

**Aneta Rybicka, Tomasz Bartłomowicz**

Uniwersytet Ekonomiczny we Wrocławiu

## **ZAGADNIENIE POZIOMU AGREGACJI DANYCH W METODACH WYBORÓW DYSKRETNYCH**

### **1. Preferencje i metody ich badań**

W badaniach preferencji konsumentów wykorzystuje się obserwacje historyczne (wtórne źródła danych) bądź informacje o charakterze antycypacyjnym, wyrażające intencje konsumentów (pierwotne źródła danych). Tym samym, co ma również odzwierciedlenie w podziale preferencji na dwa główne rodzaje, wyróżnia się metody badania preferencji ujawnionych oraz metody badania preferencji wyrażonych (por. [Bąk 2004a, s. 42]).

Przez **preferencje ujawnione** rozumieć należy te zachowania rynkowe nabywców, które stanowią odbicie ich rzeczywistych decyzji rynkowych. Podstawą analizy preferencji ujawnionych jest więc najczęściej materiał statystyczny, zgromadzony w wyniku rejestracji danych o rzeczywistych wyborach rynkowych. Alternatywnym źródłem danych mogą być przeprowadzone *a posteriori* sondaże (bezpośrednie lub pośrednie) dotyczące wyborów rynkowych wcześniej dokonanych przez nabywców. Oznacza to, iż metody badań preferencji ujawnionych (statystyczne, ekonometryczne) są z istoty rzeczy oparte na danych historycznych i wynikają często z ich charakteru (por. [Bąk 2004a, s. 42-43]).

W przypadku **preferencji wyrażonych**, które odpowiadają hipotetycznym (deklarowanym) preferencjom nabywców, metody badań najczęściej są oparte na danych zgromadzonych *a priori* za pośrednictwem sondaży pośrednich lub bezpośrednich, umożliwiających rejestrację wyrażonych intencji nabywców w momencie badań. W pomiarze preferencji wyrażonych zastosowanie znajdują metody reprezentujące podejście kompozycyjne, dekompozycyjne oraz mieszane. Wynika to z charakteru stosowanego podejścia, które ma decydujący wpływ na sposób gromadzenia danych o preferencjach (pomiar preferencji), wybór metod analizy preferencji oraz na postać modelu (por. [Bąk 2004a, s. 43]).

W **podejściu kompozycyjnym** wykorzystywana jest idea modelu postaw Fishbeina<sup>1</sup> oraz założenia, które są związane z modelem wartości oczekiwanej, gdzie użyteczność całkowita wielowymiarowego profilu jest ważoną sumą ocen poziomów zmiennych, a wagi wyrażają ważność poszczególnych zmiennych (por. [Walesiak, Bąk 1997, s. 14; Zwerina 1997, s. 3]). Modele kompozycyjne jest to klasa modeli wielu zmiennych, których przykładem są m.in. modele regresji oraz analiza dyskryminacyjna (por. [Hair i in. 1995, s. 562-563]). Analitycy, stosując modele kompozycyjne, zbierają oceny respondentów na temat wielu cech produktu lub usługi, a następnie wiążą te oceny w całkowite preferencje. Analitycy zatem określają („komponują, składają”) preferencje respondentów z ocen dokonywanych przez respondentów każdego z atrybutów produktu lub usługi.

W **podejściu dekompozycyjnym** w celu przeprowadzenia analizy preferencji konsumentów wykorzystuje się metody *conjoint analysis* oraz metody oparte na wyborach (por. [Bąk 2000, s. 76]). Modele dekompozycyjne to klasa modeli, które „rozkładają” preferencje całkowite konsumentów. Wykorzystując modele dekompozycyjne, prezentuje się respondentom zbiór profili (produktów, obiektów), zazwyczaj w formie hipotetycznych lub rzeczywistych produktów lub usług (por. [Hair i in. 1995, s. 558]). Za pomocą metod statystycznych i algorytmów komputerowych przeprowadza się dekompozycję preferencji całkowitych i oblicza się użyteczności cząstkowe<sup>2</sup> (por. [Bąk 2004a, s. 42]).

