PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 86 2009
Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania

**Ewa Witek**
Akademia Ekonomiczna w Katowicach

# MODELE MIESZANEK W TAKSONOMII I REGRESJI – GRAFICZNA PREZENTACJA WYNIKÓW

## 1. Wstęp

Modele mieszanek rozkładów (*mixture models*) należą do podstawowych typów modeli wykorzystywanych w podejściu opartym na modelach w analizie skupień. Główna idea podejścia modelowego w taksonomii opiera się na spostrzeżeniu, że obserwacje reprezentujące badane zjawisko są niejednorodne i generowane przez rozkłady o różnych parametrach [Jajuga 2007]. Najbardziej popularne w literaturze przedmiotu są mieszanki rozkładów normalnych (np. [Fraley, Raftery 2002; Witek 2009]).

W przypadku gdy zmienne obserwowane są dyskretne, a rozkłady obserwacji np. dychotomiczne lub wielomianowe, wówczas mówimy o mieszankach rozkładów Poissona (*mixtures of Poisson regressions*) lub mieszankach rozkładów dwumianowych, nazywanych także modelami klas ukrytych (*latent class models*). Celem analizy skupień opartej na modelach jest rozpoznanie właściwej struktury klas, a także oszacowanie parametrów rozkładów dla każdej z wyodrębnionych klas.

Modele uwzględniające zmienne towarzyszące, które wpływają na przynależność obserwacji do klas, zwane są modelami regresji klas ukrytych (*latent class regression models*) [Bąk 2009]. Mieszanki rozkładów warunkowych, w których wyróżnia się wpływ zmiennych objaśniających na zmienną objaśnianą i w których możliwe jest uwzględnienie zmiennych towarzyszących (wpływających na przynależność obiektów do klas), zwane są mieszankami modeli GLM (*Generalized Linear Mixture Models*) [Leish 2004]. Mieszanki rozkładów wykorzystywane w regresji mają na celu wyodrębnienie kilku podzbiorów oraz oszacowanie parametrów modelu dla każdej z nich.

Ważną zaletą podejścia modelowego w analizie skupień oraz w regresji jest możliwość uwzględnienia w opisie klasyfikowanych obiektów zmiennych mierzonych w dowolnych skalach (metrycznych i niemetrycznych) [Bąk 2009, s. 411]. W artykule przedstawiono podstawowe charakterystyki podejścia modelowego w taksonomii i regresji. Skoncentrowano się jednak głównie na zmiennych mierzo-

nych na słabych skalach pomiaru (mieszanki dla tego typu zmiennych wykorzystywane są w badaniach marketingowych). Zaprezentowano przykład wykorzystania mieszanek modeli dwumianowych oraz mieszanek funkcji regresji Poissona w badaniach marketingowych, a wyniki badań przedstawiono graficznie.

## 2. Modele mieszanek w taksonomii

Podejście modelowe w taksonomii, oparte na modelu prawdopodobieństwa, jakim jest model mieszanek, ma następujące cechy:
- obserwacje są klasyfikowane na podstawie modelu (przyjmuje się założenia o rozkładach obserwacji, a więc są to metody parametryczne),
- podstawą klasyfikacji są oszacowane prawdopodobieństwa przynależności obserwacji (obiektów) do klas, a każda klasa jest opisywana przez inny model,
- oprócz zmiennych opisujących obserwacje (obiekty) można uwzględnić w modelu zmienne towarzyszące,
- zmienne występujące w modelu mogą być ciągłe lub dyskretne [Bąk 2009].

