferty odpowiedniej do oczekiwań badanych zbiorowości.

Literatura

- Budżety gospodarstw domowych w 2009 roku*, Informacje i Opracowania Statystyczne, GUS, Warszawa 2010.
- Emerytury i renty w 2009 roku*, Informacje i Opracowania Statystyczne, GUS, Warszawa 2010.
- 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*, Wydawnictwo AE, Wrocław 2004.
- Podolec B., Ulman P., Wałęga A., *Kształtowanie się dochodów a poziom i struktura wydatków młodych małżeństw*, [w:] *Statystyka w praktyce społeczno-gospodarczej*, W. Ostasiewicz (red.), Wydawnictwo AE, Wrocław 2007.

SEGMENTATION OF PENSIONERS AND ANNUITANTS HOUSEHOLDS IN TERMS OF EXPENDITURES ON RECREATION AND CULTURE

Summary: The article presents the results of studies concerning the segmentation of pensioners and annuitants households according to their average monthly expenditures on recreation and culture. Regression tree was used for the classification of households. The unit of unidentifiable data about income and expenditures of individual households, derived from households budget surveys conducted in 2009 by the Central Statistical Office were the information basis for the studies. The segmentation was performed for the representative sample of 1308 pensioners households and 327 households of annuitants.

Keywords: households of pensioners and annuitants, segmentation, regression trees.