Tabela 1. Podobieństwa i różnice metod dekompozycyjnych

Etapy procedury badawczej	Metody <i>conjoint analysis</i>	Metody wyborów dyskretnych
Podstawy teoretyczne	<i>conjoint measurement</i>	teoria użyteczności losowej
Liczba atrybutów	do 10 (zwykle do 6)	6-8
Liczba poziomów	do 15	9-15
Forma ankiety	papierowa, komputerowa	papierowa, komputerowa
Metoda gromadzenia danych	pełnych profili, porównywania parami, porównywania atrybutów	wybór spośród zbiorów profili
Skala pomiaru preferencji	ilorazowa, przedziałowa, porządkowa	nominalna
Rodzaj modelu	liniowy, addytywny	liniowy, addytywny
Metoda estymacji użyteczności cząstkowych	KMNM, LINMAP, MONANOVA, PREFMAP, CSP	MNL, MNP, CLM, HB, NCLA
Poziom estymacji	indywidualny, segmentowy	zagregowany, segmentowy, indywidualny
Obszary wykorzystania wyników	segmentacja, symulacja udziałów w rynku	szacowanie udziałów w rynku i wielkości popytu

Źródło: [Bąk 2004a; Green, Srinivasan 1978].

<sup>1</sup> Model ten znajduje się w nurcie teorii poznawczych, których przedstawicielami w marketingu są również Lutz, Bettman, Rosenberg, Wilkie i Pessemier (por. [Sagan 2004]).

<sup>2</sup> W metodach dekompozycyjnych użyteczności całkowite są to oceny profili, które stanowią podstawę dalszej analizy, która polega na dekompozycji użyteczności całkowitych profili na użyteczności cząstkowe poziomów atrybutów oraz na oszacowaniu udziałów poszczególnych atrybutów w kształtowaniu użyteczności całkowitej każdego profilu (por. [Bąk 2004a, s. 48-49]).

W **podjęciu mieszanym** formułowane są modele łączące cechy obu opisanych powyżej podejść. Zastosowanie znajdują tu przede wszystkim modele hybrydowe *conjoint analysis* oraz adaptacyjna *conjoint analysis*. W obu tych metodach stosuje się dwufazowe procedury pomiaru preferencji (por. [Bąk 2004a, s. 44]). I etap procedury to bezpośrednie oceny atrybutów i ich poziomów, zaś II etap to ocena wybranych par lub podzbiorów profilów produktów lub usług.

Metody *conjoint analysis* i metody wyborów dyskretnych należą do wspólnej grup metod reprezentujących podejście dekompozycyjne. Ich ogólna procedura badawcza jest taka sama, lecz istotne różnice pojawiają się na etapie gromadzenia danych i na etapie estymacji użyteczności częściowych. Różne są również ich podstawy teoretyczne. Tabela 1 przedstawia podobieństwa i różnice obu grup metod.

Należy w tym miejscu zauważyć, iż zaletą metod *conjoint analysis* jest możliwość oszacowania użyteczności na poziomie indywidualnym respondenta, a co za tym idzie – możliwość przeprowadzenia segmentacji rynku. Wadą podejścia są natomiast dalekie od rzeczywistych wyborów rynkowych sposoby oceny profilów, a także ograniczona liczba atrybutów wykorzystywanych w badaniu.

W przypadku metod wyborów dyskretnych podstawową zaletą jest imitacja rzeczywistych zachowań konsumentów za sprawą m.in. możliwości rezygnacji w badaniu z dokonywania wyboru. Głównym mankamentem metod opartych na wyborach jest niedobór obserwacji wykorzystywanych do szacowania parametrów modelu dyskretnych wyborów. Tym samym w typowych zastosowaniach parametry tych modeli szacuje się przede wszystkim na poziomie zagregowanym, tzn. w przekroju całej próby (por. tab. 1).

## 2. Podstawowe problemy ze skalą pomiaru

W badaniach pomiar preferencji respondentów może być przeprowadzony na skalach mocnych (metody *conjoint analysis*) lub na skalach słabych (metody wyborów dyskretnych). Pomiar na skalach mocnych dostarcza więcej informacji, natomiast pomiar na skalach słabych wierniej odzwierciedla rzeczywiste decyzje rynkowe.

Konsekwencją pomiaru preferencji na skalach słabych, w przypadku estymacji modelu na poziomie zagregowanym, jest – oprócz abstrakcyjności skali porządkowej oraz niedoboru danych – konieczne do przyjęcia założenie jednorodności (homogeniczności) preferencji konsumentów. Uznanie homogeniczności preferencji skutkuje m.in. (por. [Bąk 2005, s. 182]):

- jednorodnością parametrów modelu,
- możliwością wystąpienia efektu tzw. złudnej większości<sup>3</sup>,
- występowaniem w szacowanych modelach własności niezależności od dodawanych opcji<sup>4</sup> (IIA – *Independence of Irrelevant Alternatives*),

<sup>3</sup> Efekt ten polega na uśrednianiu wyników dokonywanych przez respondentów wyborów – np. jeśli połowa respondentów woli małe samochody, a druga połowa duże, nie oznacza to, iż przeciętny respondent preferuje samochód o średniej wielkości (por. [Bąk 2005, s. 182]).