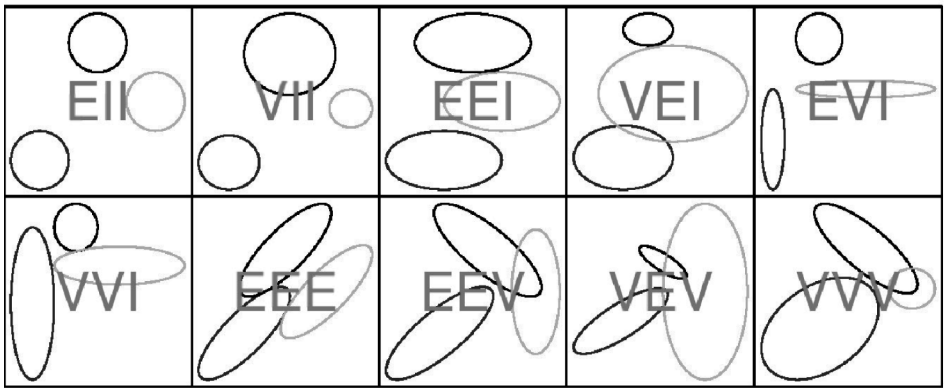
Funkcję gęstości modelu mieszanki wykorzystywaną w taksonomii można zapisać jako:

$$f(y|\boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s f_s(y|\boldsymbol{\Theta}_s), \tag{1}$$

gdzie: $f_s$ – funkcja gęstości klasy $P_s$ ($s$-tego rozkładu składowego mieszanki),
y – realizacja zmiennych obserwowanych,
$\boldsymbol{\Theta}_s$ – wektor parametrów rozkładu składowego $P_s$,
$\boldsymbol{\Theta}$ – wektor parametrów mieszanki rozkładów, $\boldsymbol{\Theta} = (\tau_s, \alpha_s, \boldsymbol{\Theta}_s)$,
$\tau_s$ – prawdopodobieństwo *a priori* – wartość prawdopodobieństwa, że dana obserwacja należy do podpopulacji $P_s$.

W przypadku gdy obiekty opisane są za pomocą zmiennych ciągłych, najczęściej stosowane są mieszanki rozkładów normalnych. Ponieważ każdy z rozkładów składowych opisuje klasę, a funkcje gęstości rozkładów normalnych zależą od dwóch parametrów (wektora wartości przeciętnych $\boldsymbol{\mu}_s$ oraz macierzy kowariancji $\boldsymbol{\Sigma}_s$), mówi się tu o graficznej interpretacji klas (macierz kowariancji ulega dekompozycji [Banfield, Raftery 1993]). Klasy mają stałe bądź różne kształty, wielkości i orientacje [Witek 2009]. Na rysunku 1 przedstawiono cechy geometryczne modeli mieszanek dostępnych w pakiecie `mclust`.

Każdy z modeli pakietu `mclust` opisany jest przez 3 litery. Pierwsza z nich odnosi się do wielkości, druga do kształtu, a trzecia do orientacji klas. Litera E (*equal*) oznacza, że dana cecha geometryczna jest taka sama dla wszystkich klas modelu, a litera V (*vary*), że klasy różnią się ze względu na tę cechę. Z kolei litera I (*identity*) oznacza kształt sferyczny bądź też brak orientacji lub orientację względem osi układu współrzędnych. W przypadku modeli sferycznych (EII, VII) mówimy o braku




Rys. 1. Modele mieszanek dostępne w pakiecie `mclust`

Źródło: opracowanie własne.

orientacji klas. Klasy modeli diagonalnych (EEI, VEI, EVI, VVI) cechuje ta sama lub różna orientacja względem osi układu współrzędnych. Klasy modeli elipsoidalnych (EEE, EEV, VEV, VVV) przyjmują tę samą lub różną orientację względem siebie. Klasy modeli sferycznych mają kształt sferyczny, a klasy modeli elipsoidalnych kształt elipsoidalny.

Mieszanki rozkładów dwumianowych czy rozkładów Poissona nie mają cech geometrycznych (rozkłady składowe zależą tylko od jednego parametru). Mieszanki tych rozkładów najczęściej znajdują zastosowanie w badaniach marketingowych. Przykład zastosowania mieszanek rozkładów dwumianowych do badania preferencji różnych gatunków whiskey zostanie przedstawiony w końcowej części pracy (przykład 1).

