PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp 9

Eugeniusz Gatnar, Balance of payments statistics and external competitiveness of Poland 15

Andrzej Sokołowski, Magdalena Czaja, Efektywność metody *k*-średnich w zależności od separowalności grup 23

Barbara Pawełek, Józef Pociecha, Adam Sagan, Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw 30

Elżbieta Gołata, Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011 39

Aleksandra Łuczak, Feliks Wysocki, Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów 49

Marek Walesiak, Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej 60

Paweł Lula, Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu *RAKE* 69

Mariusz Kubus, Propozycja modyfikacji metody złagodzonego LASSO 77

Andrzej Bąk, Tomasz Bartłomowicz, Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie `DiscreteChoice` programu R 85

Justyna Brzezińska, Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012 95

Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Zastosowanie pakietu `dcMNM` programu R w badaniach preferencji konsumentów wódki 104

Barbara Batóg, Jacek Batóg, Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010 113

Małgorzata Markowska, Danuta Strahl, Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego 121

Kamila Migdał-Najman, Krzysztof Najman, Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 131

Kamila Migdał-Najman, Krzysztof Najman, Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 139

Beata Basiura, Anna Czapkiewicz, Badanie jakości klasyfikacji szeregów czasowych 148

Michał Trzęsiok, Wybrane metody identyfikacji obserwacji oddalonych 157

Grażyna Dehnel, Tomasz Klimanek, Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej ... 167
Michał Bernard Pietrzak, Justyna Wilk, Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji ... 177
Maciej Beręsewicz, Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena ... 186
Marcin Szymkowiak, Tomasz Józefowski, Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji ... 195
Marcin Pełka, Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym ... 202
Małgorzata Machowska-Szewczyk, Ocena klas w rozmytej klasyfikacji obiektów symbolicznych ... 210
Justyna Wilk, Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych ... 220
Andrzej Dudek, Metody analizy skupień w klasyfikacji markerów map Google ... 229
Ewa Roszkowska, Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW 237
Marcin Szymkowiak, Marek Witkowski, Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych ... 248
Bartłomiej Jefmański, Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R ... 257
Karolina Bartos, Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych ... 266
Joanna Trzęsiok, Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych ... 275
Beata Bal-Domańska, Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja ... 285
Beata Bieszk-Stolorz, Iwona Markowicz, Wpływ zasiłku na proces poszukiwania pracy ... 294
Marta Dziechciarz-Duda, Klaudia Przybysz, Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni ... 303
Tomasz Klimanek, Problem pomiaru procesu dezagraryzacji wsi polskiej w świetle wielowymiarowych metod statystycznych ... 313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Wybrane metody analizy danych wzdłużnych ... 321
Artur Zaborski, Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych ... 330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy ... 338

Katarzyna Wawrzyniak, Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego 346

Summaries

Eugeniusz Gatnar, Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski 22
Andrzej Sokołowski, Magdalena Czaja, Cluster separability and the effectiveness of *k*-means method 29
Barbara Pawełek, Józef Pociecha, Adam Sagan, Multisectoral analysis of latent transitions in bankruptcy prediction models..... 38
Elżbieta Gołata, Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011 48
Aleksandra Łuczak, Feliks Wysocki, Determination of weights for features in problems of linear ordering of objects 59
Marek Walesiak, Reinforcing measurement scale for ordinal data in multivariate statistical analysis 68
Paweł Lula, Automatic identification of keywords and keyphrases in documents written in Polish..... 76
Mariusz Kubus, The proposition of modification of the relaxed LASSO method..... 84
Andrzej Bąk, Tomasz Bartłomowicz, Microeconomic multinomial logit models and their implementation in the `DiscreteChoice` R package. 94
Justyna Brzezińska, The analysis of unemployment data in Poland in 2004--2012 with application of log-linear models 103
Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Application of the MMLM package of R software for vodka consumers preference analysis..... 112
Barbara Batóg, Jacek Batóg, Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity 120
Małgorzata Markowska, Danuta Strahl, Classification of the European regional space in terms of smart growth pillars using the reference limit system..... 130
Kamila Migdał Najman, Krzysztof Najman, Formal quality assessment of group structure mapping on the Kohonen's map 138
Kamila Migdał Najman, Krzysztof Najman, Graphical quality assessment of group structure mapping on the Kohonen's map 147
Beata Basiura, Anna Czapkiewicz, Validation of time series clustering 156
Michał Trzęsiok, Selected methods for outlier detection..... 166

