PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Bąk
Akademia Ekonomiczna we Wrocławiu

MODELE ZE ZMIENNYMI UKRYTYMI W SEGMENTACJI RYNKU

1. Wstęp

Poszukuje się nieustannie, także w naukach ekonomicznych, modeli jak najlepiej opisujących rzeczywistość. Modele ze zmiennymi ukrytymi pozwalają lepiej, w porównaniu z modelami klasycznymi, opisać populacje (próby) niejednorodne, a więc zidentyfikować klasy, grupy, segmenty. Segmentacja w badaniach marketingowych oznacza podział rynku na względnie jednorodne klasy konsumentów na podstawie podobieństwa kryteriów charakteryzujących konsumentów lub kryteriów ich reakcji na oferowany produkt lub usługę.

Zmienne ukryte opisują obiekty (zjawiska lub procesy) tak jak zmienne obserwowalne. Nie można jednak bezpośrednio zmierzyć wartości tych zmiennych. Uwzględnienie w modelu takich zmiennych może jednak przynieść wymierne korzyści. Realizacje tych zmiennych są kształtowane przez zmienne, które można zaobserwować i zmierzyć.

Modele ze zmiennymi ukrytymi uwzględniają heterogeniczność konsumentów na poziomie grupowym i dlatego mogą być one wykorzystywane w segmentacji rynku. Estymacja modelu polega m.in. na oszacowaniu liczby i wielkości poszczególnych segmentów.

Celami artykułu są prezentacja podstawowych cech modeli ze zmiennymi ukrytymi, pokazanie korzyści płynących z ich stosowania w segmentacji rynku na gruncie badań marketingowych oraz wskazanie problemów związanych z ich praktycznym wykorzystaniem.

2. Segmentacja

Segmentacja rynku stała się ważnym obszarem badań marketingowych za sprawą pionierskiego artykułu Smitha [1956]. Autor ten pokazał wymierne korzyści ekonomiczne płynące z identyfikacji homogenicznych grup konsumentów wśród heterogenicznych z natury rzeczy uczestników rynku (zob. [Walesiak, Bąk 1999; Bąk 2004]). Segmentacja jest elementem marketingu strategicznego w łańcuchu: segmentacja rynku → zdefiniowanie rynku docelowego → pozycjonowanie rynku (zob. [Kotler 1999, s. 243]).

Segmentacja rynku w ujęciu klasycznym to podział rynku na względnie jednorodne klasy konsumentów na podstawie podobieństwa kryteriów charakteryzujących konsumentów lub kryteriów ich reakcji na oferowany produkt lub usługę (por. np. [Kotler 1999, s. 243; Pociecha 1996, s. 55]). Segmentację rynku prowadzi się na podstawie zbioru wyróżnionych czynników charakterystycznych (kryteriów, zmiennych), które umożliwiają wyodrębnienie jednorodnych segmentów konsumentów. Do czynników tych zalicza się kryteria geograficzne, demograficzne i kulturowe, społeczno-ekonomiczne, psychologiczne, behawioralne i inne (zob. m.in. [Kotler 1999, s. 247-255; Beane, Ennis 1987]).

W tabeli 1 zamieszczono wybrane kryteria segmentacji z uwzględnieniem podziału na zmienne dotyczące konsumenta i zmienne związane z produktem oraz zmienne obserwowalne i nieobserwowalne.

Tabela 1. Zmienne segmentacji na rynku dóbr i usług konsumpcyjnych

Zmienne	Związane z konsumentem	Związane z produktem
Obserwowalne	– **geograficzne** (zasięg terytorialny, wielkość regionu, wielkość miasta, gęstość zaludnienia, klimat) – **demograficzne i kulturowe** (wiek, płeć, wielkość rodziny, faza rozwoju rodziny, wyznanie, narodowość, rasa) – **socjoekonomiczne** (dochód, wykształcenie, zawód, klasa społeczna)	– **użytkowanie** (intensywność użytkowania, status użytkownika) – **zakup** (lojalność, częstotliwość, wielkość, forma, miejsce i czas zakupu)
Nieobserwowalne	– **psychologiczne** (styl życia, osobowość, systemy wartości)	– oczekiwane korzyści, opinie, intencje, postawy, preferencje

Źródło: opracowano na podstawie: [Wedel, Kamakura 1998, s. 7; Kotler 1999, s. 249; Beane, Ennis 1987; Walesiak, Bąk 1999; Walesiak 2000].

