Milda BURZAŁA

# KILKA UWAG NA TEMAT WYBORU MODELU DYSKRETNEGO DO PROGNOZOWANIA AKTYWNOŚCI GOSPODARCZEJ POLSKI

W modelach ekonometrycznych z reguły wykorzystujemy zmienne ilościowe. W praktyce natomiast wiele zjawisk ekonomicznych i społecznych ma charakter jakościowy z ograniczoną liczbą stanów, które można wyróżnić. Z reguły dotyczą one zachowania się poszczególnych jednostek (gospodarstw domowych, inwestorów, konsumenta na rynku) i związane są z wyborem jednej z dostępnych możliwości, kategorii, stanów (zakup danego produktu lub nie, wyjazd na wakacje w kraju lub za granicę, wybór rodzaju ogrzewania czy też sposobu finansowania rozwoju w przedsiębiorstwie). Stąd też zmienne, które je opisują, są zmiennymi dyskretnymi, a modele wyjaśniające zmiany tego typu wielkości nazywane są modelami wyboru dyskretnego lub modelami z jakościową zmienną objaśnianą. Mają one na celu określenie prawdopodobieństwa wyboru danej kategorii, jeśli znane są pewne ważne charakterystyki decydenta lub(i) charakterystyki danej kategorii wyboru.

W niniejszym artykule rozważane będą zastosowania modeli jakościowych w analizie zjawisk makroekonomicznych do prognozowania faz aktywności gospodarczej. Inwestora czy innego decydenta często bardziej interesuje kierunek i dynamika zmian gospodarczych niż konkretne wartości określonych wskaźników. Z tego punktu widzenia uzasadnione wydaje się założenie, że aktywność gospodarcza jest stanem jakościowym.

Mimo dużego zróżnicowania kolejnych cykli koniunkturalnych istniejące prawidłowości w zakresie zmian podstawowych wskaźników makroekonomicznych umożliwiają ich prognozowanie. Coraz większego znaczenia w badaniach empirycznych nabiera metoda wskaźników wiodących oraz prognostyczna wartość modeli budowanych z ich wykorzystaniem. Polega ona na doborze jednego wskaźnika lub grupy wskaźników sygnalizujących z wyprzedzeniem przyszłe stany koniunktury.

Podstawowy w tej dziedzinie jest artykuł A. Estrella, i F.S. Mishkin [8]. Autorzy na binarnym modelu probitowym testowali makroekonomiczne wskaźniki wyprzedzające nadejście kryzysu gospodarczego w Stanach Zjednoczonych. Jakościową zmienną objaśnianą zdefiniowano jako $y=1$, jeśli gospodarka w danym kwartale była w recesji, natomiast $y=0$ w przeciwnym wypadku. Zmienną dla recesji określono na podstawie danych publikowanych i opracowanych przez National Bureau of Economic Research (NBER) z lat 1959-1995. Okazało się, że różnica oprocentowania dziesięcioletnich obligacji skarbowych oraz trzymiesięcznych weksli skarbowych (tzw. *spread*) dobrze sygnalizuje z czterokwartalnym wyprzedzeniem kryzys gospodarczy w Stanach Zjednoczonych.

Bezpośrednia aplikacja tych samych modeli na gruncie polskiej gospodarki stała się niemożliwa ze względu na trudności w zdefiniowaniu faz aktywności gospodarczej, a tym samym jakościowej zmiennej objaśnianej. W badaniach empirycznych dominują dwie alternatywne metody wyodrębnienia wahań koniunkturalnych na podstawie szeregów czasowych wybranych wskaźników ([2], rozdz. 4). Koncepcja cyklu odchyleń przedstawia wahania koniunkturalne jako odchylenia wartości wskaźnika (względne lub bezwzględne) od linii trendu. Natomiast periodyzacja faz cyklu wzrostu wymaga określenia normalnej stopy wzrostu, za którą przyjmuje się zazwyczaj stopę wzrostu wynikającą z trendu lub przeciętną stopę wzrostu dla danego cyklu koniunkturalnego. W Polsce krótki okres niestabilnej gospodarki rynkowej uniemożliwia analizę aktywności ekonomicznej klasycznymi metodami głównie ze względu na trudności w określeniu długookresowej tendencji rozwojowej (trendu) naszej gospodarki lub przeciętnej stopy wzrostu dla danego cyklu. Przy krótkich szeregach czasowych istnieje niebezpieczeństwo tego, że zmiany wynikające z przebiegu cyklu koniunkturalnego będą interpretowane jako zmiany długookresowe.

W artykule przedstawiono krótki przegląd modeli wyboru dyskretnego na gruncie przyjmowanych, zróżnicowanych założeń teoretycznych. Dla prognozowania aktywności gospodarczej propozycje dotyczą w szczególności

- zdefiniowania zmiennej objaśnianej,
- wyboru modelu wielomianowego.

