METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XIX/2, 2018, s. 105 – 116

Received: 12.06.2018
Accepted: 05.07.2018

# BADANIE PREFERENCJI PRZEDSIĘBIORSTW W STOSOWANIU ZAAWANSOWANYCH METOD ANALIZY DANYCH

**Magdalena Barska (ORCID: 0000-0002-6410-7929)**
Kolegium Analiz Ekonomicznych
Szkoła Główna Handlowa w Warszawie
e-mail: d09a1997@doktorant.sgh.waw.pl

**Streszczenie:** Potrzeby firm w zakresie stosowania zaawansowanych metod przetwarzania danych są różne w zależności od branży funkcjonowania, możliwości finansowania, zachowań konkurencji, rozmiaru i zmienności gromadzonych informacji. W pewnych przypadkach technologie business intelligence, wizualizacja lub metody statystyczne stają się niezbędne do funkcjonowania firmy, w innych są sposobem zwiększenia wydajności oraz uzyskania przewagi konkurencyjnej. Celem publikacji jest analiza różnic w podejściu przedsiębiorstw do stosowania tych technologii. Sprawdzono, czy istnieją cechy powodujące, że dana grupa jest podatna na ofertę związaną z big data i data science. Realizacji tego celu służy analiza skupień, pozwalająca na wyznaczenie grup klientów o podobnej charakterystyce. Wyniki badania wskazują, że źródłem różnic są cechy demograficzne, odmienne oczekiwania oraz dotychczasowe doświadczenia.

**Słowa kluczowe:** analiza skupień, modele mieszanin rozkładów normalnych

**JEL classification:** C10

## WSTĘP

Termin big data dotyczy analizy zbiorów danych o dużych rozmiarach i zmienności. Raport McKinsey Global Institute [Manyjka 2011] jako elementy tego systemu wymienia: techniki analizy oparte na uczeniu maszynowym i przetwarzaniu języka naturalnego, technologie business intelligence, hurtownie danych, przetwarzanie danych w chmurze, graficzną prezentację danych. Termin data science obejmuje pozyskiwanie i eksplorację danych oraz wnioskowanie na ich podstawie przy użyciu modelowania predykcyjnego. Wzrost zainteresowania

DOI:10.22630/MIBE.2018.19.2.10

firm nowoczesnymi metodami przetwarzania danych ma różne przyczyny. Dysponując dużymi zbiorami informacji, przedsiębiorstwa poszukują metod ich eksploracji i wizualizacji. Konieczność wdrożenia nowych metod analizy wynika ze zmiennosci danych, ich nieustrukturyzowania oraz potrzeby przetwarzania w czasie rzeczywistym. Preferencje firm w tym zakresie kształtuje branża funkcjonowania, możliwości finansowania, konkurencji na rynku, zmienność i wolumen danych. Przeszkodą we wdrożeniu bywają koszty albo odmienne cele strategiczne. Raport [Eurostat 2017] podaje, że w 2016 ponad 75% firm Unii Europejskiej zatrudniających co najmniej 10 pracowników posiadało stronę internetową, a ponad połowa aktywnie korzystała z mediów społecznościowych. Jedynie 10% przedsiębiorstw korzystało z big data. Niemal połowa spośród tej grupy pozyskiwała dane o lokalizacji z urządzeń mobilnych, a 45% dane pochodzące z sieci społecznościowych. Najwyższy odsetek firm deklarujących stosowanie big data odnotowano dla Malty i Holandii (19%), najniższy dla Niemiec i Polski (6%) i Cypru (3%).

Celem publikacji jest analiza różnic w podejściu do stosowania zaawansowanych metod analizy danych. Pod uwagę wzięto firmy, które wykazały zainteresowanie tą dziedziną poprzez uczestnictwo w targach i śledzenie publikacji branżowych. Sprawdzono, czy istnieją cechy powodujące, że dana grupa jest podatna na ofertę związaną z nowoczesnymi technologiami. Realizacji celu służy analiza skupień, pozwalająca na wyznaczenie segmentów o podobnej charakterystyce.

