

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

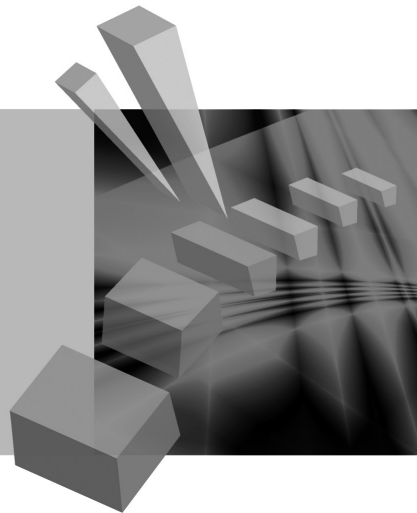
RESEARCH PAPERS

of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowci

Krzysztof Jajuga

Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

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

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 2013

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

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

| | |
|---|-----|
| Wstęp | 9 |
| Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: Sejm VI kadencji – maszynka do głosowania | 11 |
| Barbara Pawelek, Adam Sagan: Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I | 19 |
| Jan Paradysz: Nowe możliwości badania koniunktury na rynku pracy | 29 |
| Krzysztof Najman: Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze | 41 |
| Kamila Migdał-Najman: Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym | 48 |
| Aleksandra Matuszewska-Janica, Dorota Witkowska: Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych | 58 |
| Iwona Foryś, Ewa Putek-Szeląg: Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim | 67 |
| Joanna Banaś, Małgorzata Machowska-Szewczyk: Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych..... | 77 |
| Marta Jaročka: Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni | 85 |
| Anna Zamojska: Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych..... | 95 |
| Dorota Rozmus: Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i> | 106 |
| Ewa Wędrowska: Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur..... | 115 |
| Katarzyna Wójcik, Janusz Tuchowski: Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ... | 124 |
| Małgorzata Misztal: Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych..... | 135 |
| Anna Czapkiewicz, Beata Basiura: Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych | 146 |
| Tomasz Szubert: Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej” | 154 |

| | |
|--|-----|
| Marcin Szymkowiak: Konstrukcja estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości | 164 |
| Wojciech Roszka: Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie | 174 |
| Justyna Brzezińska: Metody wizualizacji danych jakościowych w programie R | 182 |
| Agata Sielska: Regionalne zróżnicowanie potencjału konkurencyjnego polskich gospodarstw rolnych w województwach po akcesji do Unii Europejskiej | 191 |
| Mariusz Kubus: Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych | 201 |
| Beata Basiura: Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości | 209 |
| Katarzyna Wardzińska: Wykorzystanie metody obwiedni danych w procesie klasyfikacji przedsiębiorstw | 217 |
| Katarzyna Dębowska: Modelowanie upadłości przedsiębiorstw oparte na próbach niezbilansowanych | 226 |
| Danuta Tarka: Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska .. | 235 |
| Artur Czech: Zastosowanie wybranych metod doboru zmiennych diagnostycznych w badaniach konsumpcji w ujęciu pośrednim | 246 |
| Beata Bal-Domańska: Ocena relacji zachodzących między inteligentnym rozwojem a spójnością ekonomiczną w wymiarze regionalnym z wykorzystaniem modeli panelowych | 255 |
| Mariola Chrzanowska: <i>Ordinary kriging</i> i <i>inverse distance weighting</i> jako metody szacowania cen nieruchomości na przykładzie warszawskiego rynku | 264 |
| Adam Depta: Zastosowanie analizy wariancji w badaniu jakości życia na podstawie kwestionariusza SF-36v2 | 272 |
| Maciej Beręsewicz, Tomasz Klimanek: Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach cen mieszkań | 281 |
| Karolina Paradysz: Benchmarkowa analiza estymacji dla małych obszarów na lokalnych rynkach pracy | 291 |
| Anna Gryko-Nikitin: Dobór parametrów w równoległych algorytmach genetycznych dla problemu plecakowego | 301 |
| Tomasz Ząbkowski, Piotr Jałowiecki: Zastosowanie reguł asocjacyjnych do analizy danych ankietowych w wybranych obszarach logistyki przedsiębiorstw przetwórstwa rolno-spożywczego | 311 |
| Agnieszka Przedborska, Małgorzata Misztal: Zastosowanie metod statystyki wielowymiarowej do oceny wydolności stawów kolanowych u pacjentów z chorobą zwyrodnieniową leczonych operacyjnie | 321 |
| Dorota Perło: Rozwój zrównoważony w wymiarze gospodarczym, społecznym i środowiskowym – analiza przestrzenna | 331 |

| | |
|--|-----|
| Ewa Putek-Szeląg, Urszula Gieraltowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej.. | 342 |
|--|-----|