## 1. Określenie jakościowej zmiennej objaśnianej

Gdy brakuje jasnych reguł dla określenia stanu aktywności gospodarczej, stajemy przed problemem określenia metody klasyfikacji obserwacji (obiektów) do poszczególnych faz aktywności gospodarczej (klas). Zadanie periodyzacji

polega na tym, aby podzielić zbiór wyników obserwacji wybranego wskaźnika referencyjnego na rozłączne podzbiory, które będzie można identyfikować z kolejnymi fazami aktywności gospodarczej. W literaturze można spotkać kilka propozycji wskaźnika, który dobrze obrazuje aktualną sytuację gospodarczą. W badaniach empirycznych najczęściej wykorzystuje się indeks produkcji przemysłowej ze względu na dostępność danych empirycznych oraz dużą korelację zmian indeksu ze zmianami realnego PKB.

Klasycznym sposobem określenia reguł klasyfikacji, gdy nie dysponujemy informacją na temat struktury klas (brakuje wzorca), jest analiza skupień (*taksonomia numeryczna*). Proponowane w literaturze algorytmy numeryczne wykorzystują kryteria klasyfikacji oparte na wartościach miary podobieństwa obiektów (np. miary odległościowe, współczynnik korelacji), które mierzą stopień wewnętrznej jednorodności klas lub(i) międzyklasową heterogeniczność[1]. Najbardziej ogólny miernik odległości przedstawia formuła Minkowskiego:

$$d_{ik}=\left[\frac{1}{N}\sum_{n=1}^{N}|y'_{in}-y'_{kn}|^{p}\right]^{1/p}, \qquad i, k=1,\ldots, T, \tag{1}$$

gdzie $y'_{in}$, $y'_{kn}$ to znormalizowane wartości zmiennej $y_n (n=1,\ldots, N)$ dla $i$ oraz $k$ obserwacji. Dla $p=1$ otrzymujemy tzw. odległość miejską (Hamminga), natomiast dla $p=2$ odległość Euklidesa.

Wyniki klasyfikacji uzyskane w ramach taksonomii numerycznej często są trudne w interpretacji. Jedyny sposób opisu polega na wymienieniu obiektów należących do danej klasy czy też określeniu ich wewnętrznego zróżnicowania, nie są bowiem znane reguły klasyfikacji (poza funkcją klasyfikacyjną). Proponowane algorytmy stawiają różnorodne wymagania, na przykład co do normalności rozkładu cech czy też jednorodności jednostek miary.

Alternatywne podejście bliższe jest taksonomii symbolicznej, gdzie klasa (faza) jest opisywana przez koniunkcję lub dysjunkcję wartości cech (zmiennych makroekonomicznych). W przeciwieństwie do taksonomii numerycznej taksonomia symboliczna umożliwia nie tylko uzyskanie *podziału na klasy*, ale również znalezienie *charakterystyk tych klas*. Nie stawia tu się żadnych wymagań co do rozkładu analizowanych wartości, oceny jakości podziału mają charakter nieparametryczny.

[1] Przegląd mierników odległości zawierają między innymi prace: *Metody statystyki międzynarodowej*, red. A. Zeliaś, PWE, Warszawa 1988, s. 170-173; T. Grabiński, S. Wydymus, A. Zeliaś, *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa 1989, s. 29-31.

W prezentowanym podejściu dla określenia reguł klasyfikacji obserwacji wykorzystano stopy wzrostu w postaci dwóch rodzajów *indeksów dynamiki produkcji przemysłowej* oczyszczone z wahań przypadkowych i wahań sezonowych.

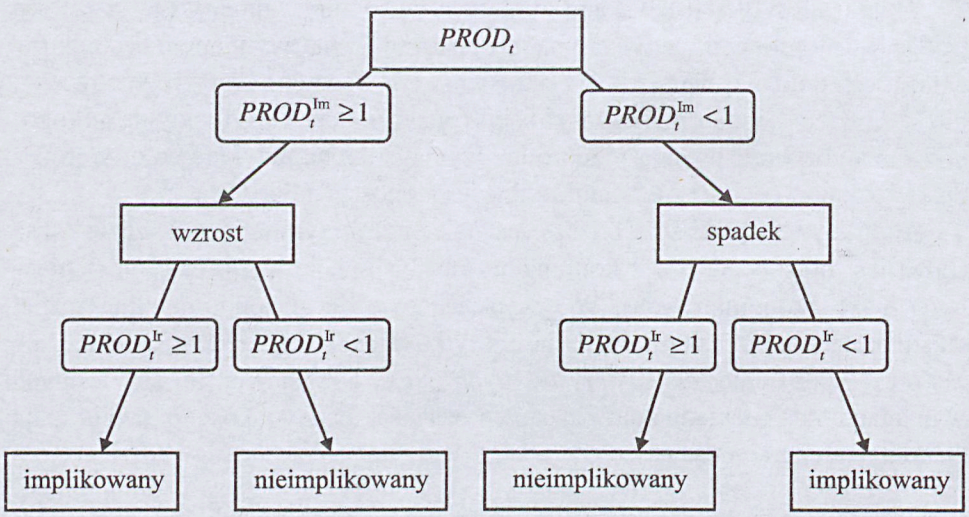
*Indeks roczny* mierzy zmianę wartości w stosunku do tego samego okresu (miesiąca) roku poprzedniego i stanowi miarę zmian „w dłuższym okresie"