## BADANIE PREFERENCJI RESPONDENTÓW

Preferencje można badać na podstawie ankiety. Warunkiem uzyskania rzetelnej informacji jest właściwe sformułowanie pytań. Do zasad konstruowania ankiety należą: niezbędność i zrozumiałość pytań, ich odpowiedni układ oraz uporządkowanie tematyczne, stosowanie pytań filtrujących [Churchill 2002]. Zbyt mała lub zbyt duża liczba wariantów odpowiedzi może powodować uzyskanie nierzetelnych danych. Uwzględnienie skali ma na celu pomiar natężenia zjawiska. Odpowiedzi na pytania otwarte wymagają interpretacji. O uzyskaniu wiarygodnych rezultatów decyduje reprezentatywność grupy. Zależy ona od liczebności próby oraz sposobu doboru respondentów. Nielosowy dobór próby polega na wyodrębnieniu jednostek o pożądanych charakterystykach i jest subiektywny [Frankfort-Nachmias i Nachmias 2010]. Dobór celowy nie daje teoretycznych podstaw do uogólnienia rezultatu.

## METODY SEGMENTACJI

Podstawą przystąpienia do segmentacji jest eliminacja współliniowości czynników wpływu. Analiza głównych składowych pozwala na przekształcenie zmiennych obserwowalnych w nieskorelowane zmienne nieobserwowalne [Gatnar

i Walesiak 2004]. Wariancje kolejnych składowych są miarą ich zasobów informacyjnych o zjawisku. Uporządkowane są tak, aby wariancje były coraz mniejsze. Zwykle kilka pierwszych składowych dostarcza większość informacji o zjawisku, co pozwala na redukcję ich liczby przy małej stracie informacji. Przyjmując, że $X = [x_{ji}]$, $j$=1,…,$m$; $i$=1,2,…,$n$ jest macierzą zmiennych wejściowych, model przyjmuje postać:

$$F = WZ \tag{1}$$

gdzie: F ($[f_{si}]$) - macierz głównych składowych o wymiarze ($m$ x $n$), W ($[w_{js}]$) - macierz współczynników głównych składowych o wymiarze ($m$ x $g$), Z ($[z_{ji}]$) - macierz standaryzowanych zmiennych wejściowych o wymiarze ($g$ x $n$).

oraz główne składowe są niezależne, to jest: $w_s^T w_{s'} = 0;\ \ s', s = 1,2,\dots,g;\ s' \neq s$.

Powszechnie stosowanym narzędziem segmentacji jest analiza skupień, która pozwala na wyodrębnienie grup o zbliżonej charakterystyce, co oprócz spełnienia funkcji poznawczych umożliwia dopasowanie strategii marketingowej i oferty przez dostawców czy instytuty badawcze. Segmenty powinny być bardziej homogeniczne niż cała populacja. Jednymi z popularniejszych metod analizy są metody hierarchiczne, dla których wyodrębnia się algorytmy aglomeracyjne i deglomeracyjne.

Na kształt skupień wpływa wybór miary odległości między obserwacjami. Metoda najdalszego sąsiedztwa generuje zwarte grupy, ale jest mało odporna na wartości odstające. Podobną tendencję wykazuje algorytm Warda, która łączy ze sobą klastry o małej liczbie obserwacji i tworzy skupienia o podobnej wielkości. Metoda opiera się na minimalizowaniu wariancji w grupie. W każdej iteracji dla istniejących skupień wyznacza się sumę kwadratów odchyleń od średniej. Miarą zróżnicowania jest błąd sumy kwadratów, wyrażony wzorem [Grabiński i in. 1989]:

$$ESS = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \tag{2}$$

gdzie: n – liczba obiektów w grupie, $x_i$ – wartość cechy charakteryzującej jednostkę i w danej grupie.

Migdał Najman i Najman [2013] przekonują o zasadności wyboru metody Warda w grupowaniu wielowymiarowych obiektów, pod warunkiem braku licznych wartości nietypowych i rzędu jednostek poniżej dziesiątek tysięcy. Skuteczność metody potwierdza też symulacja Monte Carlo, którą przeprowadzają Kutera i Lasek [2010]. Na wartości odstające odporna jest metoda średniej grupowej, dla której obserwacje są rozproszone w podobny sposób wokół skupień. Wadą wskazanych metod jest spadek efektywności wraz ze wzrostem liczby obserwacji. Kolejne iteracje nie powodują korekty błędnie przypisanych obserwacji.

Jakość hierarchicznej analizy skupień można badać wyznaczając dla każdego skupienia dwie wartości p z przedziału [0,1]: AU i BP[1]. Pierwsza wartość p wyznaczana jest na podstawie procedury bootstrap dla wielokrotnego próbkowania bez zwracania i jest lepszym przybliżeniem nieobciążonej wartości p niż BP. Dla skupień o wartości p większej niż 0,95 można odrzucić hipotezę o braku grupowanie się obserwacji na poziome ufności 0,05. Skupienie jest stabilne, tzn. występuje duże prawdopodobieństwo formowania się grupy w miarę zwiększania liczby obserwacji. Wartość p obciążona jest błędem ze względu na ograniczoną ilość prób bootstrap.