Grażyna Dehnel, Tomasz Klimanek, Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics 176
Michał Bernard Pietrzak, Justyna Wilk, Economic distance in modeling spatial phenomena with the application of gravity model..... 185
Maciej Beręsewicz, An attempt to use different distance measures in the Generalized Petersen estimator 194
Marcin Szymkowiak, Tomasz Józefowski, Construction and practical using of SPREE estimators for two-dimensional contingency tables..... 201
Marcin Pełka, The ensemble conceptual clustering for symbolic data..... 209
Małgorzata Machowska-Szewczyk, Evaluation of clusters obtained by fuzzy classification methods for symbolic objects..... 219
Justyna Wilk, Problem of determining the number of clusters in taxonomic analysis of symbolic data 228
Andrzej Dudek, Clustering techniques for Google maps markers..... 236
Ewa Roszkowska, The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure 247
Marcin Szymkowiak, Marek Witkowski, The use of correspondence analysis in analysing the financial situation of cooperative banks..... 256
Bartłomiej Jefmański, The construction of fuzzy customer satisfaction indexes using R program..... 265
Karolina Bartos, Discovering patterns of consumer behaviour by market basket analysis of the transactional data..... 274
Joanna Trzęsiok, Cluster analysis of countries with respect to fertility rate and other demographic factors 284
Beata Bal-Domańska, An attempt to identify major regional clusters and their convergence 293
Beata Bieszk-Stolorz, Iwona Markowicz, The influence of benefit on the job finding process 302
Marta Dziechciarz-Duda, Klaudia Przybysz, Education and labor market needs. Classification of university graduates 312
Tomasz Klimanek, The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods..... 320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Selected methods for an analysis of longitudinal data..... 329
Artur Zaborski, The application of distance measures for ordinal data for aggregation individual preferences 337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market 345
Katarzyna Wawrzyniak, Classification of voivodeships according to the level of the realization of priorities of *the National Development Strategy 2007-2015* with using the values of centroid of the rows 355

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Karolina Bartos
Uniwersytet Ekonomiczny we Wrocławiu

ODKRYWANIE WZORCÓW ZACHOWAŃ KONSUMENTÓW ZA POMOCĄ ANALIZY KOSZYKOWEJ DANYCH TRANSAKCYJNYCH

Streszczenie: W artykule poddano badaniu bazę danych małych sklepów typu convenience, zawierającą prawie pięć tysięcy transakcji. Celem analizy było odkrycie reguł asocjacji pomiędzy nabywanymi kategoriami produktów, a także sprawdzenie, czy istnieją zależności pomiędzy sprzedawaną kategorią produktu a porą dnia, w której dokonano transakcji. Ponadto opisano proces przygotowania danych do analizy oraz zaprezentowano podstawowe miary charakteryzujące siłę odkrytych reguł. Do badania wykorzystano algorytm *a priori*.

Słowa kluczowe: zachowanie konsumentów, analiza koszykowa, algorytm *a priori.*

1. Wstęp

Badanie asocjacji należy do klasy metod, która zajmuje się odkrywaniem zależności lub korelacji, nazywanych ogólniej asocjacjami, pomiędzy danymi w dużych zbiorach danych. Wykorzystywana jest w wielu dziedzinach, np.: w medycynie [Laxminarayan i in. 2006, s. 440-450], edukacji [Radosav i in. 2012, s. 933-944] czy do wykrywania oszustw finansowych [Sanchez i in. 2009, s. 3630-3640]. Jednak najbardziej powszechnym przykładem jej zastosowania jest tzw. analiza koszykowa (*market basket analysis*) [Pasztyła b.d.w.; Łapczyński 2009; Śniegocka-Łusiewicz 2011; Chen i in. 2005], gdzie reguły asocjacji generowane są na podstawie danych koszyka sklepowego. Celem tej analizy jest znalezienie naturalnych wzorców zachowań nabywczych konsumentów poprzez analizę produktów, które najczęściej wspólnie są przez nich kupowane (trafiają do jednego koszyka sklepowego podczas zakupów). Ich znajomość jest niezwykle cenna, gdyż pozwala na lepsze planowanie działań prowadzących do zwiększenia sprzedaży np. poprzez efektywniejsze planowanie rozmieszczenia produktów na sklepowych półkach.

Reguły asocjacji przybierają postać: „Jeżeli *poprzednik*, to *następnik*". Wynik analizy koszykowej przedstawiony jest jako zbiór reguł asocjacyjnych w postaci relacji, które prezentuje wzór (1):

$$\{(A_1 = 1) \wedge \ldots \wedge (A_k = 1)\} \rightarrow \{(B_1 = 1) \wedge \ldots \wedge (B_k = 1)\}. \quad (1)$$

Oznacza to, że jeśli klient kupił produkty A_1, A_2 aż do A_k, to z dużym prawdopodobieństwem kupi także produkty B_1, B_2 aż do B_k. Przykładowo reguła może brzmieć: „Jeżeli konsument kupił ser, kiełbasę i szynkę to prawdopodobnie kupi też masło, pomidory i chleb".