## 3. Modele mieszanek w regresji

Zakłada się, że każda składowa modelu mieszanek jest charakteryzowana przez warunkowy rozkład prawdopodobieństwa (dla danych wartości zmiennej **x**), a związek pomiędzy zmienną zależną i zmiennymi objaśniającymi jest określony za pomocą złożonego modelu, jakim jest uogólniony model liniowy (GLM). Liczba rozkładów składowych mieszanki (*components of mixture model*) może być rozumiana m.in. jako liczba segmentów rynku, w których prawdopodobieństwa zakupu w zależności od ceny czy reklamy są wyraźnie różne dla klientów każdego z wyodrębnionych segmentów. Liczba segmentów i parametry modeli GLM wyznaczane są równocześnie. Zależnie od tego, na jakiej skali mierzona jest zmienna objaśniana, wśród modeli wyróżnić można mieszanki modeli logitowych oraz mieszanki regresji Poissona. Ponieważ mieszanki rozkładów Poisssona zostaną wykorzystane w przykładzie empirycznym, poniżej przedstawiono ogólną postać modelu mieszanki regresji tych rozkładów:

$$f(y_i|\mathbf{x}_i,\boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s Po(y_i|\lambda_{is}), \tag{2}$$

$$Po(y_i|\lambda_{is}) = \frac{\lambda_{is}^{y_i}}{y_i!}\exp(-\lambda_{is}), \tag{3}$$

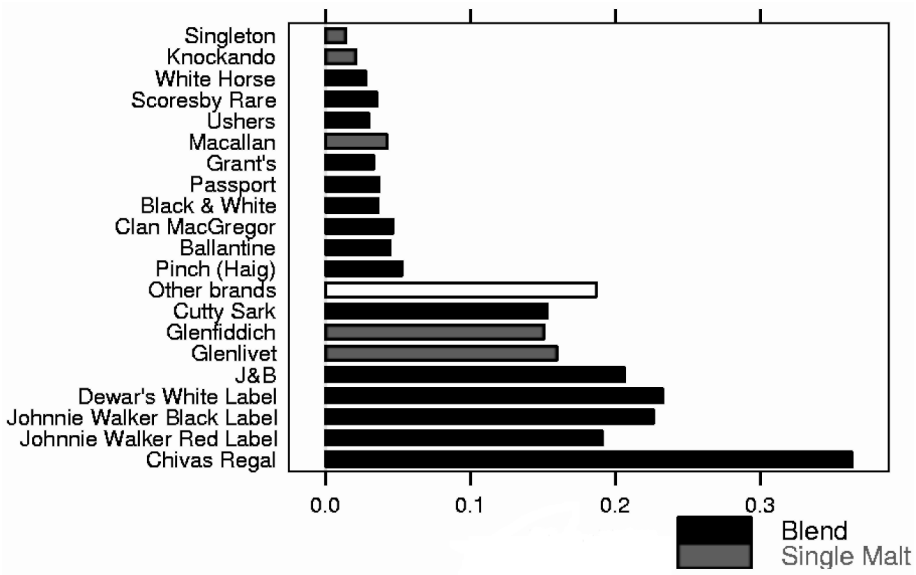
$$\lambda_{is} = \exp(\boldsymbol{\beta}'\mathbf{x}_i), \tag{4}$$

gdzie: $Po(y_i|\lambda_{is})$ – funkcja gęstości *s*-tego rozkładu Poissona,
y – zmienna zależna,
$(\boldsymbol{\beta}')\mathbf{x}_i$ – wektor zmiennych objaśniających i jej parametrów ($\boldsymbol{\beta}'$),

Parametry modeli mieszanek szacowane są najczęściej za pomocą algorytmu EM [Dempster, Laird, Rubin 1977], wybór modelu optymalnego zaś dokonywany jest na podstawie kryteriów informacyjnych BIC, AIC i ICL [Frühwirth-Schnatter 2006].