Innym narzędziem analizy skupień jest opracowana przez [MacQueena 1967] metoda k-średnich, wymagająca wyznaczenia ich liczby i środków ciężkości. Estymacji tych parametrów dokonuje się przy pomocy metod hierarchicznych, na podstawie symulacji lub wartości bayesowskiego kryterium informacyjnego (BIC) z czynnikiem karzącym za liczbę parametrów [Banfield i Raftery 1993]. Metoda jest użyteczna w przypadku dużych zbiorów danych ze względu na liniową złożoność obliczeniową [Kutera i Lasek 2010]. Obserwacje przypisywane są w kolejnych iteracjach do najbliżej położonego środka ciężkości. Przy pomocy k-średnich [Angowski i in. 2017] wyznaczają segmenty rynku produktów spożywczych w celu zbadania preferencji nabywców i dopasowania oferty do konkretnej grupy. [Pietrzykowski i Kobus 2006] wykorzystują metodę w dywersyfikacji portfela akcji.

Metoda k-średnich grupuje obserwacje na podstawie odległości i nie bazuje na modelu probabilistycznym. Badane obiekty mogą pochodzić z różnych rozkładów. Składniki mieszanin rozkładów normalnych różnią się średnią lub macierzą kowariancji, a rozkłady są zmieszane z prawdopodobieństwami $\pi_i$. Parametry mieszaniny wyznaczane są w oparciu o metodę największej wiarygodności, a maksimum funkcji wiarogodności obliczane jest przy pomocy algorytmu expectation–maximization [Biecek i in. 2012]. Różnice między modelami dotyczą parametryzacji macierzy kowariancji efektów losowych i błędów losowych w modelach mieszanych. Miarą podobieństwa dwóch rozwiązań jest skorygowany indeks Randa, przyjmujący wartości z przedziału (0,1>. O występowaniu skupień mogą świadczyć funkcje gęstości rozkładu wielowymiarowego modelu mieszanin, przy założeniu, że występuje zależność między składnikami mieszanin a występującymi skupieniami [Scrucca 2016]. Identyfikacja obszarów o wysokiej gęstości oraz formujących je obserwacji pozwala na wyznaczenie centrów skupień. Pozostałe obserwacje przypisane są na podstawie prawdopodobieństwa przynależności do danej grupy.

[1] AU – Approximately Unbiased; BP - Bootstrap Probability.

## CHARAKTERYSTYKA DANYCH

W badaniu empirycznym wykorzystano dane z 2015 roku gromadzone za pomocą ankiety na potrzeby targów poświęconych metodom analizy dużych zbiorów informacji. Wyboru respondentów dokonano metodą doboru celowego, w celu uzyskania próby bliskiej próbie reprezentatywnej. Otrzymano odpowiedzi od 1000 respondentów. Poddane maskowaniu dane prezentują ich cechy demograficzne oraz odpowiedzi na pytania związane z podejściem do wykorzystania metod analizy danych. Dotyczą obszarów takich jak: cel użycia, oczekiwania firmy, planowana strategia, zaawansowanie w użyciu metod, doświadczenie we wdrażaniu. Respondentami są przedstawiciele szeregu branż. Dominują firmy z krajów europejskich, ponad 30% ma siedzibę w Azji, pozostałe w Ameryce. Sposób doboru respondentów i konstrukcja ankiety narzucają pewne ograniczenia. Próba jest reprezentatywna dla populacji firm, które miały do czynienia z zaawansowanymi technologiami lub są zainteresowane ich wdrożeniem. Zainteresowanie to wyraża się poprzez utrzymywanie kontaktu z dostawcami rozwiązań lub śledzenie informacji o targach branżowych. Wynik badania można odnieść jedynie do firm spełniających takie kryteria.