$$PROD_t^{\mathrm{Ir}} = \frac{PROD_t}{PROD_{(t-12)}}. \tag{2}$$

*Indeks miesięczny* mierzy zmianę wartości w stosunku do poprzedniego okresu (miesiąca) i jest miarą zmian „w krótkim okresie"

$$PROD_t^{\mathrm{Im}} = \frac{PROD_t}{PROD_{(t-1)}}. \tag{3}$$

Propozycja przedstawienia reguł podziału zbioru wszystkich obserwacji na rozłączne podzbiory (fazy aktywności gospodarczej) wykorzystuje dwupoziomowe drzewo klasyfikacji obserwacji. Budowę drzewa oparto na testach sprawdzających czy krótkookresowa stopa wzrostu implikuje zmiany długoterminowe czy też nie. W zależności od tego, czy dany indeks jest




**Rys. 1. Drzewo klasyfikacji obserwacji ($PROD_t$ – indeks produkcji przemysłowej)**

większy czy mniejszy od 1 (co odpowiada dodatniej lub ujemnej stopie wzrostu) klasyfikujemy obserwację $t$ jako charakterystyczną dla danego stanu gospodarki (rysunek 1).

Podane reguły pozawalają wyróżnić cztery stany gospodarki:

$PROD_t^{Ir} \geqslant 1$ i $PROD_t^{Im} \geqslant 1$ wzrost implikowany oznacza wysoką aktywność gospodarczą (umownie – faza dobrej koniunktury, symbol m1_r1 na rysunku 2);

$PROD_t^{Ir} < 1$ i $PROD_t^{Im} \geqslant 1$ wzrost nieimplikowany oznacza wzrost indeksu rocznego po recesji (m1_r0);

$PROD_t^{Ir} < 1$ i $PROD_t^{Im} < 1$ spadek indeksu rocznego implikowany przez spadek indeksu miesięcznego (recesja, m0_r0),

$PROD_t^{Ir} \geqslant 1$ i $PROD_t^{Im} < 1$ spadek nieimplikowany oznacza spadek indeksu rocznego przed recesją (m0_r1).

Uzyskany podział zbioru wyników obserwacji porównano z podziałem dokonanym według operatorów różnicowych dla rocznego indeksu produkcji przemysłowej, których znaki wiążą się z konkretnymi zmianami wskaźników.