Summaries

| | |
|--|-----|
| Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: VIth-term Sejm – a voting machine | 18 |
| Barbara Pawelek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model | 28 |
| Jan Paradysz: New possibilities for studying the situation on the labour market | 40 |
| Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data..... | 47 |
| Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering | 57 |
| Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees..... | 66 |
| Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship... | 76 |
| Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables | 84 |
| Marta Jarocka: The impact of the method of the selection of diagnostic variables on the result of linear ordering on the example of ranking of universities in Poland..... | 94 |
| Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures..... | 105 |
| Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea..... | 114 |
| Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements | 123 |
| Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis | 134 |
| Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ... | 145 |
| Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series..... | 153 |
| Tomasz Szubert: Factors differentiating the level of satisfaction with life and the life's values of people with and without disabilities in the light of the "Social Diagnosis" survey | 162 |
| Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures | 173 |

| | |
|--|-----|
| Wojciech Roszka: Joint characteristics' estimation of variables not jointly observed..... | 181 |
| Justyna Brzezińska: Visualizing categorical data in \mathbf{R} | 190 |
| Agata Sielska: Regional diversity of competitiveness potential of Polish farms after the accession to the European Union | 200 |
| Mariusz Kubus: Regularized linear probability model as a filter | 208 |
| Beata Basiura: The Ward method in the application for classification of Polish voivodeships with different distances..... | 216 |
| Katarzyna Wardzińska: Application of Data Envelopment Analysis in company classification process..... | 225 |
| Katarzyna Dębowska: Modeling corporate bankruptcy based on unbalanced samples | 234 |
| Danuta Tarka: Influence of the features selection method on the results of objects classification using environmental data..... | 245 |
| Artur Czech: Application of chosen methods for the selection of diagnostic variables in indirect consumption research..... | 254 |
| Beata Bal-Domańska: Assessment of relations occurring between smart growth and economic cohesion in regional dimension using panel models | 263 |
| Mariola Chrzanowska: Ordinary kriging and inverse distance weighting as methods of estimating prices based on Warsaw real estate market | 271 |
| Adam Depta: Application of analysis of variance in the study of the quality of life based on questionnaire SF-36v2 | 280 |
| Maciej Beręsewicz, Tomasz Klimanek: Using indirect estimation with spatial autocorrelation in dwelling price surveys..... | 290 |
| Karolina Paradysz: Benchmark analysis of small area estimation on local labor markets | 300 |
| Anna Gryko-Nikitin: Selection of various parameters of parallel evolutionary algorithm for knapsack problems | 310 |
| Tomasz Ząbkowski, Piotr Jałowiecki: Application of association rules for the survey of data analysis in the selected areas of logistics in food processing companies | 320 |
| Agnieszka Przedborska, Małgorzata Misztal: Using multivariate statistical methods to assess the capacity of the knee joint among the patients treated surgically for osteoarthritis | 330 |
| Dorota Perło: Sustainable development in the economic, social and environmental dimensions – spatial analysis..... | 341 |
| Ewa Putek-Szeląg, Urszula Gieraltowska: Analysis and diagnosis of the volume of renewable energy production in Poland compared to EU countries | 352 |

Tomasz Ząbkowski, Piotr Jałowiecki

Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

ZASTOSOWANIE REGUŁ ASOCJACYJNYCH DO ANALIZY DANYCH ANKIETOWYCH W WYBRANYCH OBSZARACH LOGISTYKI PRZEDSIĘBIORSTW PRZETWÓRSTWA ROLNO-SPOŻYWCZEGO

Streszczenie: Artykuł prezentuje wybrany fragment badań dotyczący wykorzystania reguł asocjacyjnych do analizy danych ankietowych na temat organizacji logistyki w przedsiębiorstwach przetwórstwa rolno-spożywczego. Wykorzystanie tej techniki wynikało m.in. z dużej ilości dostępnych danych ankietowych. Nie bez znaczenia był również fakt, że otrzymane reguły w sposób niezwykle czytelny prezentują zależności znalezione w danych. W badaniu odkryto wiele reguł, które mogą stanowić cenne źródło informacji o kondycji finansowej, skali inwestycji oraz organizacji logistyki w badanych przedsiębiorstwach.

Słowa kluczowe: reguły asocjacyjne, badania ankietowe, logistyka.