## METODOLOGIA BADANIA

Pytania ankiety przełożono na zmienne obejmujące cechy demograficzne i stosunek do metod analizy danych. Dokonano standaryzacji zmiennych wyznaczonych z pytań otwartych. Na podstawie przesłanek teoretycznych zredukowano zmienne zbędne. Dokonano kodyfikacji zmiennych porządkowych do wartości od 0 do 5 w zależności od liczby odpowiedzi. Dla zmiennych kategorycznych bez skal porządkowych wprowadzono zmienne sztuczne. Uzyskano w ten sposób zbiór: zmienne demograficzne (region, przychody, branża, leader grupy), związane z celem wykorzystania metod analizy (minimalizacja ryzyka, rozwój nowych produktów, poprawa wyników finansowych, zainteresowanie data science, oczekiwania co do łatwości wdrożenia), związane z doświadczeniem w stosowaniu metod (wsparcie kierownictwa, współpraca z dostawcą, przebieg wdrożenia, spełnienie wymagań, zaawansowanie w użyciu technik, wzrost dochodów, strategia). W celu redukcji wymiaru danych oraz eliminacji korelacji między zmiennymi zbiór poddano analizie PCA i dokonano segmentacji w oparciu o wybrany model.

## WYNIKI BADAŃ

W oparciu o kryterium wyjaśnienia wariancji ustalone na poziomie 75%, wybrano 10 głównych składowych, które przedstawia tabela 1.

Tabela 1. Macierz głównych składowych

| Czynniki | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Ameryka Pd. | -0,017 | 0,100 | -0,397 | 0,472 | -0,121 | 0,067 | -0,058 | -0,096 | -0,112 | -0,127 |
| Ameryka Pn. | -0,006 | 0,357 | -0,270 | -0,241 | 0,160 | -0,263 | 0,186 | 0,069 | 0,110 | 0,558 |
| Azja | -0,085 | 0,352 | 0,533 | 0,251 | 0,091 | 0,257 | -0,043 | -0,014 | 0,120 | -0,139 |
| Europa | 0,090 | -0,622 | -0,120 | -0,265 | -0,145 | -0,076 | -0,068 | 0,005 | -0,139 | -0,216 |
| Przychody | 0,181 | 0,141 | -0,335 | -0,305 | 0,173 | 0,103 | -0,076 | 0,021 | 0,346 | -0,280 |
| Finanse | 0,103 | 0,381 | 0,110 | -0,243 | -0,196 | -0,191 | -0,133 | -0,005 | -0,131 | -0,471 |
| Energetyka | -0,128 | -0,072 | 0,048 | -0,008 | 0,129 | -0,027 | -0,646 | 0,534 | 0,026 | 0,239 |
| Zdrowie | -0,060 | -0,014 | 0,021 | 0,032 | 0,260 | -0,168 | 0,596 | 0,412 | -0,386 | -0,157 |
| Sprzedaż | 0,053 | -0,153 | 0,129 | -0,008 | 0,511 | -0,126 | -0,114 | -0,669 | -0,123 | 0,132 |
| Telekomunikacja | -0,013 | -0,101 | -0,032 | -0,193 | -0,176 | 0,783 | 0,227 | -0,013 | 0,020 | 0,227 |
| Usługi | -0,060 | -0,159 | 0,173 | 0,059 | -0,433 | -0,353 | 0,215 | -0,079 | 0,589 | 0,086 |
| Ubezpieczenia | 0,077 | 0,080 | -0,478 | 0,437 | -0,090 | 0,003 | -0,067 | -0,060 | 0,016 | -0,037 |
| Wzrost dochodów | 0,318 | 0,161 | -0,118 | -0,270 | 0,192 | 0,118 | 0,015 | -0,001 | 0,221 | -0,187 |
| Oczekiwania | 0,398 | -0,134 | 0,121 | 0,150 | 0,105 | 0,012 | 0,005 | 0,172 | 0,053 | 0,103 |
| Wsparcie kierownictwa | 0,364 | 0,090 | 0,136 | -0,010 | -0,269 | -0,030 | -0,031 | -0,042 | -0,237 | 0,084 |
| Data science | 0,350 | 0,141 | 0,081 | -0,001 | -0,246 | -0,071 | 0,048 | -0,054 | -0,120 | 0,219 |
| Współpraca z dostawcą | 0,189 | -0,140 | 0,037 | 0,291 | 0,301 | 0,049 | 0,189 | 0,136 | 0,347 | -0,112 |
| Przebieg wdrożenia | 0,431 | -0,155 | 0,117 | 0,141 | 0,111 | -0,045 | -0,029 | 0,141 | 0,106 | -0,023 |
| Spełnienie wymagań | 0,416 | 0,046 | -0,005 | 0,037 | -0,075 | 0,036 | -0,062 | 0,004 | -0,196 | 0,181 |