## 4. Przykład I

W badaniu wykorzystano zbiór danych zgromadzonych w trakcie badań rynku przeprowadzonych przez Simmonsa. Zbiór ten został szczegółowo opisany w pracy Edwardsa i Allenby'ego [2003]. Badaniem objęte zostały gospodarstwa domowe, których członkowie proszeni byli o wskazanie tych gatunków whiskey, które kupo-

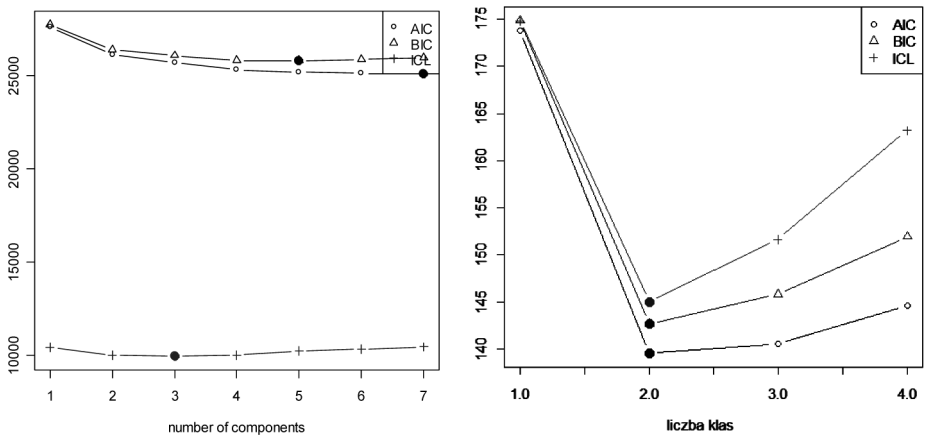



Rys. 2. Graficzna prezentacja zbioru `whiskey`

Źródło: opracowanie własne.

wali w ciągu ostatniego roku. Zbiór dostępny jest także w pakiecie `flexmix` programu **R**[1], gdzie przedstawiony został w postaci macierzy danych binarnych (respondenci przypisywali wartość 0 lub 1 dla 21 gatunków whiskey). Graficzną ilustrację zbioru przedstawiono na rys. 2.

Podziału gospodarstw domowych dokonano za pomocą analizy skupień opartej na mieszankach rozkładów dwumianowych. Wyboru optymalnej liczby klas modelu mieszanek rozkładów dwumianowych dokonano na podstawie kryteriów AIC, BIC oraz ICL (rys. 3a)).




a) dla zbioru `whiskey` (przykład 1)

b) dla zbioru „wizyty na stronie internetowej" (przykład 2)

Rys. 3. Graficzna ilustracja kryteriów informacyjnych BIC, AIC, ICL

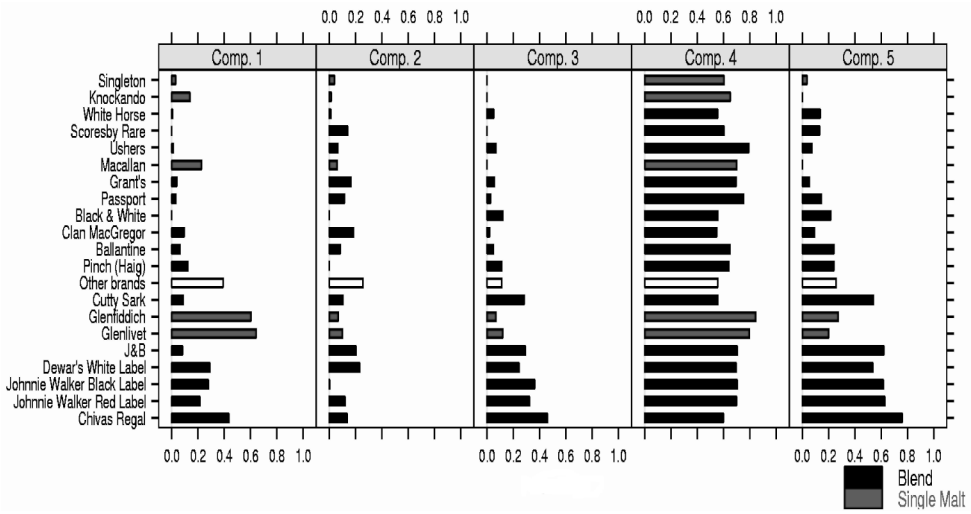
Źródło: opracowanie własne.