1. Cel badań

Celem prezentowanych badań było zastosowanie reguł asocjacyjnych do analizy danych ankietowych pochodzących z badania dotyczącego organizacji logistyki w przedsiębiorstwach przetwórstwa rolno-spożywczego. Wykorzystanie tej techniki w pracy wynika z kilku przesłanek. Po pierwsze, w przypadku analiz dużych zbiorów danych nasuwa się pytanie, w jaki sposób racjonalnie wykorzystać zgromadzone dane. Może to stanowić problem, zwłaszcza wtedy, gdy do analizy mamy setki ankiet, a każda z nich zawiera odpowiedzi na dziesiątki pytań, często z możliwością wielokrotnego wyboru. Po drugie, wykorzystanie reguł asocjacyjnych stwarza możliwości odkrycia zupełnie nowych i nieznanych dotąd zależności (asocjacji) pomiędzy badanymi obiektami, którymi są odpowiedzi respondentów. Po trzecie, nie bez znaczenia jest fakt, że wyniki algorytmu asocjacyjnego można przedstawić w postaci prostych reguł opisujących znalezione zależności, co pozwala w sposób czytelny opisać najważniejsze prawidłowości występujące w wybranych obszarach logistyki przedsiębiorstw.

Przeprowadzone badanie, poza aspektem poznawczym, ma również zastosowanie praktyczne. Stanowi ono cenne źródło informacji o kondycji finansowej, skali inwestycji oraz organizacji logistyki w przedsiębiorstwach przetwórstwa rolno-spożywczego, co może zostać wykorzystane do identyfikacji pożądaných kierunków doskonalenia istniejących rozwiązań w zakresie organizacji i zarządzania oraz umożliwia zaproponowanie systemowych rozwiązań poprawiających ich funkcjonowanie.

2. Dane empiryczne

Dane źródłowe pochodzą z badań ankietowych przeprowadzonych wśród przedsiębiorstw przetwórstwa rolno-spożywczego. Ankiety zostały rozesłane na przełomie 2009 i 2010 r. do 10 tys. losowo wybranych przedsiębiorstw zajmujących się przetwórstwem mięsa, mleka, zbóż oraz przetwórstwem owoców i warzyw, przy czym odpowiedzi uzyskano z 507 przedsiębiorstw. Ankieta wystosowana do przedsiębiorstw obejmowała 58 pytań o zróżnicowanej formie (głównie pytania zamknięte), które podzielono na 7 obszarów tematycznych, tj. (1) ogólne informacje o przedsiębiorstwie; (2) organizacja i zarządzanie logistyką; (3) zapasy; (4) magazynowanie i magazyny; (5) opakowania i logistyka zwrotna; (6) transport; (7) zarządzanie informacją i informatyka. Badane przedsiębiorstwa zajmowały się produkcją wyrobów piekarskich i mącznych (42%), przetwórstwem mięsa (23%), produkcją pozostałych artykułów spożywczych (9%), wytwarzaniem produktów przemiału zbóż i skrobi (8%), przetwórstwem owoców i warzyw (6%), wytwarzaniem wyrobów mleczarskich (5%), produkcją gotowych pasz dla zwierząt (3%), produkcją napojów (3%), produkcją olejów i tłuszczów (1%). W badanej próbie przedsiębiorstw według liczby zatrudnionych odnotowano przedsiębiorstwa mikro (do 9 osób), małe (od 10 do 49 osób), średnie (od 50 do 249 osób) i duże (powyżej 250 osób). Strukturę przedsiębiorstw według liczby zatrudnionych prezentuje tab. 1.

Tabela 1. Struktura próby przedsiębiorstw przetwórstwa rolno-spożywczego według wielkości

| Wyszczególnienie | Mikro | Małe | Średnie | Duże | Ogółem |
|------------------------|-------|------|---------|------|--------|
| Liczba przedsiębiorstw | 61 | 317 | 100 | 29 | 507 |
| Udział w próbie (w %) | 12 | 62 | 20 | 6 | 100 |

Źródło: badania własne.

Spośród 58 pytań w ankiecie, aby uniknąć problemu z analizą zbyt wielu wymiarów, do dalszej analizy wybrano 9, które opisywały m.in. wielkość firmy, skalę inwestycji, sytuację finansową na tle branży, koszty logistyki, jakość obsługi logistycznej, plany i zamierzenia w zakresie logistyki, rynki zbytu, rynki zaopatrzenia, planowanie produkcji i zaopatrzenia. Szczegółowe zestawienie badanych obszarów wraz z kategoriami zawiera tab. 2.