Oznaczmy pierwszą i drugą różnicę jako

$$PROD_t^{D1} = PROD_t^{Ir} - PROD_{(t-1)}^{Ir} \tag{4}$$

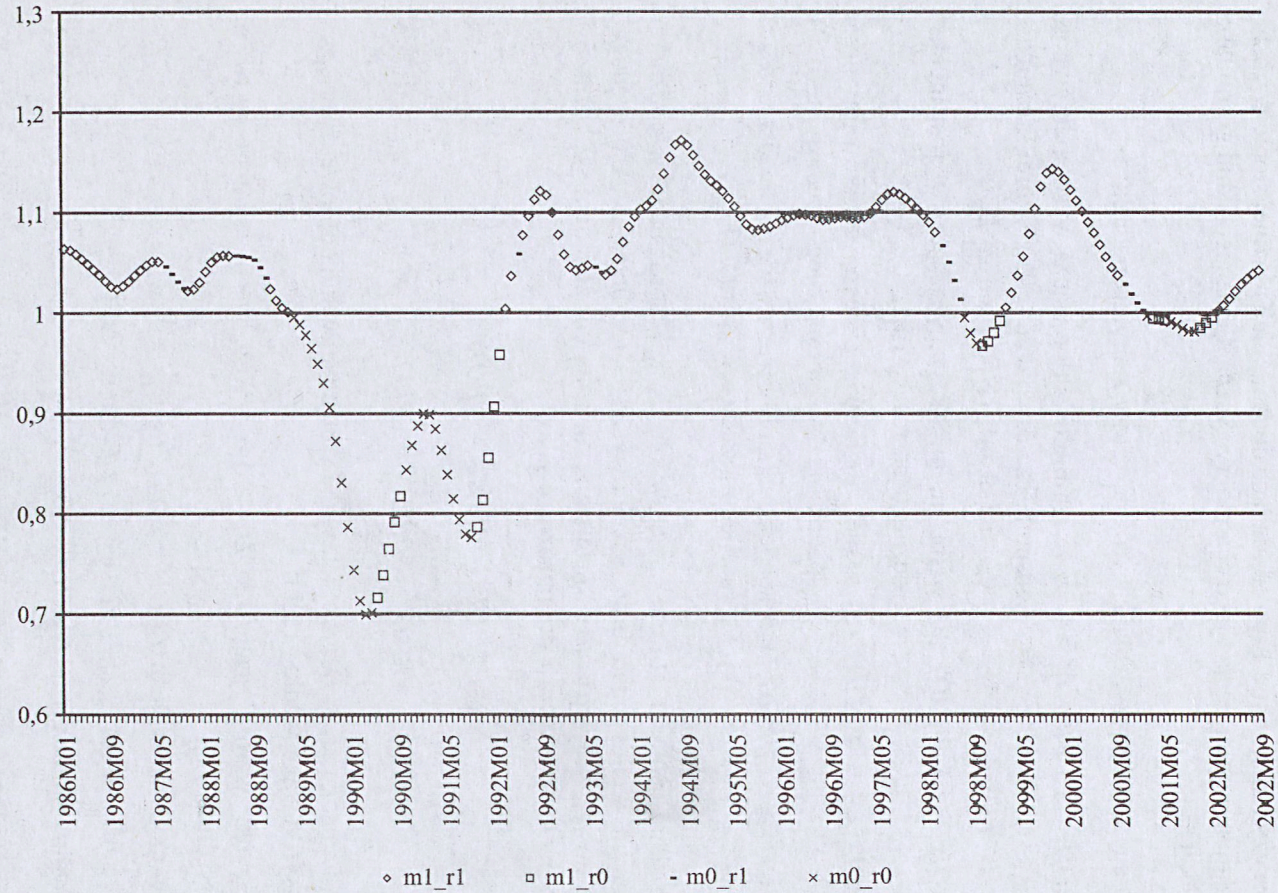
oraz

$$PROD_t^{D2} = PROD_t^{D1} - PROD_{(t-1)}^{D1}. \tag{5}$$

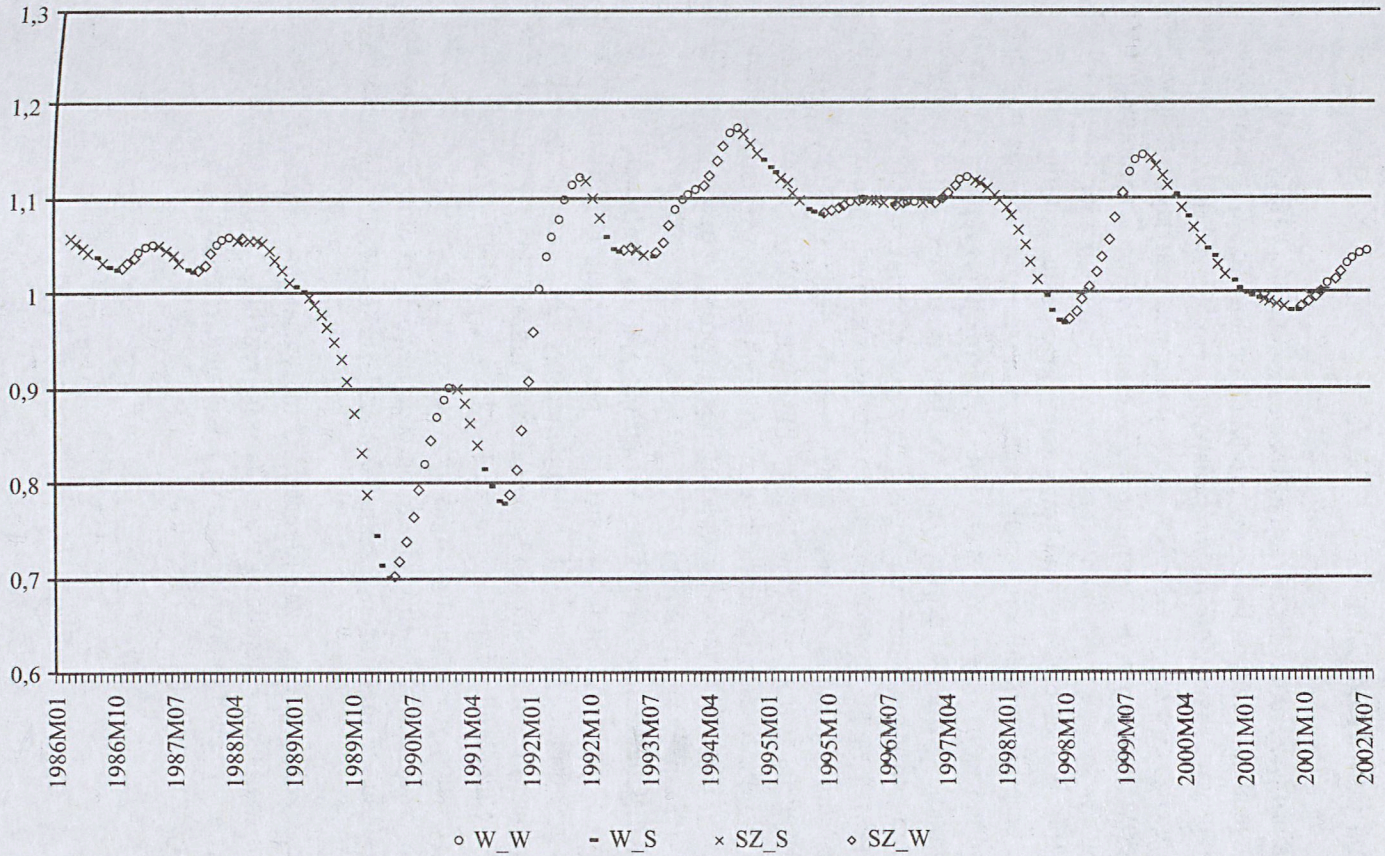
W tym wypadku drzewo klasyfikacyjne byłoby też dwupoziomowe, a obserwacje klasyfikowane według reguł:

- szybki wzrost, jeśli $PROD_t^{D1} \geqslant 0$ i $PROD_t^{D2} \geqslant 0$ (symbol sz_w na rysunku 3),
- wolny wzrost, jeśli $PROD_t^{D1} \geqslant 0$ i $PROD_t^{D2} < 0$ (w_w),
- szybki spadek, jeśli $PROD_t^{D1} < 0$ i $PROD_t^{D2} < 0$ (sz_s),
- wolny spadek, jeśli $PROD_t^{D1} < 0$ i $PROD_t^{D2} \geqslant 0$ (w_s).

Zaletą proponowanych podziałów jest niezmienność wyróżnionych faz. Zauważmy, że parametry trendu długookresowego, w miarę napływu informacji, mogą ulec zmianie, co wiąże się ze zmianą klasyfikacji. Przebieg współczesnego cyklu koniunkturalnego opisywany w literaturze jest dwufazowy, stąd dalszą rozbudowę drzewa można uznać za niecelową.




**Rys. 2. Fazy aktywności gospodarczej Polski w latach 1986-2002, wyznaczone na podstawie wartości rocznych i miesięcznych indeksów produkcji przemysłowej**




**Rys. 3. Fazy aktywności gospodarczej Polski w latach 1986-2002, wyznaczone na podstawie znaków operatorów różnicowych rocznych indeksów produkcji przemysłowej**

## 1.1. Ocena jakości podziału

Podział zbioru wszystkich obserwacji na rozłączne podzbiory (fazy aktywności) umożliwia analizę rocznych indeksów produkcji przemysłowej w podzbiorach.

W przypadku danych ilościowych podstawą oceny jakości podziału najczęściej jest analiza wariancji ANOVA[2]. Na podstawie statystyki testowej $F$, odnoszącej wariancje między grupami do wariancji wewnątrzgrupowej, rozstrzygamy, czy średnie w uzyskanych podzbiorach (przynajmniej dwie) istotnie się od siebie różnią czy nie. Wykorzystanie tego testu jest możliwe przy założeniu, że próby z populacji zostały pobrane niezależnie, a każda populacja ma rozkład normalny o tej samej wariancji. Jeśli istnieje przypuszczenie, że nie jest spełnione założenie o normalności rozkładu populacji, należy stosować testy nieparametryczne, które nie wymagają założeń, dotyczących rozkładu.

Do sprawdzenia poprawności podziału szeregu rocznych indeksów produkcji wykorzystano nieparametryczny test Kruskala-Wallisa[3]. Uzasadnieniem jest nieparametryczna metoda podziału zbioru obserwacji. Poza tym w modelach wyboru dyskretnego przyjmuje się różne założenia co do rozkładu składnika losowego w populacji. Tylko w przypadku modelu probitowego wymagane jest założenie o normalności rozkładu.

Test Kruskala-Wallisa jest swego rodzaju analizą wariancji, w której zamiast wartości obserwacji używa się rang, przypisanych kolejnym uporządkowanym wartościom obserwacji, od najmniejszej do największej. Jedynym założeniem jest niezależność prób losowych, pochodzących z $k$ populacji. Test jest wrażliwy na różnice w położeniu, stąd wykorzystywany jest do testowania równości średnich w $k$ populacjach. W przypadku gdy $k>2$ nie jest zalecane testowanie wszystkich możliwych kombinacji dwóch populacji, chociażby z tego względu, że nie możemy zakładać niezależności par testów w ramach $k$ populacji.