Metody segmentacji klasyfikuje się ze względu na podejście stosowane w celu wyodrębnienia segmentów oraz wykorzystywane metody statystyczno-ekonometryczne (zob. tab. 2).

Tabela 2. Klasyfikacja metod segmentacji

Metody segmentacji	*A priori*	*Post hoc*
Deskryptywne	– tabele kontyngencji – modele log-liniowe	– metody klasyfikacji: • rozłącznej (hierarchiczne aglomeracyjne i deglomeracyjne, obszarowe i gęstościowe, optymalizacji iteracyjnej) • nierozłącznej • rozmytej
Predyktywne	– tabele krzyżowe – analiza regresji – analiza logitowa – analiza dyskryminacyjna	– drzewa klasyfikacyjne – sieci neuronowe – regresja skupieniowa – modele ze zmiennymi ukrytymi

Źródło: opracowano na podstawie: [Wedel, Kamakura 1998, s. 18; Walesiak, Bąk 1999; Walesiak 2000].

Wśród metod segmentacji wyodrębnianych ze względu na stosowane podejście w celu identyfikacji segmentów wyróżnia się:

1. Metody segmentacji *a priori*, w których:
   - ustala się „z góry" zarówno liczbę segmentów, jak i ich charakterystyki,
   - podstawę segmentacji stanowią na ogół zmienne związane z konsumentem.
2. Metody segmentacji *post hoc*, w których:
   - segmenty wyodrębnia się za pomocą statystycznych metod klasyfikacji danych,
   - podstawę segmentacji stanowią zmienne obserwowalne i nieobserwowalne.
3. Metody segmentacji hybrydowej (dwuetapowej), w których:
   - etap I to segmentacja *a priori*,
   - etap II to segmentacja *post hoc*.

Metody segmentacji ze względu na stosowane metody statystyczno-ekonometryczne podzielić można na dwie grupy:

1. Metody opisowe (deskryptywne) – w zbiorze zmiennych nie wyodrębnia się zmiennej objaśnianej (segmenty identyfikuje się np. za pomocą metod klasyfikacji).

2. Metody predyktywne – w zbiorze zmiennych wyodrębnia się zmienną objaśnianą i zmienne objaśniające (segmenty identyfikuje się np. za pomocą metod regresji).

3. Modele ze zmiennymi ukrytymi

Założenie o jednorodności obserwacji, przyjmowane w klasycznych modelach regresji, może prowadzić do „dopasowania" jednego modelu do danych, podczas gdy właściwym rozwiązaniem jest „dopasowanie" dwóch lub więcej modeli do istniejących w próbie (populacji) segmentów. Modele ze zmiennymi ukrytymi uwzględniają heterogeniczność obserwacji i umożliwiają identyfikację w badanej próbie grup (segmentów) względnie jednorodnych.

Zakłada się, że w badanej próbie istnieje skończona liczba grup obserwacji o podobnych właściwościach (np. konsumentów o podobnych preferencjach). Między grupami natomiast występują istotne różnice. Grupy te nie są znane *a priori* (są „ukryte"), ponieważ nie jest znana ani przynależność poszczególnych obserwacji (konsumentów) do określonych segmentów, ani liczba grup. Estymacja modelu ze zmiennymi ukrytymi polega m.in. na oszacowaniu liczby i wielkości poszczególnych segmentów. Ogólną postać modelu ze zmiennymi ukrytymi (modelu klas ukrytych) reprezentuje zależność w postaci rozkładów warunkowych (por. [DeSarbo, Wedel 1994; Vriens 2001]):

$$f(\mathbf{y}\,|\,\boldsymbol{\Phi}) = \sum_{c=1}^{C} \pi_c f(\mathbf{y}\,|\,\boldsymbol{\theta}_c), \tag{1}$$