Źródło: opracowanie własne

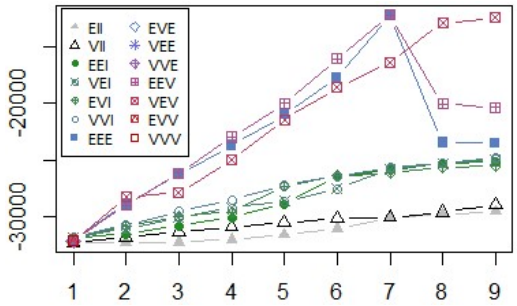
Poszczególne składowe zawierają informacje, na podstawie których można identyfikować cechy respondentów:

- Składowa 1: dominują respondenci o pozytywnych doświadczeniach we wdrażaniu nowych technologii, ich wymagania zostały spełnione, a kierownictwo firmy było przychylne przedsięwzięciu. Zastosowanie nowych rozwiązań wiązało się ze wzrostem przychodów.
- Składowa 2: respondenci z sektora finansowego, głównie z Ameryki Pn. lub Azji.
- Składowa 3: respondenci z Azji, spoza sektora ubezpieczeń, o raczej niskich przychodach.
- Składowa 4: respondenci z Ameryki Południowej lub z sektora ubezpieczeń.
- Składowa 5: respondenci z branży sprzedażowej.
- Składowa 6: respondenci reprezentujący branżę telekomunikacyjną.
- Składowa 7: respondenci reprezentujący głównie opiekę zdrowotną, rzadko sektor energetyczny.
- Składowa 8: respondenci reprezentujący służbę zdrowia lub sektor energetyczny.

- Składowa 9: respondenci z sektora usługowego, mający dobre doświadczenia we współpracy z dostawcą oraz odnotowujący wysokie przychody.
- Składowa 10: respondenci z Ameryki Północnej, reprezentujący głównie przemysł energetyczny i telekomunikację, zainteresowani data science.

Przy pomocy pakietu mclust programu R estymowano parametry modeli mieszanin rozkładów normalnych dla głównych skladowych. Najwyższe wartości bayesowskiego kryterium informacyjnego[2] uzyskano dla modeli wielowymiarowych rozkładów normalnych EEV i EEE[3] dla 7 skupień, odpowiednio -12221,3 i -12289,5 (rys. 1). Modele zakładają te same rozkłady eliptyczne dla obserwacji tworzących dane skupienie oraz jednakowy kształt i wymiary tych rozkładów. W modelu EEE dodatkowo zakłada się ich jednakowe położenie. Skorygowany indeks Randa dla modeli wynosi 0,79, co świadczy o dużym podobieństwie segmentacji uzyskanej z obu modeli.

Rysunek 1. Wartości kryterium informacyjnego BIC w zależności od liczby skupień




Źródło: opracowanie własne

Wyznaczono 7 segmentów na podstawie najlepszego modelu EEV. Ze względu na dużą liczbę głównych składowych wizualizacji skupień dokonano jedynie dla 3 pierwszych. Rysunek 2 ilustruje rozkład obiektów. Elipsy określają kształt gęstości rozkładów, a ich środki wskazują na środki skupień. Rysunek 3 ilustruje prawdopodobieństwo przypisania do danego skupienia wyznaczone dla każdej obserwacji i wyrażone wartościami z przedziału (0,1). Pogrubione symbole na rysunku wskazują na obserwacje o wysokim współczynniku niepewności zaklasyfikowania do grupy, wyznaczonym na podstawie prawdopodobieństwa przynależności do danego rozkładu. Obserwacje te występują głównie na brzegach skupień, dlatego można traktować je jako wartości odstające, nie dające się

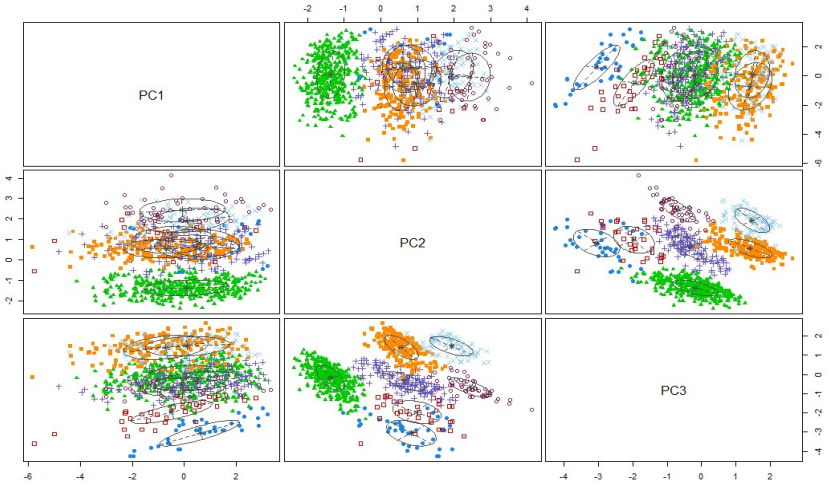
[2] Kryterium BIC jest stosowane w ocenie modeli mieszanin, w klasycznej analizie skupień powszechnie stosowany jest np. indeks Calińskiego-Harabasza.