Z każdą regułą asocjacji związane są dwie podstawowe miary charakteryzujące jej statystyczną ważność i siłę:
- wsparcie (*support*),
- ufność (*confidence*), zwana także pewnością.

Wsparcie dla reguły asocjacyjnej A→B jest procentem transakcji w Z, które zawierają A i B [Larose 2006, s. 188-189]:

$$wsparcie = P(A \cap B) = \frac{liczba\ transakcji\ zawierających\ A\ i\ B}{całkowita\ liczba\ transakcji}. \quad (2)$$

Ufność dla reguły A→B jest miarą dokładności reguły, określa, jaki procent transakcji zawierających A, również zawiera B [Larose 2006, s.188-189]:

$$ufność = \frac{P(A \cap B)}{P(A)} = \frac{liczba\ transakcji\ zawierających\ A\ i\ B}{liczba\ transakcji\ zawierających\ A}. \quad (3)$$

Jest to inaczej prawdopodobieństwo, że losowo wybrany klient, który nabył produkt A, również zakupi produkt B [Pasztyła 2005, s. 62].

Istnieje jeszcze wiele innych miar wykorzystywanych w analizie koszykowej. Wśród nich warto wymienić korelację (*correlation*) i przyrost (*lift*). Korelacja informuje, w jakim stopniu fakt, że klient wybrał produkt A, zwiększa (dodatnia korelacja) lub zmniejsza (ujemna korelacja) prawdopodobieństwo, iż wybierze on również produkt B. Przyrost jest modyfikacją miary korelacji i również określa wpływ sprzedaży produktu A na prawdopodobieństwo sprzedaż produktu B.

2. Przygotowanie danych empirycznych do badania

Celem badania było odkrycie reguł asocjacji pomiędzy nabywanymi produktami, a także sprawdzenie, czy istnieją zależności pomiędzy kupowanym produktem a porą dnia, w której dokonywano transakcji. Przedmiotem analizy były dane transakcyjne z trzech małych samoobsługowych osiedlowych sklepów, należących do jednej sieci. Placówki zlokalizowane były w dużym polskim mieście i oferowały głównie asortyment spożywczy oraz w niewielkiej ilości artykuły przemysłowe. Dane dotyczyły okresu trzech dni i opisywały 4991 transakcji (9781 zakupionych produktów).

Proces obliczeń musiał zostać poprzedzony wstępną obróbką danych. Połączono w całość dane z trzech analizowanych sklepów. Ze względu na unikatową wartość numerów paragonów (numery paragonów z różnych sklepów nie powtarzały się), nie było potrzeby tworzenia osobnego pola z numerem ID transakcji. Ponadto

usunięto wiersze zawierające sumę wartości poszczególnych transakcji. Kolumnę *Godzina na paragonie* przekształcono na kolumnę zawierająca informację o porze dnia dokonanej transakcji: wczesny ranek (od otwarcia sklepu około godziny 6^{00} do 9^{00}); rano $<9^{00}$-11^{00}); południe $<11^{00}$-14^{00}); popołudnie $<14^{00}$-18^{00}); wieczór $<18^{00}$-21^{00}); noc $<21^{00}$ do zamknięcia sklepu, w zależności od dnia i sklepu była to godzina 23^{00} -2^{00}). Asortyment sklepów (1455 rodzajów produktów) został pogrupowany w 30 następujących kategorii (kolejność według malejącej ilości sprzedaży produktów z danej kategorii): piwo (badane sklepy nie oferowały innych produktów alkoholowych); pieczywo (chleb, bułki, rogale, drożdżówki, pączki, kanapki); batony (do tej kategorii zostały zakwalifikowane także wafle w czekoladzie, jak np. Prince Polo, Grzesiek); napoje bezalkoholowe; papierosy; woda; zupki błyskawiczne (tzw. „zupki chińskie"); słone przekąski (chipsy, paluszki, orzeszki, słonecznik itp.); artykuły przemysłowe (kosmetyki, detergenty, chusteczki higieniczne, papier toaletowy, torby okolicznościowe, zapalniczki, golarki, baterie, karmy dla zwierząt, leki); prasa; jogurty (tutaj także: śmietana, maślanka, kefir, śmietanka do kawy); cukierki (tutaj także: lizaki, draże, dropsy); Doładowania telefonu (także startery); warzywa i owoce; obiadowe (kasze, ryż, makarony, mąka, cukier, przyprawy, sosy w proszku, zupy w proszku, oleje, ketchupy, majonezy, musztardy); sery; gumy do żucia; torba foliowa; mleko; lody i desery; kawa i herbata; wędliny (tutaj także konserwy i pasztety); ryby (także sałatki rybne); ciastka; czekolada; masło; dania gotowe (słoiki z gotowymi daniami, mrożonki); dżemy i pasty do kanapek; bilety; płatki śniadaniowe i kaszki.