Tabela 2. Badane obszary w zakresie organizacji logistyki

| Badany obszar | Kategoria | Kod |
|--|--|--------|
| 1 | 2 | 3 |
| Wielkość firmy (liczba zatrudnionych) | Do 9 osób | P3/1 |
| | Od 10 do 49 osób | P3/2 |
| | Od 50 do 249 osób | P3/3 |
| | 250 osób i więcej | P3/4 |
| Skala procesów inwestycyjnych | Inwestycje przewyższają wartość zużycia majątku trwałego | P4/1 |
| | Inwestycje wystarczają na odtworzenie majątku trwałego | P4/2 |
| | Inwestycje są poniżej wartości zużycia majątku trwałego | P4/3 |
| | Nie inwestujemy w majątek trwały | P4/4 |
| Sytuacja finansowa na tle branży | Bardzo dobra | P5/1 |
| | Raczej dobra | P5/2 |
| | Raczej zła | P5/3 |
| | Bardzo zła | P5/4 |
| Udział kosztów logistyki w całkowitych kosztach | Bardzo wysoki udział (powyżej 15%) | P12/1 |
| | Raczej wysoki udział (10-14%) | P12/2 |
| | Raczej niski udział (5-9%) | P12/3 |
| | Bardzo niski udział (1-4%) | P12/4 |
| Miary poziomu obsługi logistycznej klientów (obsługi dostawczej) | Czas realizacji zamówienia | P13A/1 |
| | Dostępność zamówionego towaru bezpośrednio z magazynu | P13B/1 |
| | Procent w pełni poprawnie zrealizowanych zamówień | P13C/1 |
| | Nie mierzymy poziomu obsługi klienta | P13D/1 |
| Plany i zamierzenia w zakresie logistyki | Wdrożenie nowoczesnych rozwiązań informatycznych | P15A/1 |
| | Wdrożenie rachunku kosztów logistyki | P15B/1 |
| | Inwestycje w powierzchnie magazynowe | P15C/1 |
| | Inwestycje w środki transportu | P15D/1 |
| | Outsourcing magazynowania (przekazanie innej firmie) | P15E/1 |
| | Outsourcing transportu (przekazanie transportu innej firmie) | P15F/1 |
| | Inwestycje w urządzenia do pakowania | P15G/1 |
| | Centralizacja zadań przez wyodrębnienie działu logistyki | P15H/1 |
| | Udział w specjalistycznych kursach, szkoleniach | P15I/1 |
| Inne | P15J/1 | |

Tabela 2, cd.

| 1 | 2 | 3 |
|--|--|-------|
| Główne rynki zbytu | Lokalne (kilka powiatów) | P41/1 |
| | Regionalne (kilka województw) | P41/2 |
| | Obejmują cały kraj | P41/3 |
| | Obejmują Polskę i kraje sąsiadujące | P41/4 |
| | Obejmują Polskę i kraje UE | P41/5 |
| | Obejmują cały świat | P41/6 |
| Główne rynki zaopatrzenia | Lokalne (kilka powiatów) | P42/1 |
| | Regionalne (kilka województw) | P42/2 |
| | Obejmują cały kraj | P42/3 |
| | Obejmują Polskę i kraje sąsiadujące | P42/4 |
| | Obejmują Polskę i kraje UE | P42/5 |
| | Obejmują cały świat | P42/6 |
| Do planowania produkcji i wielkości zapotrzebowania na surowce i wyroby gotowe wykorzystywane są | Tylko dane archiwalne z firmy | P55/1 |
| | Tylko dane z opracowań o rynku | P55/2 |
| | Dane z firmy i dane z rynku (w tym prognozy rynkowe) | P55/3 |
| | Nie sporządzamy formalnych prognoz | P55/4 |
| | Produkujemy w zależności od podaży surowca | P55/5 |
| | Produkujemy na podstawie otrzymywanych zamówień | P55/6 |

Źródło: opracowanie własne.

3. Zastosowane techniki (reguły asocjacyjne)

Eksploracja reguł asocjacyjnych wykorzystywana jest do analizy problemów w wielu dziedzinach, m.in. w nauce, marketingu bezpośrednim, handlu elektronicznym oraz wielu innych. Najbardziej znanym przykładem wykorzystania asocjacji jest analiza koszyka zakupów (*market basket analysis*), jednak algorytmy te znajdują zastosowanie wszędzie tam, gdzie współwystępują pewne dobra, usługi, zdarzenia, charakterystyki, w formie koszyków, por. m.in. [Kukliński, Śniegocka-Łusiewicz 2009; Kurzawa, Wysocki 2008; Lasek i in. 2008; Migdał-Najman 2011; Pasztyła 2005]. Koszyki mogą stanowić więc np. zbiór transakcji kartą płatniczą, zbiór danych transakcyjnych (paragonów), zbiór produktów bankowych lub ubezpieczeniowych, zbiór cech podmiotów gospodarczych czy też zbiór odpowiedzi z badań ankietowych.