[3] E – (ang. equal) równy, V – (ang. variable) zmienny - współczynniki dla określenia podobieństwa wymiarów, kształtu oraz położenia rozkładów dla obserwacji tworzących skupienie. Wyjaśnienie różnic między modelami m.in. u [Dang i in. 2017].

przypisać do żadnej grupy. Centra skupień pokrywają się z wierzchołkami funkcji gęstości rozkładów, z których pochodzą obserwacje.

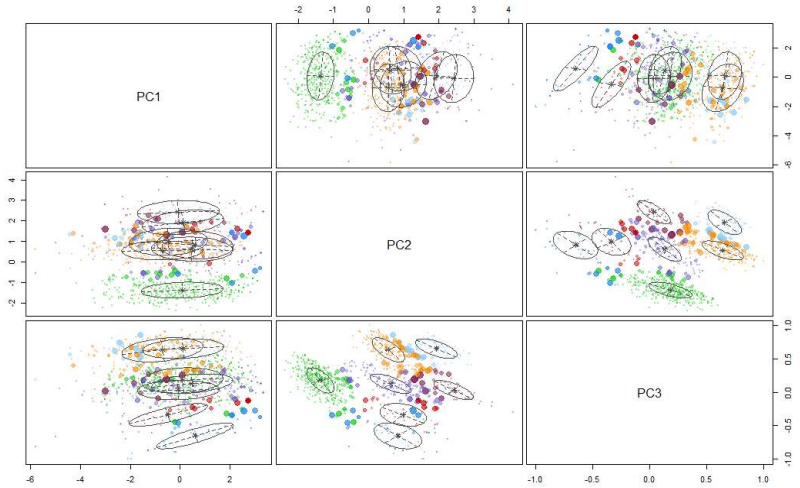
Optymalną liczbę segmentów badano też metodą Warda, ustalając ją na 7 ze względu na powolne tempo spadku wariancji w grupach dla większej liczby skupień. Wariancje w grupach w zależności od ich liczby przedstawia wykres na rysunku 4. Poziom odcięcia dla 7 skupień przedstawia dendrogram na rysunku 5.