Ze względu na duży udział sprzedaży batonów w sprzedaży ogółem zdecydowano o wyodrębnieniu osobnej kategorii batony i niewłączeniu jej do kategorii czekolada. Podobnie, ze względu na dużą sprzedaż zupek błyskawicznych, stworzono osobną grupę zawierającą tylko ten produkt.

3. Wykrywanie wzorców zachowań klientów – wyniki analizy

Do badania wykorzystano program SAL (*Sequence, Association and Link Analysis*) pakietu Statistica 10.0, który pozwala na przeprowadzenie analizy koszykowej z zastosowaniem algorytmu *a priori* (szerzej na temat algorytmu *a priori* w: [Bartos 2012, s. 281-283]). Eksplorację danych przeprowadzono bez analizy sekwencji, ustalając minimalne wsparcie na 1% oraz zaufanie na poziomie 20%.

Pierwszy etap analiz dotyczył wyszukania reguł asocjacji wyłącznie pomiędzy kategoriami produktów (tab. 1), nie uwzględniono pory dnia, w jakiej dokonywano transakcji.

Wyraźnie widać bardzo dużą zależność pomiędzy produktami: woda, zupki błyskawiczne i batony. Przypuszcza się, że w badanym okresie w analizowanych sklepach prowadzona była promocja jednego lub kilku artykułów z tych kategorii. Prawie co dziesiąty klient (wsparcie 9,78% – 488 takich transakcji) zakupił w jednym koszyku produkty z tych trzech kategorii. Analizując reguły z tabeli 1, można

Tabela 1. Reguły asocjacji pomiędzy kategoriami produktów

	Podsumowanie reguł asocjacji (Analiza-30 kategorii) Min: wsparcie= 1,0%, zaufanie = 20,0% Maks. liczność zestawu = 10				
	Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
11	**WODA, ZUPKI BŁYSKAWICZNE**	**==>**	**BATONY**	**9,77760**	**97,40519**
13	BATONY, ZUPKI BŁYSKAWICZNE	==>	WODA	9,77760	96,25247
14	BATONY, WODA	==>	ZUPKI BŁYSKAWICZNE	9,77760	94,39072
7	ZUPKI BŁYSKAWICZNE	==>	BATONY	10,15828	82,17180
18	ZUPKI BŁYSKAWICZNE	==>	WODA	10,03807	81,19935
9	ZUPKI BŁYSKAWICZNE	==>	BATONY, WODA	9,77760	79,09238
15	WODA	==>	BATONY	10,35865	72,51052
19	WODA	==>	ZUPKI BŁYSKAWICZNE	10,03807	70,26648
10	WODA	==>	BATONY, ZUPKI BŁYSKAWICZNE	9,77760	68,44320
16	BATONY	==>	WODA	10,35865	60,82353
8	BATONY	==>	ZUPKI BŁYSKAWICZNE	10,15828	59,64706
12	BATONY	==>	WODA, ZUPKI BŁYSKAWICZNE	9,77760	57,41176
6	**SER**	**==>**	**PIECZYWO**	**1,40252**	**54,68750**
3	**TORBA FOLIOWA**	**==>**	**PIWO**	**1,10198**	**41,35338**
5	**JOGURTY**	**==>**	**PIECZYWO**	**1,44260**	**40,67797**
4	**SŁONE PRZEKĄSKI**	**==>**	**PIWO**	**1,44260**	**26,18182**
2	**PAPIEROSY**	**==>**	**PIWO**	**3,60649**	**23,56021**
17	SŁONE PRZEKĄSKI	==>	NAPOJE BEZALKOHOLOWE	1,26227	22,90909
1	ART.PRZEMYSŁOWE	==>	PAPIEROSY	1,14206	22,00772

Źródło: opracowanie własne za pomocą pakietu Statistica 10.

powiedzieć, że 97,4% kupujących wodę i zupkę błyskawiczną nabyło także w tej samej transakcji produkt z kategorii batony. 96,25% klientów kupujących batona i zupkę sięgnęło jeszcze po wodę, a 94,39% klientów mających w koszyku batona i wodę zakupiło również zupkę. Duże zaufanie mają także reguły: jeżeli ser, to pieczywo (prawie 55%), jeżeli torba foliowa, to piwo (41,35%), jeżeli jogurt, to pieczywo (40,68%), jeżeli słone przekąski, to piwo (26,18%) oraz jeżeli papierosy, to piwo (23,56%).