gdzie: $f(\mathbf{y}\,|\,\boldsymbol{\Phi})$ – funkcja rozkładu obserwacji (np. preferencji konsumentów); $\sum_{c=1}^{C} \pi_c$ – rozkład prawdopodobieństw bezwarunkowych wyrażających przynależności do poszczególnych klas ukrytych (reprezentuje w modelu rozkład zmiennej ukrytej); $f(\mathbf{y}\,|\,\boldsymbol{\theta}_c)$ – funkcja opisująca prawdopodobieństwa warunkowe (reprezentuje w modelu rozkład zmiennych obserwowanych lub zmiennej objaśnianej); $\boldsymbol{\Phi} = (\boldsymbol{\pi}, \boldsymbol{\theta})$ – wszystkie nieznane parametry modelu; $\boldsymbol{\theta}_c$ – wektor nieznanych parametrów w c-tej klasie (np. μ_c i σ_c dla rozkładu normalnego).

Na podstawie modelu (1) metodą największej wiarygodności szacuje się parametry π_c i $\boldsymbol{\theta}_c$ w poszczególnych segmentach. Funkcja największej wiarygodności dla próby liczącej S konsumentów określona jest wzorem (postać funkcji jest zależna od rozkładu preferencji):

$$L(\mathbf{y};\boldsymbol{\Phi}) = \prod_{s=1}^{S} f(y_s\,|\,\boldsymbol{\Phi}). \tag{2}$$

Dopasowanie funkcji do danych jest przeprowadzane metodą największej wiarygodności z wykorzystaniem algorytmów optymalizacyjnych (np. Newtona-Raphsona, E-M).

Modele ze zmiennymi ukrytymi charakteryzują się określonymi właściwościami formalnymi, które są istotne z punktu widzenia ich zastosowań w segmentacji rynku, a mianowicie:

- umożliwiają identyfikację segmentów na podstawie zmiennych obserwowanych lub zmiennej objaśnianej,
- zawierają jedną kategorialną zmienną ukrytą (liczba kategorii jest równa liczbie segmentów),
- podstawą klasyfikacji obiektów do segmentów są oszacowane na podstawie modelu prawdopodobieństwa przynależności,
- zmienne zaobserwowane mogą być mierzone na różnych skalach,
- do modelu można włączyć zmienne towarzyszące i zmienne objaśniające.

W modelach ze zmiennymi ukrytymi można wyróżnić następujące typy zmiennych:

- zmienne ukryte (*latent variables*), które mogą być mierzone na skalach nominalnych lub porządkowych; model musi zawierać co najmniej jedną zmienną ukrytą,
- zmienne obserwowane (*manifest variables*) lub objaśniane (*dependent variables*), które mogą być mierzone na różnych skalach; model musi zawierać co najmniej jedną z takich zmiennych,
- zmienne towarzyszące (*concomitant variables*, *covariates*) i zmienne objaśniające (*predictor variables*), które mogą być mierzone na różnych skalach; te zmienne nie muszą w modelu występować.

Między tymi zmiennymi zachodzą następujące podstawowe zależności:

- zmienne objaśniające (atrybuty produktów lub usług, zmienne sztuczne reprezentujące układ czynnikowy) → zmienna objaśniana (np. preferencje konsumentów),
- zmienne towarzyszące (np. charakterystyki konsumentów, cechy demograficzne) → zmienna ukryta (przynależność do segmentu),
- zmienna ukryta → zmienna objaśniana.

Zmienne ukryte opisują te zjawiska, których bezpośrednio nie można zmierzyć. W literaturze przedmiotu podkreśla się jednak, że zmienne ukryte nie wnoszą żadnych dodatkowych informacji ponad te, które są zawarte w zmiennych obserwowanych (mierzonych). Zmienne ukryte są kategorią abstrakcyjną, mogącą służyć do syntezy lub agregacji właściwości (informacji) zawartych w zmiennych obserwowanych (por. [Everitt, Dunn 2001, s. 305]).