Rysunek 2. Rozkład obiektów dla 3 głównych składowych



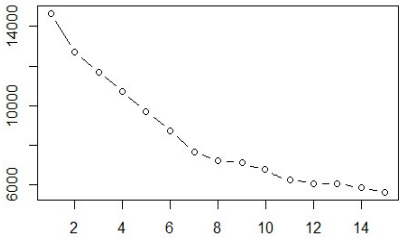

Źródło: opracowanie własne

Rysunek 3. Prawdopodobieństwo przypisania dla skupienia

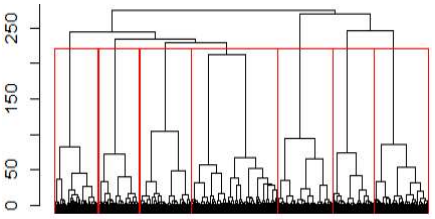



Źródło: opracowanie własne

Rysunek 4. Suma kwadratów błędów w grupach



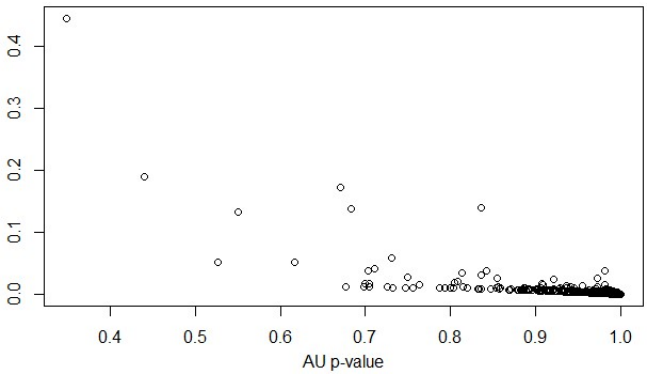

Rysunek 5. Dendrogram




Źródło: opracowanie własne

Korzystając z pakietu pvclust programu R dla każdej grupy wyodrębnionej na podstawie hierarchicznej analizy skupień wyznaczono dwie wartości p z przedziału [0,1] (AU i BP) dla 10000 prób bootstrap. Dla wszystkich skupień oprócz 4-ego wartość p dla AU jest większa lub równa 0,94, zatem odrzucono dla nich hipotezę o braku grupowania się obserwacji. W przypadku 4-ego skupienia p wynosi 0,70, co nie pozwala na odrzucenie tej hipotezy. Wartość p AU obciążona jest błędem próby. Błąd standardowy nigdy nie przekracza jednak 0,2, poza jednym przypadkiem, gdzie wartość p AU nie przekracza 0,4, co pokazano na rysunku 6. Daje to podstawy wnioskować o występowaniu 7 segmentów.

Rysunek 6. Błąd standardowy dla wartości p AU dla 10000 prób bootstrap




Źródło: opracowanie własne

Na podstawie wartości kryterium informacyjnego i metody Warda zdecydowano o wyborze 7 skupień. Segmentacji dokonano przy pomocy modelu EEV. Tabela 2 pokazuje średnie wartości składowych dla skupień. Pogrubiono składowe o największym udziale w tworzeniu danej grupy, a kursywą oznaczono te o udziale najniższym. Pierwsza grupa kształtowana jest przez respondentów z sektora finansowego z Ameryki Północnej lub Azji o raczej wysokich przychodach. Trudno jednoznacznie określić ich doświadczenia we wdrażaniu nowych technologii. Najliczniejszy jest segment drugi. Silny dodatni wpływ składowej 2 oraz silny negatywny wpływ składowej 10 wskazują na kształtowanie

tej grupy przez respondentów z branży finansowej, głównie z Azji, rzadko z branży telekomunikacyjnej oraz przemysłu energetycznego. Grupa ta charakteryzuje sią bardzo dobrym doświadczeniem we wdrażaniu nowych technologii. Segment trzeci zdominowany jest przez branżę telekomunikacyjną, gdzie obserwujemy zainteresowanie data science, występują też przedstawiciele sektora opieki zdrowotnej i energetyki. Na kształtowanie czwartej grupy największy wpływ ma składowa 9, gdzie dominują respondenci z sektora usługowego, mający dobre doświadczenia we współpracy z dostawcą oraz odnotowujący wysokie przychody. W grupie piątej obserwujemy silny wpływ składowej 8, co wskazywałoby na przewagę respondentów reprezentujących służbę zdrowia oraz sektor energetyczny. Jednak silny negatywny wpływ składowej 7 wyklucza obecność tych pierwszych. Przedstawiciele tej grupy mają raczej negatywne doświadczenia we wdrażaniu big data. Segment szósty jest najmniej liczny. Stanowią go w większości przedstawiciele opieki zdrowotnej o nienajlepszych doświadczeniach we wdrażaniu technologii, rzadko pochodzący z Ameryki Północnej. Grupa siódma skupia przedstawicieli branży sprzedażowej o umiarkowanie dobrych doświadczeniach i wsparciu kierownictwa. Dominujący wpływ na tworzenie segmentów mają zatem region i branża działania, a także wcześniejsze doświadczenia.

Tabela 2. Centra skupień

| Grupa | Liczność | SG1 | SG2 | SG3 | SG4 | SG5 | SG6 | SG7 | SG8 | SG9 | SG10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 128 | -0,053 | **1,674** | *-1,855* | 0,468 | 0,009 | -0,630 | 0,406 | 0,062 | -0,157 | 0,265 |
| 2 | 205 | **0,893** | 0,672 | 0,152 | 0,070 | -0,519 | -0,157 | -0,406 | -0,004 | -0,240 | *-1,034* |
| 3 | 170 | -0,078 | -0,415 | -0,115 | *-0,660* | -0,487 | **2,088** | 0,583 | -0,034 | 0,045 | 0,501 |
| 4 | 120 | -0,456 | -0,803 | 0,758 | 0,247 | *-1,472* | -1,153 | 0,677 | -0,244 | **1,666** | 0,232 |
| 5 | 130 | -0,936 | -0,345 | 0,200 | -0,031 | 0,419 | -0,083 | *-1,939* | **1,577** | 0,070 | 0,618 |
| 6 | 78 | -0,542 | -0,429 | 0,397 | 0,222 | 0,895 | -0,466 | **1,958** | 1,397 | *-1,275* | -0,832 |
| 7 | 169 | 0,330 | -0,631 | 0,462 | -0,029 | **1,423** | -0,336 | -0,294 | *-1,693* | -0,284 | 0,294 |