Tabela 2 przedstawia najczęstsze zbiory kategorii produktów występujących w analizowanych transakcjach. Wspomniana wcześniej grupa: woda, zupki błyskawiczne i batony jest najczęstszym zbiorem trzyelementowym i wystąpiła aż 488 razy. Koszyk z batonem i wodą występował częściej (517 razy) niż koszyk z batonem i zupką (507) oraz wodą i zupką (501). Klienci często decydowali się też na łączny zakup papierosów i piwa (180) oraz pieczywa z napojami bezalkoholowymi (129).

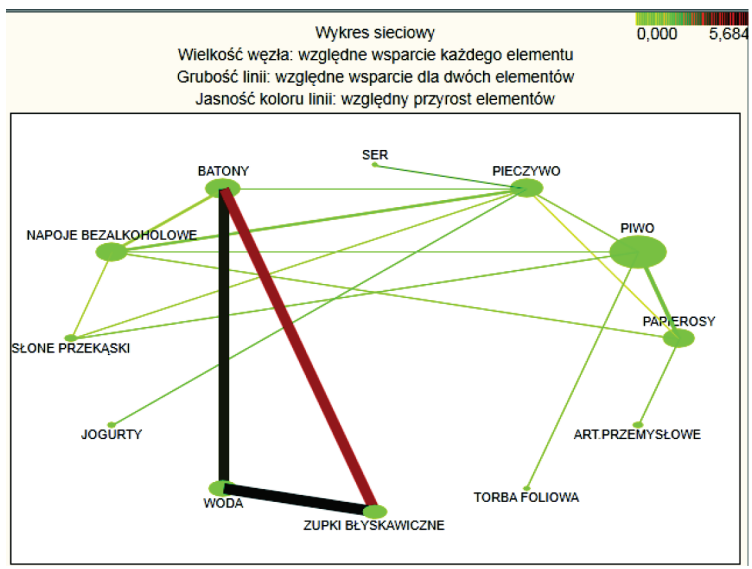
Na wykresie sieciowym (rys. 1) pokazane są zależności między wybranymi kategoriami produktów. Ze względu na dużą liczbę kategorii oraz to że wiele grup nie ma powiązań między sobą, zdecydowano się na pozostawienie tylko najistotniejszych kategorii. Wielkość węzła symbolizuje tutaj wsparcie poszczególnej kategorii produktów, a grubość łączących je linii wsparcie dla danego zbioru kategorii produktów.

Oprócz widocznych tutaj i opisanych już wcześniej zależności można dodatkowo zauważyć, że pieczywo jest kategorią, która łączy się z największą liczbą innych kategorii produktów (ma aż 7 powiązań). Oznacza to, że często razem z pieczywem klienci kupują różne inne artykuły. Nie dziwi, więc fakt, że właści-

Tabela 2. Częstość występowania 3- i 2-elementowych zbiorów kategorii produktów

	Wyznaczone popularne zestawy (Analiza-30 kategorii) Min: wsparcie= 1,0%, zaufanie = 20,0% Maks. liczność zestawu = 10			
	Popularne zestawy	Liczba elementów	Liczność	Wsparcie%
41	{ BATONY, WODA, ZUPKI BŁYSKAWICZNE }	3,000000	**488,000**	9,77760
42	{ BATONY, WODA }	2,000000	**517,000**	10,35865
40	{ BATONY, ZUPKI BŁYSKAWICZNE }	2,000000	**507,000**	10,15828
45	{ WODA, ZUPKI BŁYSKAWICZNE }	2,000000	**501,000**	10,03807
30	{ PAPIEROSY, PIWO }	2,000000	180,000	3,60649
37	{ PIECZYWO, NAPOJE BEZALKOHOLOWE }	2,000000	129,000	2,58465
43	{ BATONY, NAPOJE BEZALKOHOLOWE }	2,000000	86,000	1,72310
33	{ PIWO, NAPOJE BEZALKOHOLOWE }	2,000000	80,000	1,60289
34	{ PIWO, PIECZYWO }	2,000000	77,000	1,54278
32	{ PIWO, SŁONE PRZEKĄSKI }	2,000000	72,000	1,44260
35	{ PIECZYWO, JOGURTY }	2,000000	72,000	1,44260
28	{ PAPIEROSY, NAPOJE BEZALKOHOLOWE }	2,000000	70,000	1,40252
39	{ PIECZYWO, SER }	2,000000	70,000	1,40252
44	{ NAPOJE BEZALKOHOLOWE, SŁONE PRZEKĄSKI }	2,000000	63,000	1,26227
29	{ PAPIEROSY, PIECZYWO }	2,000000	61,000	1,22220
38	{ PIECZYWO, BATONY }	2,000000	61,000	1,22220
27	{ PAPIEROSY, ART.PRZEMYSŁOWE }	2,000000	57,000	1,14206
31	{ PIWO, TORBA FOLIOWA }	2,000000	55,000	1,10198
36	{ PIECZYWO, SŁONE PRZEKĄSKI }	2,000000	54,000	1,08195