Hipoteza zerowa i hipoteza alternatywna mają następującą postać:

$H_0$: wszystkie $k$ populacji mają taki sam rozkład,

$H_1$: nie wszystkie populacje mają takie same rozkłady.

Statystyką testu jest:

$$H=\frac{12}{n(n+1)}\left(\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right)-3(n+1), \tag{6}$$

---

[2] Szczegółowy opis testu można znaleźć w: [11], s. 99, oraz w [1], s. 390.

[3] Omówienie testu na podstawie [1], s. 731.

gdzie:

$n_j$ – liczebność próby z populacji $j$ $(j=1,\ldots,k)$,

$n=\sum_{j=1}^{k} n_j$ – liczebność całej próby,

$R_j$ – suma rang w obrębie próby populacji $j$.

Dla prób większych od 5 rozkład statystyki $H$ jest dobrze przybliżony przez rozkład $\chi^2$ z $k-1$ stopniami swobody. Hipotezę $H_0$ odrzucamy dla $H>\chi^2_{(k-1)}$ przy danym poziomie istotności $\alpha$.

W przypadku odrzucenia hipotezy $H_0$ na korzyść hipotezy $H_1$ powstaje pytanie, które z populacji istotnie się różnią. W tym celu obliczamy średnie rangi prób:

$$\bar{R}_i=\frac{R_i}{n_i} \quad \text{oraz} \quad \bar{R}_j=\frac{R_j}{n_j}. \tag{7}$$

Sprawdzianem hipotezy zerowej, że populacje $i$ oraz $j$ są takie same, jest:

$$D=|\bar{R}_i-\bar{R}_j|. \tag{8}$$

Hipotezę $H_0$ odrzucamy, gdy statystyka testowa $D$ jest większa od wartości krytycznej $C_{KW}$ $(D>C_{KW})$ obliczonej ze wzoru:

$$C_{KW}=\sqrt{\chi^2_{\alpha,k-1}\left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{9}$$

gdzie $\chi_{\alpha,k-1}$ oznacza wcześniej odczytaną wartość krytyczną rozkładu $\chi^2$. Porównując wartości $D$ i $C_{KW}$ możemy sprawdzić wszystkie pary populacji.

W tabeli 1 znajdują się wyniki testu Kruskala-Wallisa w celu weryfikacji hipotezy o jednakowym rozkładzie w 4 populacjach wyróżnionych na podstawie rocznych i miesięcznych indeksów produkcji przemysłowej Polski zaobserwowanych od stycznia 1986 roku do grudnia 2002 roku.

**Tabela 1. Wyniki I części testu Kruskala-Wallisa dla sprawdzenia hipotezy zerowej o braku różnic między populacjami**

| Kryterium podziału | $H$ | $\chi^2_{0,05,k-1}$ |
|---|---|---|
| Wartości indeksów | 130,68 | 7,81 |
| Znaki operatorów różnicowych | 15,39 | 7,81 |

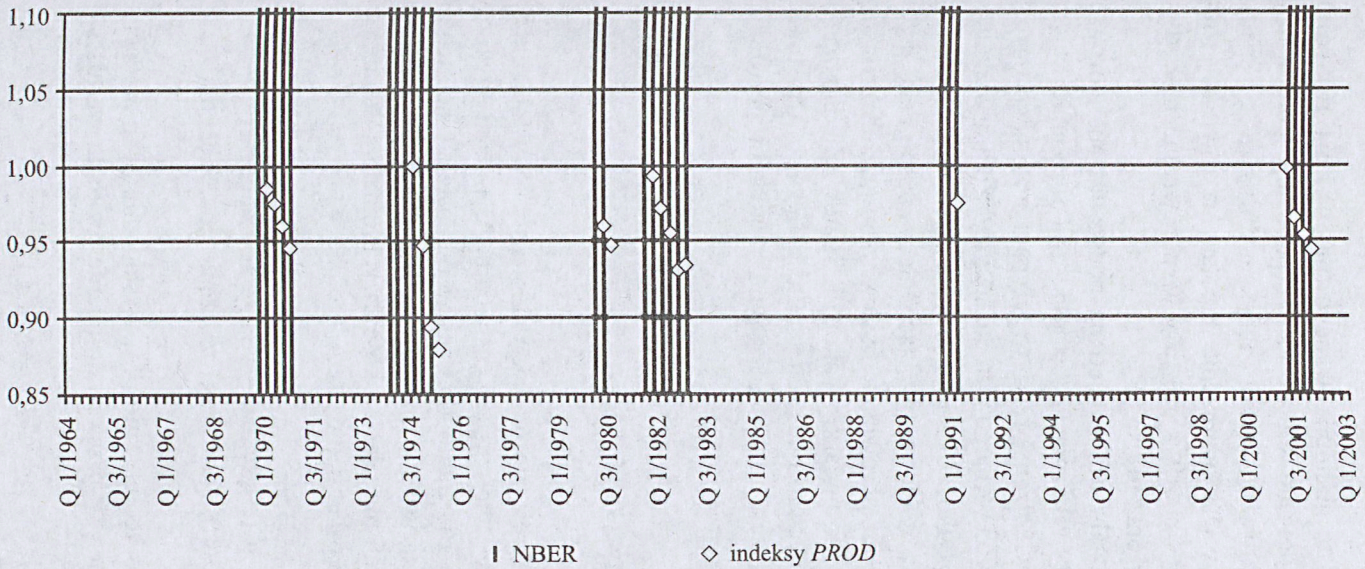
Wartość empirycznej statystyki $H$ jest w obu wypadkach większa od wartości krytycznej $\chi^2$ dla $\alpha=0{,}05$ i 3 stopni swobody. Hipotezę zerową o równości rozkładów odrzucamy zarówno w przypadku podziału według wartości indeksów, jak i znaków operatorów różnicowych. Oznacza to, że cztery wyróżnione populacje rocznych indeksów produkcji nie mają takich samych rozkładów (przynajmniej dwa z nich są różne).