W segmentacji znajdują zastosowanie głównie dwa typy modeli ze zmiennymi ukrytymi (zob. [Wedel, Kamakura 1998; Wedel 1999; Vermunt, Magidson 2003]):

1. Model klas ukrytych:
   - zawiera więcej niż jedną zmienną obserwowaną (dychotomiczną, wielokategorialną),
   - nie zawiera zmiennych objaśniających,
   - w modelu mogą (ale nie muszą) występować zmienne towarzyszące.
2. Model regresji klas ukrytych:
   - zawiera jedną zmienną obserwowaną (zmienną objaśnianą),
   - w modelu występują zmienne objaśniające,
   - w modelu mogą występować zmienne towarzyszące.

Model klas ukrytych w ogólnej postaci przedstawia zależność (1), natomiast wśród modeli regresji klas ukrytych można wyróżnić dwa warianty:

1) modele ze zmiennymi objaśniającymi:

$$f(\mathbf{y} \mid \mathbf{x}) = \sum_{c=1}^{C} \pi_c f(\mathbf{y} \mid \pi_c, \mathbf{x}), \tag{3}$$

gdzie: $\mathbf{y}$ – wektor obserwacji; $\mathbf{x}$ – zmienne objaśniające (atrybuty, predyktory) wpływające na $\mathbf{y}$; π_c – prawdopodobieństwo przynależności do c-tej klasy lub rozmiar segmentu (zmienna ukryta);

2) modele ze zmiennymi objaśniającymi i zmiennymi towarzyszącymi:

$$f(\mathbf{y}|\mathbf{x}) = \sum_{c=1}^{C} (\pi_c|\mathbf{z}) f(\mathbf{y}|\pi_c, \mathbf{x}), \tag{4}$$

gdzie: $\mathbf{z}$ – zmienne towarzyszące wpływające na zmienną ukrytą (przynależność do klas ukrytych).

Ogólne modele regresji ze zmiennymi ukrytymi wymagają, na potrzeby ich oszacowania, parametryzacji polegającej na przyjęciu określonych założeń o rozkładach $f(\mathbf{y}|\pi_c, \mathbf{x})$ i π_c. W tym celu wykorzystuje się koncepcję uogólnionych modeli liniowych (GLM – *generalized linear models*) (zob. np. [Everitt, Dunn 2001; Wedel, Kamakura 1998]).

4. Przykład

W przykładzie wykorzystano model ze zmiennymi ukrytymi (model klas ukrytych) w dwóch wersjach: w pierwszej model zawiera tylko zmienne obserwowane i zmienną ukrytą, natomiast w drugiej model zawiera zmienne obserwowane, zmienną ukrytą i zmienne towarzyszące. Celem badania jest identyfikacja wpływu zmiennych towarzyszących na jakość modelu w sensie zarówno zbieżności, jak i dopasowania modelu do danych.

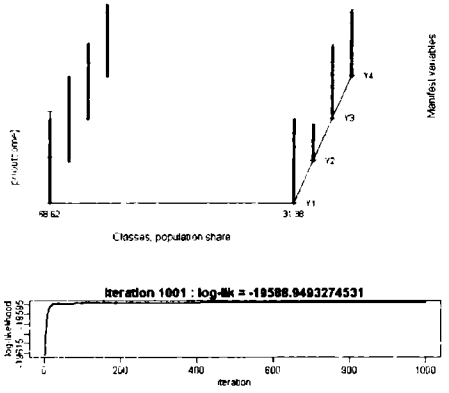
Do estymacji modelu został użyty pakiet poLCA [Linzer, Lewis 2006] pracujący w środowisku programistycznym R (stosowanym w obliczeniach matematycznych, statystycznych i ekonometrycznych). Pakiet poLCA jest przeznaczony do estymacji modeli klas ukrytych, w których mogą występować wielokategorialne zmienne obserwowane oraz zmienne towarzyszące. Parametry modelu są szacowane metodą największej wiarygodności z wykorzystaniem dwóch algorytmów maksymalizacji funkcji wiarygodności: E-M i Newtona-Raphsona.