Wynikiem procesu odkrywania asocjacji w danych jest zbiór reguł asocjacyjnych opisujących znalezione zależności następującej postaci: *JEŻELI A [poprzednik], TO B [następnik]*. Jeśli określona transakcja (rekord), czyli pojedynczy przy-

padek, pasuje do reguły, co oznacza, że spełnia warunki poprzednika i następnika, to wtedy możemy mówić, że reguła zawiera tę transakcję lub że transakcja wspiera regułę asocjacji.

Do oceny reguł stosuje się szereg miar, m.in. [Tan i in. 2005; Pasztyła 2005]:

(1) Wsparcie reguły, *support* ($A \rightarrow B$) – oznacza stosunek liczby transakcji zawierających daną regułę do ogółu transakcji:

$$\text{wsparcie (jeśli } A \text{ to } B) = \frac{\text{liczba wystąpień } A \text{ i } B}{\text{liczba obserwacji w zbiorze}};$$

(2) Ufność reguły, *confidence* ($A \rightarrow B$) – oznacza stosunek liczby transakcji zawierających daną regułę do liczby transakcji zawierających dany element:

$$\text{ufność (jeśli } A \text{ to } B) = \frac{\text{liczba wystąpień } A \text{ i } B}{\text{liczba obserwacji } A \text{ w zbiorze}};$$

(3) Przyrost reguły, *lift* ($A \rightarrow B$) – informuje o tym, jaki jest wpływ elementu A na występowanie elementu B. Jest to miara często wykorzystywana przez profesjonalne systemy drążenia danych. Przyrost równy 1 oznacza, że zdarzenia są niezależne, większy od 1 wskazuje natomiast na pozytywne skorelowanie zdarzeń. Przyrost wyraża się jako:

$$\text{przyrost (jeśli } A \text{ to } B) = \frac{\text{ufność (jeśli } A \text{ to } B)}{\text{liczba obserwacji } B \text{ w zbiorze}}.$$

Najogólniej problem odkrywania reguł asocjacyjnych w danych sprowadza się do wygenerowania wszystkich tych reguł, które posiadają pewne minimalne miary wsparcia i pewności ustalone na wstępie celem redukcji problemu do mniejszego podzbioru.

W niniejszym badaniu został wykorzystany algorytm Apriori [Agrawal i Srikanth 1994]. Jest on jednym z najpopularniejszych obecnie algorytmów i jego istota sprowadza się do wygenerowania zbiorów częstych, a następnie konstrukcji reguł asocjacyjnych z tych właśnie zbiorów. Algorytm ten jest algorytmem iteracyjnym, który w kolejnych krokach znajduje zbiory częste o rozmiarach 1, 2, ..., k . Pierwszym etapem algorytmu jest wyodrębnienie z danych wszystkich zbiorów jednoelementowych, które występują w transakcjach, i sprawdzenie, które z nich są częste, tzn. posiadają co najmniej minimalne wsparcie. W kolejnym etapie, na podstawie zbiorów częstych, algorytm generuje zbiory kandydujące dwuelementowe, które, potencjalnie, mogą być zbiorami częstymi. Dla każdego wygenerowanego zbioru kandydującego obliczane jest jego wsparcie w bazie danych i jeśli spełnia on warunek minimalnego wsparcia, trafia do listy zbiorów częstych i w kolejnym kroku zostanie on wykorzystany do generowania zbiorów kandydujących trzelementowych. Następnie zbiory częste trzelementowe są wykorzystywane do generowania zbiorów kandydujących czteroelementowych itd. Działanie algorytmu kończy się,

gdy nie można już wygenerować kolejnych zbiorów kandydujących, a wynikiem działania algorytmu jest suma k -elementowych zbiorów częstych ($k = 1, 2, \dots$).

4. Badanie

Tabela 3. Format danych wymaganych przez algorytm

| Nr ankiety | Odpowiedź |
|------------|-----------|
| 1 | P15D/1 |
| 1 | P41/3 |
| 1 | P42/3 |
| 1 | P55/4 |
| 2 | P3/1 |
| 2 | P4/2 |
| 2 | P5/2 |
| 2 | P12/4 |

Źródło: badania własne.