Źródło: opracowanie własne

## PODSUMOWANIE

Analiza skupień pozwala na wyznaczenie segmentów o podobnej charakterystyce oraz na dopasowanie strategii marketingowych przez dostawców nowoczesnych technologii. Do głównych czynników pogłębiających lub ograniczających zainteresowanie tego typu technologiami należą branża, region działania oraz wcześniejsze doświadczenia. Na podstawie segmentacji można wyodrębnić 3 główne grupy o zdecydowanie pozytywnych (skupienie 2) i zdecydowanie negatywnych doświadczeniach (skupienie 5 i 6). Ograniczeniem modelu jest korzystanie z badań wtórnych i opieranie się na pytaniach sformułowanych w ankiecie. Ponadto rezultatu nie można uogólnić ze względu na

dobór celowy respondentów. Opracowanie przedstawia możliwe kroki postępowania w tego typu badaniu oraz prezentuje metody sprawdzenia jakości segmentacji.

## BIBLIOGRAFIA

Biecek P., Szczurek E., Vingron M., Tiuryn J. (2011) The R Package bgmm: Mixture Modeling with Uncertain Knowledge. Journal of Statistical Software, 47 (3).

Banfield J. D., Raftery A. E. (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803-821.

Churchill G.A. (2002) Badania marketingowe. Podstawy metodologiczne, PWN, Warszawa, 372-379.

Dang U., Punzo A., McNicholas P., Ingrassia S., Browne R. (2017) Multivariate Response and Parsimony for Gaussian Cluster-Weighted Models. Journal of Classification, 34(1), 4-34.

Eurostat (2017) 1 in 10 EU businesses analyses big data. http://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/EDN-20170516-1.

Frankfort-Nachmias C., Nachmias D. (2001) Metody badawcze w naukach społecznych, Wydawnictwo Zysk i S-ka, Poznań, 200-205.

Gatnar M., Walesiak E. (2004) Metody statystycznej analizy wielowymiarowej w badaniach marketingowych. AE, Wrocław.

Grabiński T, Wydymus S., Zeliaś A. (1989) Metody taksonomii numerycznej w badaniu zjawisk społeczno-gospodarczych. PWN, Warszawa.

Kutera M., Lasek M. (2010). Zastosowanie metod analizy skupień w przeprowadzaniu segmentacji klientów na potrzeby kampanii reklamowych. Współczesna Ekonomia, 3(15).

Manyika J. at al. (2011) Big Data: The Next Frontier For Innovation, Competition, And Productivity. McKinsey & Company [dostęp 2017-01-25].

Migdał-Najman K., Najman K. (2013) Analiza porównawcza wybranych metod analizy skupień w grupowaniu jednostek o złożonej strukturze grupowej. Zarządzanie i finanse, 11(3) cz. 2, 179-194.

Pakiet mclust programu R: https://cran.r-project.org/web/packages/mclust/mclust.pdf

Pakiet pvclust program R: https://cran.r-project.org/web/packages/pvclust/pvclust.pdf

Scrucca L. (2016) Identifying Connected Components in Gaussian Finite Mixture Models for Clustering. Computational Statistics & Data Analysis, 93, 5-17.

## SURVEY ANALYSIS ON ENTERPRENEURS' PREFERENCES TOWARDS ADVANCED DATA ANALYSIS METHODS

**Abstract:** Enterpreneurs' needs in terms of advanced data analysis methods vary depending on the business sector, funding flexibility, competitors' behavior, volume and volatility of stored information. Business intelligence, visualisation or statistical methods become essential for performing daily

operations in some cases, while in the others they develop into a mean of increasing efficiency or gaining competitive advantage. This publication analyses the differences in enterprises' attitude towards application of hot technologies. An attempt is made to distinguish certain features that potentially make a particular group prone to use offered solutions. This objective is accomplished with a cluster analysis carried out to determine client segments sharing similar characteristics. The results indicate that main differences arise from demographic features, varied expectations and past experiences.

**Keywords:** cluster analysis, Gaussian mixture models