Dalsza analiza dotyczy sprawdzenia, gdzie występują różnice. W tym celu porównano wszystkie możliwe pary populacji. Wyniki dotyczące sprawdzenia hipotezy zerowej o równości rozkładów w dwóch populacjach znajdują się w tabeli 2.

Analiza wartości statystyk $D$ i $C_{KW}$ umożliwia porównanie par *łącznie* na poziomie istotności $\alpha$. Hipotezę o równości średnich odrzucamy, gdy $D > C_{KW}$. W przypadku populacji wyróżnionych na podstawie wartości indeksów brak różnicy między populacjami stwierdzamy tylko w przypadku spadku im-

**Tablica 2. Statystyki empiryczne $D$ i wartości krytyczne $C_{KW}$ (w nawiasie) dla II części testu Kruskala-Wallisa**

| Pary populacji uzyskane na podstawie wartości indeksów | | | | |
|---|---|---|---|---|
| Średnie rangi R | m1_r1 | m1_r0 | m0_r1 | m0_r0 |
| | 136,4 | 30,0 | 90,9 | 26,9 |
| m1_r1 | × | **106,4** (39,7) | **45,5** (38,9) | **109,5** (31,5) |
| m1_r0 | × | × | **60,9** (51,6) | 3,06 (46,3) |
| m0_r1 | × | × | × | **64,0** (45,6) |
| Pary populacji uzyskane na podstawie znaków operatorów różnicowych | | | | |
| Średnie rangi R | sz_w | w_w | w_s | sz_s |
| | 91,5 | 127 | 79,5 | 108,4 |
| sz_w | × | **35,5** (35,4) | 12,1 (34,3) | 16,9 (30,0) |
| w_w | × | × | **47,5** (37,4) | 18,6 (33,5) |
| w_s | × | × | × | 28,9 (32,4) |




**Rys. 4. Porównanie faz kryzysów gospodarczych Stanów Zjednoczonych wyznaczonych na podstawie rocznych i miesięcznych indeksów produkcji przemysłowej oraz danych publikowanych przez NBER** (www.census.gov)

plikowanego i wzrostu nieimplikowanego (m0_r0 i m1_r0). Mimo opisanego wyniku testu zdecydowano się nie łączyć populacji, ponieważ jedna z nich związana jest ze wzrostem wartości indeksu, druga zaś ze spadkiem. Dodatkowym argumentem jest odrzucenie hipotezy o równości rozkładu jednej i drugiej populacji (m0_r0 i m1_r0) w porównaniach z dwoma innymi populacjami (m0_r1 oraz m1_r1).

Wyniki dla podziału uzyskanego na podstawie operatorów różnicowych (druga część tabeli 2) wskazują, że jest to podział dużo mniej skuteczny. Istotną różnicę populacji stwierdzamy tylko w przypadku wolnego wzrostu i wolnego spadku. Dla szybkiego i wolnego wzrostu statystyka empiryczna jest większa od wartości krytycznej tylko o 0,1 i co prawda pozwala na odrzucenie hipotezy o równości rozkładów w tych populacjach, ale w sposób mało przekonywający.

Ciekawe okazało się porównanie faz kryzysów wyznaczonych przez NBER dla ustabilizowanej gospodarki rynkowej Stanów Zjednoczonych oraz faz kryzysów wyznaczonych na podstawie wartości indeksów (rysunek 4). Zauważmy, że fazy wyznaczone dwoma różnymi metodami prawie dokładnie się pokrywają.

## 2. Wybór modelu dyskretnego

Gdy interesuje nas określenie prawdopodobieństwa wystąpienia tylko jednego z możliwych stanów aktywności gospodarczej, można wykorzystać model dwumianowy. Tego rodzaju podejście, w odniesieniu do fazy kryzysu, zastosowali cytowani na wstępie artykułu A. Estrella i F.S. Mishkin.