Charakterystyka badania:

1) dane: próba pseudolosowa licząca `1000` obserwacji (wygenerowana za pomocą funkcji `poLCA.simdata`),
2) zmienne: (a) cztery zmienne obserwowane o wartościach dychotomicznych: `Y1`, `Y2`, `Y3`, `Y4`; (b) cztery zmienne towarzyszące: `X1`, `X2`, `X3`, `X4`,
3) liczba segmentów: `2`,
4) maksymalna liczba iteracji: `1000`,
5) mierniki jakości dopasowania modelu: kryterium informacyjne Akaikego (`AIC`) i kryterium Schwarza (`BIC`).

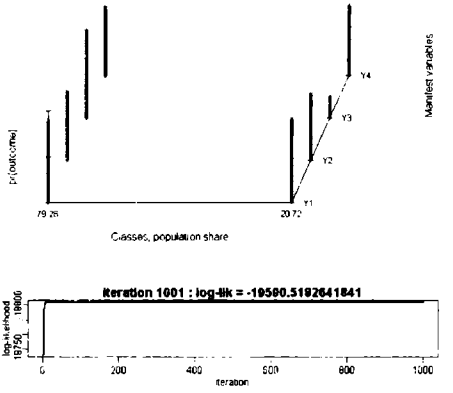
Rysunek 1 przedstawia wyniki estymacji modelu klas ukrytych w wersji pierwszej, tj. bez zmiennych towarzyszących. Przy założonej liczbie 1000 iteracji algorytm maksymalizacji funkcji wiarygodności nie znajduje maksimum. Po dodaniu czterech zmiennych towarzyszących (ich zadaniem jest „lepsze" dopasowanie obserwacji do segmentów) model także nie osiąga progu zbieżności przy tej samej liczbie iteracji (rys. 2).

```
f2x0 <- cbind(Y1,Y2,Y3,Y4)~1
brak zbieżności
```



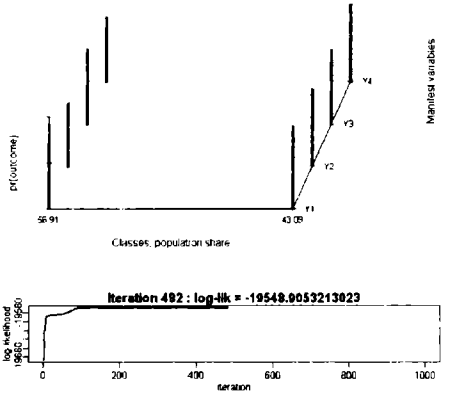
Rys. 1. Model bez zmiennych towarzyszących
Źródło: opracowanie własne.

```
f2x4<-cbind(Y1,Y2,Y3,Y4)~X1+X2+X3+X4
brak zbieżności
```



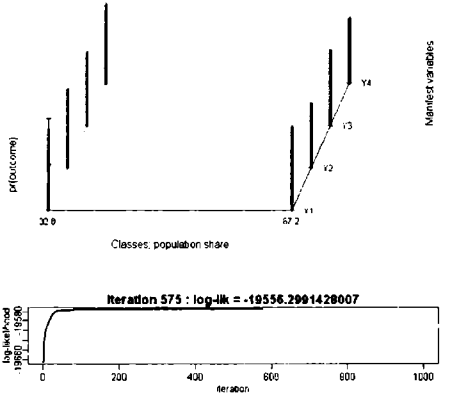
Rys. 2. Model zawierający 4 zmienne towarzyszące
Źródło: opracowanie własne.

```
f2x3<-cbind(Y1,Y2,Y3,Y4)~X1+X2+X3
AIC( 2 ): 39121.81
BIC( 2 ): 39208.33
```



Rys. 3. Model zawierający 3 zmienne towarzyszące
Źródło: opracowanie własne.

```
f2x2c <- cbind(Y1,Y2,Y3,Y4)~X2+X3
AIC( 2 ): 39134.6
BIC( 2 ): 39213.91
```



Rys. 4. Model zawierający 2 zmienne towarzyszące
Źródło: opracowanie własne.