Dane pochodzące z badań ankietowych zostały zestawione w tab. 3, w której kolumny zawierały odpowiedzi na pytania, natomiast wiersze odpowiadały poszczególnym ankietom przypisanym do przedsiębiorstw. Konieczna była transpozycja danych do formatu transakcyjnego, aby w jednej kolumnie mieć informację o transakcji (numerze ankiety), zaś w drugiej kolumnie odpowiedzi na pytania (por. tab. 3). W pracy rozważano jedynie reguły dwuelementowe. Ustalony został także minimalny poziom wsparcia na poziomie 3% oraz pewności reguły na poziomie 30%.

Otrzymano reguły asocjacyjne, których część zawiera tabela wynikowa (por. tab. 4)

Tabela 4. Wybrane reguły asocjacyjne według wsparcia (support)

| Przyrost | Wsparcie (%) | Ufność (%) | Reguła |
|----------|--------------|------------|----------------|
| 1,02 | 45,01 | 72,56 | P3/2=>P5/2 |
| 1,02 | 45,01 | 63,01 | P5/2=> P3/2 |
| 1,11 | 41,88 | 79,55 | P15D/1=>P5/2 |
| 1,11 | 41,88 | 58,63 | P5/2=> P15D/1 |
| 1,22 | 35,42 | 75,42 | P41/1=>P3/2 |
| 1,22 | 35,42 | 57,10 | P3/2=> P41/1 |
| 1,04 | 32,68 | 52,68 | P3/2=>P13D/1 |
| 1,04 | 32,68 | 64,73 | P13D/1 => P3/2 |
| 1,02 | 31,90 | 73,09 | P55/6=>P5/2 |
| 1,02 | 31,90 | 44,66 | P5/2=> P55/6 |
| 1,76 | 29,94 | 63,75 | P41/1=>P42/1 |
| 1,76 | 29,94 | 82,70 | P42/1=>P41/1 |
| 1,08 | 29,35 | 47,32 | P3/2=>P55/6 |
| 1,08 | 29,35 | 67,26 | P55/6 => P3/2 |

Źródło: badania własne na podstawie obliczeń z programu SAS Enterprise Miner.

Na podstawie uzyskanych wyników można stwierdzić, że najczęściej występującym schematem odpowiedzi wśród analizowanych obszarów była reguła $P3/2 \Rightarrow P5/2$ (por. opis reguł w tab. 2). Reguła ta mówi, że firmy małe wskazywały na „raczej dobrą”, ich zdaniem, sytuację finansową. W szczególności na podstawie pewności reguły (confidence) 72,56% małych firm oceniało, że ich sytuacja finansowa jest „raczej dobra”. Reguła odwrotna postaci $P5/2 \Rightarrow P3/2$ posiada ufność na poziomie 63,01%, co oznacza, że 63% spośród oceniających sytuację finansową na „raczej dobrą” stanowiły firmy małe.

Kolejną często występującą regułą było powiązanie $P15D/1 \Rightarrow P5/2$. Reguła ta mówi, że plany inwestycyjne w środki transportu były najczęściej wskazywane (79,55%) wśród firm, które oceniły sytuację finansową jako „raczej dobrą”. Z kolei reguła odwrotna postaci $P5/2 \Rightarrow P15D/1$ posiada ufność na poziomie 58,63%, co oznacza, że prawie 59% spośród firm planujących inwestycje w środki transportu stanowiły firmy o „raczej dobrej” kondycji finansowej.

Trzecim najczęściej występującym schematem odpowiedzi w ankietach była reguła postaci $P41/1 \Rightarrow P3/2$. Oznacza ona, że lokalne rynki zbytu były najczęściej wskazywane (75,42%) wśród firm małych. Tym samym, rozpatrując regułę odwrotną, można zauważyć, że prawie 57,1% spośród firm z lokalnymi rynkami zbytu stanowiły firmy małe.

Dokonując porządkowania reguł malejąco według miary wyrażającej ufność, możemy zbudować interesujący obraz analizowanych przedsiębiorstw, por. tab. 5.