<sup>4</sup> Własność ta zakłada stałość ilorazu dwóch prawdopodobieństw wyboru niezależnie od innych opcji dodawanych do zbioru. Literatura przedmiotu dostarcza przykłady braku konsekwencji w dokonywanych przez respondentów wyborach w przypadku pojawiających się innych profilów – *red bus/blue bus problem* (por. [Green, Srinivasan 1978; Zwierna 1997]).

- brakiem możliwości przeprowadzenia segmentacji konsumentów,
- niską trafnością prognostyczną tzw. symulatorów wyboru.

Jak podaje literatura przedmiotu, podstawową przyczyną problemów związanych z dyskretną skalą pomiaru jest występująca w rzeczywistości niejednorodność preferencji konsumentów. Do jej źródeł zalicza się (por. [Bąk 2005, s. 183; DeSarbo i in. 1997, s. 337-338]):

- niejednorodność reakcji respondentów wynikającą z odmiennego traktowania przez respondentów poszczególnych wartości skali preferencji,
- niejednorodność strukturalną, która wynika z różnej wagi przywiązywanej przez konsumentów do poszczególnych poziomów atrybutów,
- niejednorodność percepcji cech produktu wynikającą z odmiennego postrzegania atrybutów produktu lub usługi,
- niejednorodność postaci modelu wynikającą z indywidualnych sposobów oceny przez respondentów użyteczności całkowitej produktu lub usługi,
- niejednorodność rozkładu błędu wynikającą z różnych rozkładów składnika losowego dla różnych respondentów lub ich segmentów,
- niejednorodność w czasie, która wynika ze zmienności w czasie postaw i preferencji indywidualnych respondentów.

W przypadku niejednorodności preferencji zastosowanie znajdują metryczne modele dekompozycyjne – *conjoint analysis*, w których użyteczności szacowane są na poziomie indywidualnym (dla każdego respondenta szacowany może być odrębny model). W przypadku założenia jednorodności preferencji wykorzystywane są niemetryczne modele wyborów dyskretnych, w których parametry szacowane są na poziomie zagregowanym (jeden model dla całej próby). Należy podkreślić, iż w swoich klasycznych postaciach modele te nie uwzględniają zjawiska niejednorodności preferencji. Oznacza to potrzebę rozszerzenia klasy modeli o modele niemetryczne uwzględniające niejednorodność preferencji.

### 3. Niemetryczne modele heterogenicznych preferencji

Wśród propozycji modeli umożliwiających oszacowanie parametrów również na poziomie segmentowym lub indywidualnym, bez konieczności przyjmowania założenia o jednorodności preferencji, literatura przedmiotu wskazuje na modele z parametrami losowymi oraz modele klas ukrytych (por. [Bąk 2004b, s. 183]).

W **modelach z parametrami losowymi** przyjmuje się założenie, iż szacowane parametry modelu nie są nieznanymi wielkościami stałymi, lecz zmiennymi losowymi o określonych rozkładach (por. [Bąk 2004b, s. 190-192]). W swoich klasycznych zastosowaniach modele te są wykorzystywane do badania preferencji, gdy występuje niedobór danych z próby oraz znane są dodatkowe informacje o preferencjach konsumentów<sup>5</sup>.

---

<sup>5</sup> Na przykład większość konsumentów preferuje niższą cenę.

Przedstawicielem tych modeli jest **hierarchiczny model Bayesa**, który jest bardzo efektywną techniką „pożyczania danych”, pozwalającą na estymację indywidualnych użyteczności cząstkowych (por. tab. 2), wykorzystując informacje pochodzące nie tylko od danego respondenta, lecz również od innych respondentów z badanej grupy. Model ten opiera się na twierdzeniu Bayesa o prawdopodobieństwie warunkowym wystąpienia przyczyny (hipotezy), jeżeli nastąpił skutek (zaszło określone zdarzenie).