W dwóch kolejnych przypadkach (rys. 3 i 4) włączono do modelu odpowiednio trzy i dwie zmienne towarzyszące. Obydwa modele są zbieżne, przy czym model z trzema zmiennymi towarzyszącymi wykazuje lepsze dopasowanie do danych (większą zawartość informacyjną) w świetle kryteriów AIC i BIC oraz szybciej osiąga zbieżność.

Wyniki badań zilustrowane na rys. 1-4, jak też przebiegi symulacyjne, których rezultaty ze względu na ograniczenia objętościowe artykułu nie zostały zamieszczone, wskazują na to, iż zbieżność modelu i jakość jego dopasowania do danych zależą od liczby zmiennych towarzyszących oraz ich konfiguracji. Pojawia się jednak problem wyboru liczby tych zmiennych oraz ich optymalnej konfiguracji. Zagadnienia te wymagają dalszych badań.

Reasumując, do podstawowych problemów badawczych związanych z wykorzystaniem modeli ze zmiennymi ukrytymi w segmentacji rynku należy zaliczyć:

- wybór zmiennych towarzyszących (wykorzystanie metod doboru zmiennych, analiza zasobu informacyjnego macierzy danych),
- badania porównawcze modeli klas ukrytych z modelami regresji klas ukrytych,
- badania porównawcze z innymi metodami klasyfikacji,
- wybór liczby segmentów i problem dużej liczby segmentów,
- zbieżność iteracyjnych algorytmów optymalizacyjnych.

Literatura

Bąk A. (2004), *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*, Prace Naukowe AE we Wrocławiu nr 1013, Monografie i Opracowania nr 157, AE, Wrocław.

Beane T.T., Ennis D.M. (1987), *Market Segmentation: A Review*, „European Journal of Marketing" 21 (5), s. 20-42.

DeSarbo W.S., Wedel M. (1994), *A Review of Recent Developments in Latent Class Regression Models*, [w:] R.P. Bagozzi (red.), *Advanced Methods of Marketing Research*, s. 352-388. Blackwell, Cambridge.

Everitt B.S., Dunn G. (2001), *Applied Multivariate Data Analysis*, Arnold, London.

Kotler P. (1999), *Marketing. Analiza, planowanie, wdrażanie i kontrola*, Felberg SJA, Warszawa.

Linzer D.A., Lewis J. (2006), *poLCA: Polytomous Variable Latent Class Analysis*, http://dlin zer.bol.ucla.edu/poLCA.

Pociecha J. (1996), *Metody statystyczne w badaniach marketingowych*, PWN, Warszawa.

Smith W.R. (1956), *Product Differentiation and Market Segmentation as Alternative Marketing Strategies*, „Journal of Marketing" 21 (3), s. 3-8.

Vermunt J.K., Magidson J. (2003), *Technical Appendix for Latent GOLD 3.0*, http:// www.statisticalinnovations.com, Belmont.

Vriens M. (2001), *Market Segmentation. Analytical Developments and Application Guidelines*, Millward Brown IntelliQuest.

Walesiak M. (2000), *Segmentacja rynku. kryteria i metody*, [w:] A. Zeliaś (red.), *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, AE, Kraków, s. 191-201.

Walesiak M., Bąk A. (1999), *Segmentacja rynku z wykorzystaniem metody conjoint measurement*, [w:] S. Mynarski (red.), *Zastosowanie metod wielowymiarowych w badaniach segmentacji i selektywności rynku*, materiały z II Warsztatów Metodologicznych, AE, Kraków, s. 83-93.

Wedel M. (1999), *Concomitant Variables in Mixture Models*, http://www.ub.rug.nl/eldoc/som/b/99B24. SOM, Groningen.

Wedel M., Kamakura W.A. (1998), *Market Segmentation. Conceptual and Methodological Foundations*, Kluwer Academic Publishers, Boston-Dordrecht--London.

LATENT VARIABLES MODELS IN MARKET SEGMENTATION

Summary

Latent variable models account for heterogeneity of consumer at the group (segment) level. Therefore they may be used as a tool of market segmentation. Model estimation involves, among others, estimation of the number and size of individual segments.

The paper presents basic feathers of latent variables models and the benefits of their use in market segmentation on the marketing research field. Then it presents the problems connected with those models which should be the subject of further research.