Źródło: opracowanie własne za pomocą pakietu Statistica 10.

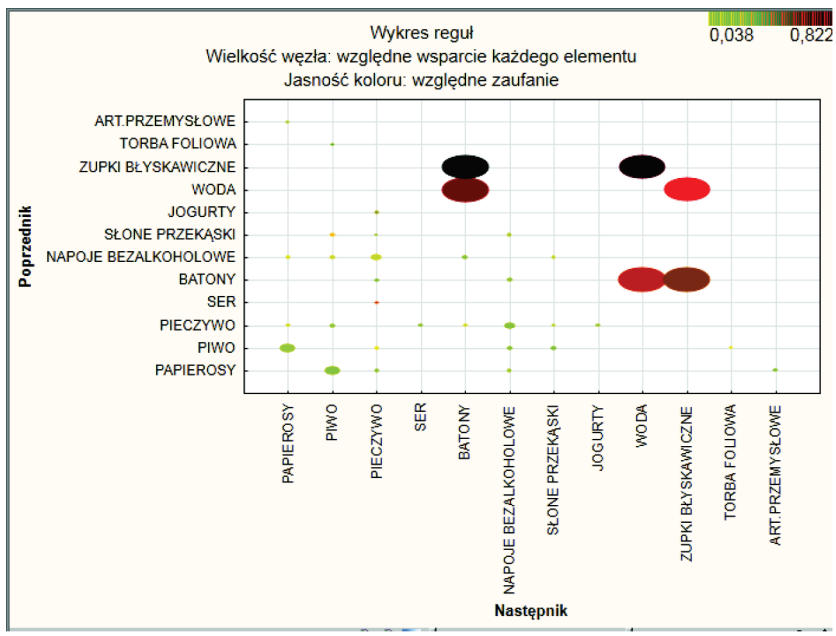


Rys. 1. Wykres sieciowy dla 12 najważniejszych kategorii

Źródło: opracowanie własne za pomocą pakietu Statistica 10.

ciele sklepów spożywczych zabiegają o dobry asortyment pieczywa. Wiedzą oni bowiem, że jego sprzedaż pociągnie za sobą również sprzedaż produktów z innych kategorii. Po 5 powiązań mają piwo i napoje bezalkoholowe, a po 4 batony i papie-

rosy. Widoczne jest to także na rys. 2, gdzie dodatkowo kolor symbolizuje względne zaufanie pomiędzy kategoriami produktów: im ciemniejszy kolor węzła, tym większe zaufanie. Najciemniejszy węzeł przy regule: „jeśli zupka błyskawiczna, to baton", wskazuje na największe zaufanie dla tej reguły (ponad 82%) spośród wszystkich analizowanych par produktów. Ciemniejszy kolor węzła przy regule: „jeśli papierosy, to piwo" niż przy regule „jeśli piwo, to papierosy", świadczy, że częściej konsumenci wybierający papierosy sięgają dodatkowo po piwo niż osoby kupujące piwo po papierosy. Podobna sytuacja jest w przypadku produktów z kategorii pieczywo i napoje bezalkoholowe. Klienci chętniej wraz z pieczywem kupują napoje bezalkoholowe niż odwrotnie.



Rys. 2. Wykres reguł dla 12 najważniejszych kategorii

Źródło: opracowanie własne za pomocą pakietu Statistica 10.

Kolejnym etapem analizy było włączenie do badania danych dotyczących pory dnia, w której dokonana została transakcja. Umożliwiło to sprawdzenie, czy istnieją silne zależności pomiędzy porą dnia a kategorią produktów, która najchętniej była wtedy kupowana.