Na rysunku 3 a) widać, że kryteria informacyjne nie zawsze dają wyniki jednoznaczne (wartość minimalna dla AIC = 7, dla BIC = 5, dla ICL = 3). Kryterium BIC w analizie skupień opartej na modelach mieszanek dało bardzo dobre wyniki m.in. w pracach [Fraley, Raftery 2002; Stanford, Raftery 2000], dlatego w dalszej analizie za optymalny przyjęto wynik kryterium BIC (dla liczby klas równej pięć).

Liczba obserwacji w pięciu klasach wynosi odpowiednio: 283, 791, 953, 25, 166, a oszacowane prawdopodobieństwa przynależności obiektów do poszczególnych klas to: 0,142; 0,330; 0,431; 0,011; 0,085.

W gospodarstwach domowych klasy pierwszej (14,25%) najczęściej kupuje się gatunki typu whiskey słodowej (*Single-Malt*). Klasę drugą (33%) tworzą gospodar-

[1] Program **R** oraz pakiet `mclust` są bezpłatnie dostępne w Internecie na stronie: http://www.r-project.org/.




Rys. 4. Wynik podziału zbioru `whiskey` na 5 klas

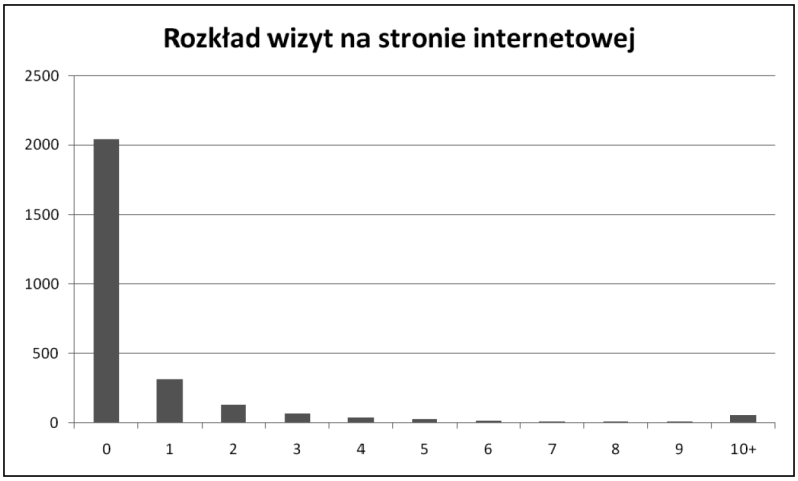
Źródło: opracowanie własne.

stwa, w których najczęściej kupuje się te marki whiskey, które nie zostały objęte badaniem, prawie wcale zaś nie kupuje się gatunku Johnnie Walker Black Label. W gospodarstwach klasy piątej (8,5%) kupowane są różnorodne gatunki whiskey, ale unika się zakupu gatunków typu *Single-Malt*. Klienci klasy trzeciej – największej (43,1%), kupują te same gatunki whiskey co klienci klasy piątej, tyle tylko, że w mniejszych ilościach. Klasę czwartą (1,1%) tworzą gospodarstwa, w których nabywa się największą liczbę różnych gatunków whiskey, a liczba zakupionych butelek whiskey dla różnych gatunków jest bardzo podobna.

## 5. Przykład II

W odróżnieniu od przykładu pierwszego analizowany zbiór obserwacji składa się z zmiennych zależnych oraz objaśniających. Zbiór „wizyty na stronie internetowej" zawiera 2728 obserwacji (2728 osób, które wzięły udział w badaniu). Badano częstość ponownych wizyt na stronie internetowej odzieżowego sklepu internetowego Khaki Chinos, Inc. w ciągu II połowy 2000 r. Zarząd firmy chciał sprawdzić, czy na częstość ponownych wizyt na stronie www.khakichinos.com ($y$) miały wpływ cechy, takie jak: $x_1$– dochód; $x_2$ – płeć; $x_3$ – wiek; $x_4$ – liczba osób w gospodarstwie domowym. Częstość wizyt na stronie internetowej przedstawiono na rys. 5.