Tabela 5. Wybrane reguły asocjacyjne według pewności (confidence)

| Przyrost | Wsparcie (%) | Ufność (%) | Reguła |
|----------|--------------|------------|----------------------------|
| 1,18 | 25,83 | 84,62 | $P4/2 \Rightarrow P5/2$ |
| 1,76 | 29,94 | 82,70 | $P42/1 \Rightarrow P41/1$ |
| 1,12 | 14,87 | 80,00 | $P42/3 \Rightarrow P5/2$ |
| 1,11 | 41,88 | 79,55 | $P15D/1 \Rightarrow P5/2$ |
| 1,10 | 23,29 | 78,81 | $P42/2 \Rightarrow P5/2$ |
| 1,09 | 18,20 | 78,15 | $P12/2 \Rightarrow P5/2$ |
| 1,08 | 15,46 | 77,45 | $P55/3 \Rightarrow P5/2$ |
| 1,08 | 14,68 | 77,32 | $P4/1 \Rightarrow P5/2$ |
| 1,06 | 8,02 | 75,93 | $P41/5 \Rightarrow P5/2$ |
| 1,22 | 35,42 | 75,42 | $P41/1 \Rightarrow P3/2$ |
| 1,06 | 19,77 | 75,37 | $P15C/1 \Rightarrow P5/2$ |
| 1,05 | 6,46 | 75,00 | $P42/5 \Rightarrow P5/2$ |
| 1,48 | 8,61 | 74,58 | $P55/4 \Rightarrow P13D/1$ |

Źródło: badania własne na podstawie obliczeń z programu SAS Enterprise Miner.

W szczególności możemy zauważyć, że:

- 84,62% spośród firm, które oceniły, że inwestycje wystarczają na odtworzenie majątku trwałego, to były te, które określiły swoją sytuację finansową jako „raczej dobrą” (P4/2 => P5/2);
- 82,70% spośród firm wskazujących na lokalne rynki zaopatrzenia to te, które operowały również na lokalnych rynkach zbytu (P42/1 => P41/1);
- 80% wśród firm zaopatrujących się w całym kraju to te, które określiły swoją sytuację finansową jako „raczej dobrą” (P42/3 => P5/2);
- plany inwestycyjne w środki transportu były najczęściej wskazywane (79,55%) wśród firm, które oceniły sytuację finansową jako „raczej dobrą” (P15D/1 => P5/2);
- 78,81% firm wskazujących na regionalny zakres rynku zaopatrzenia to te, które określiły swoją sytuację finansową jako „raczej dobrą” (P42/2 => P5/2);
- 78,15% firm, które określiły udział kosztów logistyki w całkowitych kosztach jako „raczej wysoki (10–14%)”, to te, które oceniły swoją sytuację finansową jako „raczej dobrą” (P12/2 => P5/2);
- 77,45% firm, które do planowania produkcji wykorzystują dane z firmy i dane z rynku (w tym prognozy rynkowe), to te, które oceniły swoją sytuację finansową jako „raczej dobrą” (P55/3 => P5/2).

Równie ciekawe spostrzeżenia nasuwają się w efekcie zawężenia reguł, biorąc pod uwagę wielkość firmy. Rozpatrzmy przykładowo kategorię P55, czyli jakiego typu dane i informacje wykorzystują firmy do planowania produkcji i wielkości zapotrzebowania na surowce i wyroby gotowe (por. tab. 6). W przypadku tej kategorii możemy zauważyć, że firmy mikro (P3/1) i małe (P3/2) planowały produkcję na podstawie otrzymywanych zamówień (P55/3) i stanowiło to 44,26% oraz 47,32% odpowiednio dla firm mikro i małych. Z kolei firmy średnie (P3/3) i duże (P3/4) do planowania produkcji brały pod uwagę dane z firmy oraz dane z rynku, w tym prognozy rynkowe (P55/3). Wśród firm średnich było to 34%, natomiast wśród firm dużych było to 72,41%.

Tabela 6. Wybrane reguły według wielkości firmy dla kategorii P55 (sposób planowania produkcji)

| Przyrost | Wsparcie (%) | Ufność (%) | Reguła |
|----------|--------------|------------|---------------|
| 1,01 | 5,28 | 44,26 | P3/1 => P55/6 |
| 1,08 | 29,35 | 47,32 | P3/2 => P55/6 |
| 1,70 | 6,65 | 34,00 | P3/3 => P55/3 |
| 3,63 | 4,11 | 72,41 | P3/4 => P55/3 |

Źródło: badania własne.

Podobna analiza została również przeprowadzona dla kategorii opisującej sytuację finansową firm (P5). W tym przypadku otrzymano następujące reguły (por. tab. 7).

Tabela 7. Wybrane reguły według wielkości firmy dla kategorii P5 (sytuacja finansowa na tle branży)

| Przyrost | Wsparcie (%) | Ufność (%) | Reguła |
|----------|--------------|------------|--------------|
| 1,95 | 3,91 | 32,79 | P3/1 => P5/3 |
| 1,02 | 45,01 | 72,56 | P3/2 => P5/2 |
| 1,04 | 14,48 | 74,00 | P3/3 => P5/2 |
| 1,16 | 4,70 | 82,76 | P3/4 => P5/2 |

Źródło: badania własne.