Model hierarchiczny Bayesa składa się z dwóch poziomów. Na wyższym poziomie przyjmuje się założenie, iż parametry (indywidualne użyteczności cząstkowe) określone są wielowymiarowym rozkładem normalnym. Na niższym poziomie przyjmuje się, iż indywidualne prawdopodobieństwa wyboru określonych profili opisane są za pomocą wielomianowego modelu logitowego (bądź też za pomocą regresji liniowej). Estymacja modelu polega na oszacowaniu indywidualnych użyteczności cząstkowych według przyjętych założeń o określonych rozkładach. Należy zauważyć, iż w praktyce, ze względu na konieczność znalezienia rozkładu wielu zmiennych, jest to trudne obliczeniowo zadanie. Stosowane algorytmy pozwalają jednak na dobre oszacowania parametrów modelu<sup>6</sup>.

Tabela 2. Charakterystyka niemetrycznych modeli uwzględniających niejednorodność preferencji

Wyszczególnienie	Model hierarchiczny Bayesa	Modele klas ukrytych
Zastosowanie	identyfikacja preferencji indywidualnych	segmentacja
Wykorzystywane informacje	dane z próby, informacje <i>a priori</i>	dane z próby, „brakujące dane”
Poziom agregacji	indywidualny	segmentowy, indywidualny
Metoda estymacji	metoda MCMC	metoda E-M
Liczba parametrów	większa niż liczba obserwacji	zależna od liczby i wielkości segmentów
Oprogramowanie komputerowe	CBC/HB Sawtooth Software, autorskie programy źródłowe w językach: FORTRAN, C, GAUSS lub w środowisku R	LCSM Sawtooth Software, SPSS, SAS/STAT, LATENT GOLD, GLIMMIX

Źródło: opracowanie własne na podstawie [Bąk 2004b, s. 192].

Do podstawowych cech modelu z parametrami losowymi (hierarchicznego modelu Bayesa) zalicza się (por. [Bąk 2004b, s. 192]):

- założenie o heterogeniczności preferencji,
- estymację indywidualnych użyteczności cząstkowych na podstawie wyborów,
- możliwość estymacji dużej liczby parametrów (większej niż liczba obserwacji),
- wykorzystywanie informacji *a priori*,
- dużą złożoność obliczeniową algorytmów estymacji parametrów,
- brak konkurencyjnego oprogramowania.

W badaniach z wykorzystaniem **modeli klas ukrytych** zakłada się, że w badanej próbie istnieje skończona liczba grup konsumentów o podobnych preferen-

<sup>6</sup> Wśród stosowanych metod symulacyjnych dobre wartości szacowanych parametrów dostarczają metody Monte Carlo, łańcuchy Markowa oraz metody Gibbsa i Metropolisa-Hastingsa (por. [Bąk 2004b, s. 191]).

cjach, natomiast między grupami występują istotne różnice (por. [Bąk 2004b, s. 187-190]). Grupy te nie są znane *a priori*, tylko są „ukryte”, gdyż nie są znane przynależności poszczególnych konsumentów do określonych grup ani liczba grup. Tym samym „ukryte” są także liczebności poszczególnych segmentów.

W statystyce modele klas ukrytych zalicza się do grupy modeli rozkładów mieszanych, które tworzone są przez określoną liczbę rozkładów składowych. Udział każdego rozkładu składowego jest określony przez tzw. parametr mieszający, który w przypadku modeli klas ukrytych wskazuje na rozmiar segmentu. Estymacja modelu polega tym samym na oszacowaniu liczby i wielkości poszczególnych segmentów.

Modele te uwzględniają heterogeniczność preferencji konsumentów na poziomie grupowym (segmentowym). Modele klas ukrytych „rozdzielają” próbę na segmenty danego rynku (różniące się preferencjami respondentów) oraz oszacowują użyteczności przedstawiające preferencje każdego z segmentów. Segmentowe użyteczności cząstkowe szacowane są z wykorzystaniem metody największej wiarygodności. Wraz z wyznaczeniem wielkości segmentów w dalszym etapie umożliwia to oszacowanie wartości użyteczności cząstkowych także na poziomie indywidualnym (por. tab. 2).

Do podstawowych cech modeli klasy ukrytych zalicza się (por. [Bąk 2004b, s. 190]):

- założenie o heterogeniczności preferencji,
- zastosowanie w segmentacji konsumentów,
- możliwość powiększenia zasobu informacyjnego danych empirycznych,
- zastosowanie w estymacji algorytmu E-M (dekompozycyjnych modeli metrycznych i niemetrycznych),
- monotoniczną poprawę wartości funkcji wiarygodności,
- generowanie dużej liczby segmentów,
- problem ustalenia liczby segmentów oraz maksimum lokalnego.