Za pomocą kryteriów informacyjnych dokonano wyboru optymalnej liczby rozkładów składowych. Wartości kryteriów przedstawione zostały na wykresie 3b). Tym razem uzyskano jednoznaczne wyniki – wszystkie kryteria wskazały optymal-




Rys. 5. Rozkład częstości wizyt na stronie internetowej

Źródło: opracowanie własne.

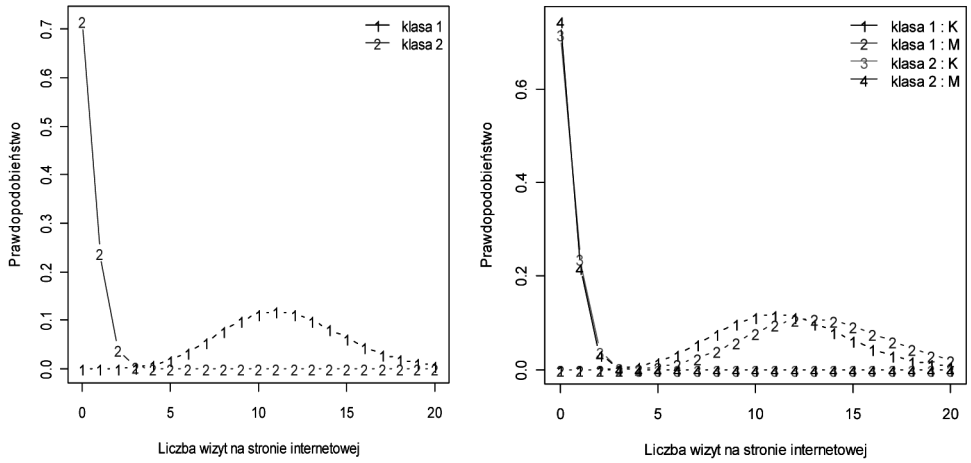
ną liczbę segmentów równą dwa. Dla każdego z nich oszacowano funkcję regresji Poissona za pomocą funkcji `stepFlexmix()` pakietu `flexmix` programu **R**. Następnie zbadano istotność parametrów za pomocą testu *t*-Studenta. Zmienna $x_4$ okazała się nieistotna. Dlatego parametry mieszanki funkcji regresji Poissona oszacowane zostały po raz kolejny. Wyniki przedstawiono w tab. 1.

Tabela 1. Wyniki podziału na 2 klasy, $\tau_s$ – prawdopodobieństwo przynależności do klasy $P_s$

| Klasa | $\tau_s$ | Liczebność | Model GLM |
|---|---|---|---|
| I | 0,05 | 150 | $\lambda_1 = \exp(-0{,}29x_1 - 0{,}25x_2 + 0{,}14x_3 + 6{,}61)$ |
| II | 0,95 | 2578 | $\lambda_2 = \exp(-0{,}21x_1 - 0{,}68x_3 - 4{,}78)$ |

Źródło: opracowanie własne.

Największy (negatywny) wpływ na liczbę odwiedzających stronę internetową w klasie pierwszej ma dochód ($x_1$) oraz płeć. Pozytywny wpływ na liczbę odwiedzających stronę ma wiek. W klasie drugiej w miarę wzrostu dochodu wzrasta liczba odwiedzających stronę. Płeć nie wpływa istotnie na przeciętną liczbę odwiedzających www.khakichinos.com, największy zaś wpływ na liczbę odwiedzających ma wiek. Prawie wszyscy klienci zostali zaklasyfikowani do klasy II . Są to osoby, które z wysokim prawdopodobieństwem (równym 0,7) nie odwiedzą strony www.khakichinos.com po raz kolejny. Istnieje niskie prawdopodobieństwo, że liczba ponownych wizyt wyniesie 10. To prawdopodobieństwo wynosi zaledwie 0,1 wśród nielicznej 150-osobowej klienteli grupy II. Wyniki badań wskazują na dość pesymistyczną




a) Wykres z podziałem tylko na klasy

b) wykres z podziałem na klasy z uwzględnieniem płci

Rys. 6. Wykres prawdopodobieństw przynależności klientów do klas I i II

Źródło: opracowanie własne.

prognozę, jednakże należy zauważyć, że w drugiej połowie 2000 r. najczęstsza liczba ponownych wizyt wynosiła 0 (rys. 5). Z rysunku 6b) można odczytać, że dla klasy II zmienna płeć jest zmienną nieistotną.