Wśród mikrofirm 32,79% z nich określiło swoją sytuację finansową jako „raczej złą”. Wśród firm małych 72,56% z nich określało swoją sytuację finansową jako „raczej dobrą”. Z kolei 74% firm średnich i aż 82,76% firm dużych oceniło swoją sytuację finansową jako „raczej dobrą”.

5. Wnioski

Na podstawie przeprowadzonych badań można sformułować następujące wnioski.

W szczególności możemy stwierdzić, że objęte badaniem małe, średnie i duże przedsiębiorstwa oceniały swoją sytuację finansową jako dość dobrą, w przeciwieństwie do mikrofirm, które wskazywały na raczej złą sytuację.

Biorąc pod uwagę plany inwestycyjne, należy stwierdzić, że środki transportu były najczęściej wskazywane jako obszar inwestycji wśród firm, które oceniły sytuację finansową jako „raczej dobrą”.

Skala zasięgu badanych firm, jeśli chodzi o główne rynki zbytu i zaopatrzenia, była niewielka (lokalne rynki zbytu były najczęściej wskazywane przez firmy małe).

Ciekawych spostrzeżeń dostarczają reguły eksplorujące odpowiedzi na pytanie, jakiego typu dane i informacje wykorzystują firmy do planowania produkcji. W przypadku tej kategorii możemy zauważyć, że firmy mikro i małe planowały produkcję na podstawie otrzymywanych na bieżąco zamówień. Z kolei firmy średnie i duże do planowania produkcji brały pod uwagę dane z firmy oraz dane z rynku, w tym także prognozy rynkowe.

Przytoczone powyżej reguły główne oraz wiele innych, pomniejszych stanowią cenne źródło informacji o kondycji finansowej, skali inwestycji oraz organizacji logistyki w badanych przedsiębiorstwach. Nie bez znaczenia pozostaje fakt, że otrzymane reguły w sposób niezwykle przystępny ujmują najważniejsze prawidłowości w bardzo dużym zbiorze danych ankietowych. Stąd celem dalszych badań autorów będzie zastosowanie analizy powiązań (*link analysis*), by wizualnie, za pomocą grafów, przedstawić zależności pomiędzy odpowiedziami respondentów.

Literatura

- Agrawal R., Srikant R. (1994), *Fast Algorithms for Mining Association Rules*, IBM Research Report RJ9839, IBM Almaden Research Center San Jose, California.
- Kukliński M., Śniegocka-Lusiewicz M. (2009), *Miary asocjacji w analizie koszykowej – przykład empiryczny*, *Acta Universitatis Nicolai Copernici. Oeconomia*, 389, 307-316.
- Kurzawa I., Wysocki F. (2008), *Wykorzystanie analizy koszykowej do identyfikacji zachowań konsumpcyjnych gospodarstw domowych w Polsce*, [w:] *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu* nr 7(1207), *Taksonomia* 15, 527-534.
- Lasek M., Nowak E., Pęczkowski M. (2008), *Zastosowanie reguł asocjacji i sekwencji zdarzeń do analizy działalności inwestycyjnej gospodarstw agroturystycznych*, „*Turyzm*” 18/2, 57-73.
- Migdał-Najman K. (2011), *Analiza porównawcza samouczących się sieci neuronowych typu SOM i GNG w poszukiwaniu reguł asocjacyjnych*, [w:] *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu* nr 176, *Taksonomia* 18, 272-281.
- Paszyła A. (2005), *Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej. Data mining: poznaj siebie i swoich klientów* (pub. elektr.) Statsoft, Kraków.
- Tan P., Steinbach M., Kumar V. (2005), *Introduction to Data Mining*, Addison-Wesley, Boston.

APPLICATION OF ASSOCIATION RULES FOR THE SURVEY OF DATA ANALYSIS IN THE SELECTED AREAS OF LOGISTICS IN FOOD PROCESSING COMPANIES

Summary: This paper presents a selected part of the research with association rules application to the survey data exploring the organization of logistics in food processing companies. The application of association rules is due to the large volume of available survey data. Furthermore, the other important aspect was the possibility to present rules in a very clear and meaningful way. The analysis resulted in a number of interesting rules that can be a valuable source of information on companies' financial condition, the scale of investments and organization of logistics.

Keywords: association rules, survey data, logistics.