#### 4. Podsumowanie

Modele z parametrami losowymi oraz modele klas ukrytych stanowią uzupełnienie dotychczasowych możliwości pomiaru preferencji wyrażonych. Do podstawowych zalet obydwu grup modeli zaliczyć należy możliwość uwzględnienia heterogeniczności preferencji w metodach wyborów dyskretnych w przypadku pomiaru preferencji na skali nominalnej. Jak wskazuje na to literatura przedmiotu (por. [Bąk 2005, s. 188]), obie grupy modeli umożliwiają osiągnięcie podobnych celów badawczych. Wśród „dyskretnych” i nielicznych różnic między obydwoma grupami wyróżnić należy następujące zalety wykorzystania modeli klas ukrytych<sup>7</sup>:

- w porównaniu z modelem hierarchicznym Bayesa zastosowanie analizy skupień dostarcza lepszych rezultatów w przypadku modeli klas ukrytych,
- w przypadku podziału respondentów na zwarte i zróżnicowane segmenty modele klas ukrytych pozwalają modelować wybory respondentów wierniej niż przy wykorzystaniu hierarchicznego modelu Bayesa;

<sup>7</sup> Na podstawie: [The CBC... 2004].

oraz atuty modelu hierarchicznego Bayesa:

- w większości przypadków oszacowanie indywidualnych preferencji konsumentów za pomocą hierarchicznego modelu Bayesa przewyższa dokładnością wyniki uzyskane z modeli klas ukrytych,
- symulacje rynkowe wskazują na bardziej efektywną redukcję problemu niezależności od dodawanych opcji (IIA – independence of irrelevant alternatives) przy zastosowaniu hierarchicznego modelu Bayesa.

## Literatura

- Bąk A. (2000), *Conjoint analysis jako metoda pomiaru postaw i preferencji konsumentów*, [w:] M. Walesiak (red.), *Pomiar w badaniach rynkowych i marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 856, AE, Wrocław, s. 69-81.
- Bąk A. (2004a), *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1013, AE, Wrocław.
- Bąk A. (2004b), *Estymacja parametrów w modelach dyskretnych wyborów*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1022, AE, Wrocław, s. 183-194.
- Bąk A. (2005), *Problemy estymacji parametrów w modelach dekompozycyjnych z dyskretną zmienną objaśnianą*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław, s. 180-189.
- DeSarbo W.S., Ansari A., Chintagunta P., Himmelfarb C., Jedidi K., Johnson R., Kamakura W., Lenk P., Srinivasan K., Wedel M. (1997), *Representing heterogeneity in consumer response models*, „Marketing Letters” no 8:3, s. 335-348.
- Green P.E., Srinivasan V. (1978), *Conjoint analysis in consumer research: issues and outlook*, „Journal of Consumer Research” no 5, September.
- Hair J.F., Anderson R.E., Tatham R.L., Black W.C. (1995), *Multivariate data analysis with readings*, Englewood Cliffs, Prentice-Hall.
- Sagan A. (2004), *Modele zachowań konsumentów*, [www.cem.pl/?a=pages&id=42](http://www.cem.pl/?a=pages&id=42), CEM.
- The CBC Latent Class Technical Paper* (2004), Technical Paper from Sawtooth Software, <http://www.sawtoothsoftware.com/download/techpap/lctech.pdf>, Sawtooth Software.
- Walesiak M., Bąk A. (1997), *Realizacja badań marketingowych metodą conjoint analysis z wykorzystaniem pakietu statystycznego SPSS for Windows*, AE, Wrocław.
- Zwerina K. (1997), *Discrete choice experiments in marketing*, Heidelberg-New York, Physica-Verlag.

## PROBLEM OF DATA AGGREGATION LEVEL IN DISCRETE CHOICE MODELS

### Summary

The paper presents a decompositional approach to measurement of consumers' preferences: *conjoint analysis methods* and *discrete choice methods*. The utilities are estimated at individual level in *conjoint analysis*. In *discrete choice methods* utilities are estimated at aggregate level, but with latent class models we can estimate them at segmentation level. In order to estimate utilities at individual level there should be used hierarchical Bayes models.

In the article, the basic characteristics of latent class models and hierarchical Bayes models as well as the benefits of their use in *discrete choice methods* were described. Additionally, software being used with both kinds of models was presented.