Najwięcej produktów badane sklepy sprzedają wieczorem <18:00 do 21:00), nocą <21:00 do zamknięcia) oraz po południu <14:00 do 18:00) (tabela 3). W tym okresie sprzedano łącznie ponad 66% produktów. Nie dziwi więc fakt wystąpienia wielu reguł o wysokim wsparciu i zaufaniu właśnie o tej porze (tabela 4). Blisko 42% klientów kupujących zupkę błyskawiczną dokonało zakupu w nocy. Podobnie

wodę (ponad 40%) oraz batony prawie 33%. Słone przekąski i piwo wkładano do koszyka najczęściej wieczorem, a pieczywo wczesnym rankiem (od około 6:00 do 9:00).

Tabela 3. Liczba sprzedanych produktów a pora dnia

	Popularne zestawy zawierające wybrane elementy Min: wsparcie= 1,0%, zaufanie = 20,0% Maks. liczność zestawu = 2			
	Popularne zestawy	Liczba elementów	Liczność	Wsparcie%
5	(Wieczór)	1,000000	2267,000	23,17759
6	(Noc)	1,000000	2115,000	21,62356
4	(Popołudnie)	1,000000	2104,000	21,51109
3	(Południe)	1,000000	1440,000	14,72242
1	(Wczesny ranek)	1,000000	1165,000	11,91085
2	(Rano)	1,000000	690,000	7,05449

Źródło: opracowanie własne za pomocą pakietu Statistica 10.

Tabela 4. Reguły asocjacji między kategorią produktu jako poprzednik a porą dnia jako następnik

	Podsumowanie reguł asocjacji dla wybranych elementów (Analiza-30 kategorii) Min: wsparcie= 1,0%, zaufanie = 20,0% Maks. liczność zestawu = 2				
	Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
14	**ZUPKI BŁYSKAWICZNE**	**==>**	**Noc**	**2,740006**	**41,61491**
12	**WODA**	**==>**	**Noc**	**3,046723**	**40,82192**
11	**SŁONE PRZEKĄSKI**	**==>**	**Wieczór**	**1,114405**	**34,16928**
7	**BATONY**	**==>**	**Noc**	**3,077395**	**32,86026**
5	**PIWO**	**==>**	**Wieczór**	**5,970760**	**31,26338**
1	**PIECZYWO**	**==>**	**Wczesny ranek**	**2,801350**	**28,78151**
2	PAPIEROSY	==>	Wieczór	2,218587	27,46835
10	NAPOJE BEZALKOHOLOWE	==>	Popołudnie	2,433289	26,41509
4	PIWO	==>	Noc	4,989265	26,12420
8	BATONY	==>	Wieczór	2,147020	22,92576
9	NAPOJE BEZALKOHOLOWE	==>	Wieczór	2,106124	22,86349
3	PAPIEROSY	==>	Popołudnie	1,768735	21,89873
6	PIWO	==>	Popołudnie	4,099785	21,46681
16	ZUPKI BŁYSKAWICZNE	==>	Południe	1,390451	21,11801
15	ZUPKI BŁYSKAWICZNE	==>	Wieczór	1,349555	20,49689
13	WODA	==>	Wieczór	1,502914	20,13699

Źródło: opracowanie własne za pomocą pakietu Statistica 10.

W tabeli 5, obrazującej reguły asocjacji między porą dnia jako poprzednikiem a kategorią produktu jako następnikiem, można zauważyć, że wieczorem i nocą sprzedawano głównie piwo z zaufaniem odpowiednio 26% i 23%. Natomiast najczęściej kupowanym artykułem wczesnym rankiem i rano było pieczywo (zaufanie 23,5% i 22%).

Tabela 5. Reguły asocjacji między porą dnia jako poprzednik a kategorią produktu jako następnik

	Podsumowanie reguł asocjacji dla wybranych elementów Min: wsparcie= 1,0%, zaufanie = 20,0% Maks. liczność zestawu = 2				
	Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)
3	Wieczór	==>	PIWO	5,970760	25,76092
1	Wczesny ranek	==>	PIECZYWO	2,801350	23,51931
2	Noc	==>	PIWO	4,989265	23,07329
4	Rano	==>	PIECZYWO	1,533586	21,73913

Źródło: opracowanie własne za pomocą pakietu Statistica 10.