## 6. Podsumowanie

W artykule przedstawiono przykład wykorzystania modeli mieszanek w badaniach marketingowych. Podejście modelowe pozwoliło na segmentację respondentów na podstawie postaw i preferencji wyrażonych w badaniu. Dla danych z przykładu 1 (dotyczących wykorzystania modeli mieszanek w taksonomii) wyodrębniono segmenty o podobnych wzorcach zachowań dla klientów kupujących gatunki whiskey. Dla danych z przykładu 2 (dotyczących wykorzystania modeli mieszanek w regresji) dokonano wyboru optymalnej liczby grup klientów, oszacowano parametry funkcji regresji Poissona dla każdej z nich oraz dokonano oceny wpływu zmiennych demograficznych na częstość wizyt na stronie odzieżowego sklepu internetowego.

## Literatura

Banfield J.D., Raftery A.E., *Model-Based Gaussian and Non-Gaussian Clustering*, „Biometrics" 1993 nr 49, s. 803-821.

Bąk A., *Podejście modelowe w analizie skupień – zastosowania na przykładach danych symulacyjnych*, [w:] *Współczesne problemy modelowania i prognozowania zjawisk społeczno-gospodarczych*, J. Pociecha (red.), UE, Kraków 2009, s. 411-421.

DeSarbo W.S., Cron W.L., *A Maximum Likelihood Methodology for Clusterwise Linear Regression*, „Journal of Classification" 1988 nr 5, s. 249-282.
Dempster A.P., Laird N.M., Rubin D.B., *Maxiumum Likelihood for Incomplete Data Via The EM Algorithm (With Discussion)*, „Journal of the Royal Statistical Society" 1977 Ser.B, nr 39, s. 1-38.
Edwards Y., Allenby G., *Multivariate Analysis of Multiple Response Data*, "JMR" 2003 nr 30, s. 321--334.
Fraley C., Raftery A.E., *Model-based Clustering, Discriminant Analysis, and Density Estimation*, „Journal of the American Statistical Association" 2002 nr 97, s. 611-631.
Frühwirth-Schnatter S., *Finite Mixture and Markov Switching Models*, Springer, 2006, s. 116-118.
Jajuga K., *From Multivariate Distribiton to Data Analysis – Model Based Clustering* [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia nr 14, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław 2007.
McLachlan G.J., Peel D., *Finite Mixture Models*, Wiley, New York 2000.
Leisch R., *Flexmix: A General Framework for Finite Mixture Models and Latent Class Regression in R*, „Journal of Statistical Software" 2004 nr 11(8), s. 1-18, http://www,jstatsoft,org/v11/i08/.
Stanford D.C., Raftery A.E., *Principal Curve Clustering With Noise*, „IEEE Transactions on Pattern Analysis and Machine Analysis" 2000 nr 22, s. 601-609.
Wang P., Cockburn I.M., Puterman M.L., *Analysis of Patent Data – A Mixed-Poisson-Regression-Model Approach*, „Journal of Business & Economic Statistics" 1998 nr 16 (1), s. 27-41.
Wedel M., *Concomitant Variables in Finite Mixture Models*, „StatisticaNeerlandica" 2002 nr 55, s. 362--375.
Witek E., *Analiza skupień – podejście modelowe*, [w:] *Statystyczna analiza danych z wykorzystaniem programu R*, M. Walesiak, E. Gatnar (red.), PWN, Warszawa 2009, s. 434-462.

## MIXTURE MODELS IN CLUSTERING AND REGRESSION – GRAPHICAL RESULTS

### Summary

The paper focuses on finite mixture models and their application in clustering and regression. These models are often used to capture overdispersion in the data which can occur for example if important covariates are omitted in the regression. It is then assumed that the influence of these covariates can be captured by allowing a random distribution for the intercept. The author presents the graphical results of research.