4. Podsumowanie

Na podstawie przeprowadzonego badania wykryto bardzo silną zależność między kupnem: wody, zupki błyskawicznej i batonu. Prawie co dziesiąty klient kupił łącznie te trzy produkty. Reguła: jeżeli woda i zupki błyskawiczne, to baton, osiągnęła zaufanie ponad 97%. Ponadto duży wskaźnik zaufania dla każdej reguły zawierającej te trzy elementy świadczy o dużym prawdopodobieństwie łącznego zakupu artykułów z tych trzech kategorii. Odnaleziono także dość silną regułę: jeżeli ser, to pieczywo (zaufanie prawie 55%), jeżeli torba foliowa, to piwo (41,35%), jeżeli jogurt, to pieczywo (40,68%), jeżeli słone przekąski, to piwo (26,18%), oraz jeżeli papierosy, to piwo (23,56%). Duga część analizy, do której włączono dane dotyczące pory dnia zakupu, wykazała, że kupno poszczególnych kategorii produktów silnie zależy od czasu, w którym przebiega transakcja. Blisko 42% klientów kupujących zupkę błyskawiczną dokonało zakupu w nocy. Podobnie wodę (ponad 40%) oraz batony prawie 33%. Słone przekąski i piwo wkładano do koszyka najczęściej wieczorem (zaufanie 34%, 31%). Ponadto wyniki badania pokazują, że wieczorem i nocą sklepy sprzedawały głównie piwo (zaufanie 26%, 23%), natomiast rano i wczesnym rankiem – pieczywo (23,5%, 22%).

Odkryte reguły dostarczają zarządzającym badanymi sklepami ważnych informacji o zachowaniach ich klientów. Znajomość odnalezionych zależności może się przyczynić m.in. do dużo efektywniejszego wykorzystania narzędzi marketingu krzyżowego, co pociągnie za sobą zwiększenie sprzedaży.

Literatura

Bartos K. (2013), *Association rules in the study of consumer behaviour*, Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 763, Szczecin, s. 279-286.

Chen Y.L., Tang K., Shen R.J., Hu Y.H. (2005), *Market basket analysis in a multiple store environment*, „Decision Support System", Vol. 40, Issue 2, s. 339-354.

Kurzawa I., Wysocki F. (2008), *Wykorzystanie analizy koszykowej do identyfikacji zachowań konsumpcyjnych gospodarstw domowych w Polsce*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja*

i analiza danych – teoria i zastosowania, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 7 (1207), Taksonomia 15, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 527-534.

Larose D.T. (2006), *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Laxminarayan P., Alvarez S.A., Ruiz C. (2006), *Mining statistically significant associations for exploratory analysis of human sleep data*, „IEEE Transactions on Information Technology in Biomedicine", Vol. 10, Issue 3, s. 440-450.

Łapczyński M. (2009), *Analiza koszykowa i analiza sekwencji – wielki brat czuwa*, www.statsoft.pl (29.01.2014).

Łapczyński M. (2010), *Web Usage Mining, czyli jak sprzedać sukienkę ciążową w Internecie*, www.statsoft.pl (29.01.2014).

Olson D.L., Delen D. (2008), *Advanced Data Mining Techniques*, Springer, Berlin, s. 53-67.

Pasztyła A. (b.d.), *Analiza koszykowa danych transakcyjnych – cele i metody*, „Magazyn Systemy IT", s. 51-54, www.statsoft.pl (10.04.2013).

Pasztyła A. (2005), *Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej*, StatSoft Polska, www.statsoft.pl (21.12.2012).

Radosav D., Brtka E., Brtka V. (2012), *Association Rules from Empirical Data in the Domain of Education*, „International Journal of Computers Communications & Control", Vol. 7, Issue 5, s. 933-944.

Sanchez D., Vila M.A., Cerda L. (2009), „Expert Systems with Applications", Vol. 36, Issue 2, s. 3630-3640.

Śniegocka-Łusiewicz M. (2011), *Analiza koszykowa w badaniu cykliczności reguł asocjacji w handlu detalicznym*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Taksonomia 28, Wydawnictwo Uniwersytetu Ekonomicznego, Wrocław, s. 621-628.

DISCOVERING PATTERNS OF CONSUMER BEHAVIOUR BY MARKET BASKET ANALYSIS OF THE TRANSACTIONAL DATA

Summary: The article depicts a study analysing small convenience stores' database, which contains nearly five thousand transactions. The first aim of the research is to discover association rules between the categories of purchased products. The second goal is to verify whether there is any relationship between the selling category of products and the time of day when transactions were made. Moreover, the process of preparing data for analysis is described and the basic measures characterizing the strength of discovered rules are presented. The *a priori* algorithm was used during the research.

Keywords: consumer behaviour, market basket analysis, *a priori* algorithm.