W dwumianowych modelach jakościowych, gdzie zmienna objaśniana jest z reguły zmienną zero-jedynkową, tzn.

$$y_t = \begin{cases} 1, & \text{gdy dane zdarzenie zachodzi,} \\ 0, & \text{gdy dane zdarzenie nie zachodzi,} \end{cases}$$

prawdopodobieństwo teoretyczne $P_{t1}$ wystąpienia w momencie $t$ pierwszej możliwości określone jest przez dystrybuantę rozkładu wartości $(\boldsymbol{\beta}'\mathbf{x}_t)$

$$P_{t1} = F(\boldsymbol{\beta}'\mathbf{x}_t), \tag{10}$$

gdzie $\mathbf{x}_t$ oznacza wektor wartości wskaźników wyprzedzających, $\boldsymbol{\beta}$ – wektor parametrów.

Rozkład losowej zmiennej objaśnianej generowany jest przez rozkład składnika losowego, zmienne objaśniające bowiem traktujemy jako zmienne z góry ustalone. Stąd

$$P_{t1} = F(\boldsymbol{\beta}' \mathbf{x}_t) \int_{-\infty}^{\boldsymbol{\beta}' \mathbf{x}_t} f(\xi_t) d\xi_t, \tag{11}$$

gdzie $f(\xi_t)$ to funkcja gęstości o znanych parametrach zmiennej $\xi_t$. Przyjęty typ rozkładu dla zmiennej losowej $\xi_t$ determinuje typ modelu.

W modelu probitowym zakłada się, że prawdopodobieństwo $P_{t1}$ określone jest za pomocą dystrybuanty standaryzowanego rozkładu normalnego $\xi_t \sim N(0, 1)$:

$$P_{t1} = F(\boldsymbol{\beta}' \mathbf{x}_t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boldsymbol{\beta}' \mathbf{x}_t} e^{-\xi_t^2/2} d\xi_t, \tag{12}$$

natomiast w modelu logitowym, że za pomocą dystrybuanty standaryzowanego rozkładu logistycznego $\xi_t \sim L(0, \pi^2/3)$

$$P_{t1} = F(\boldsymbol{\beta}' \mathbf{x}_t) = \int_{-\infty}^{\boldsymbol{\beta}' \mathbf{x}_t} \frac{e^{-\xi_t}}{(1+e^{-\xi_t})^2} d\xi_t = \frac{1}{1+e^{-\boldsymbol{\beta}' \mathbf{x}_t}}. \tag{13}$$

Jeśli istnieje wybór między modelami opisującymi kształtowanie się zmiennej jakościowej, to powstaje pytanie, czym kierować się w zastosowaniach praktycznych. Podstawowym kryterium wyboru powinny być względy merytoryczne, na przykład wiedza na temat dystrybuanty $F(\xi_t)$. Wiedzę na temat badanego zjawiska możemy czerpać z teorii, wcześniejszych badań empirycznych, analizy dostępnego materiału statystycznego. Bardzo często jednak nie dysponujemy dostateczną wiedzą, a materiał statystyczny jest albo ubogi, albo niepełny. Stawiamy pewne hipotezy co do modelu, a następnie poddajemy go weryfikacji przy wykorzystaniu wybranych mierników dopasowania lub – pośrednio – poprzez weryfikację hipotez o istotności parametrów itp.

W przypadku modeli wyboru dyskretnego najczęściej stosowane są modele logitowe. W pracach porównawczych podkreśla się niewielkie różnice teoretycznych wartości prawdopodobieństw uzyskanych z różnych modeli[4]. Amemiya w swojej pracy (1981) podaje, że pomiędzy parametrami $\beta$ w modelu logitowym i probitowym zachodzi relacja:

[4] Zob. K. Jajuga, *Modele z dyskretną zmienną objaśnianą*, w: *Estymacja modeli ekonometrycznych*, red. S. Bartosiewicz, PWE, Warszawa 1990.

$$\beta_{logit} = 1{,}6\ \beta_{probit}, \tag{14}$$

co pozwala przeliczyć wyniki estymacji z jednego modelu na drugi. Warto dodać, że rozróżnienie pomiędzy rozkładem logistycznym i rozkładem normalnym jest empirycznie trudne[5]. Za wyborem w badaniach empirycznych modeli logitowych przemawiają również ich własności interpretacyjne ([12], s. 96, oraz [15], rozdz. 3 i 5). Dalszy przegląd możliwości wyboru modelu dyskretnego dla prognozowania faz aktywności gospodarczej ograniczono do modeli logitowych. Przeprowadzona klasyfikacja wyników obserwacji na 4 fazy aktywności gospodarczej sugeruje wybór spośród modeli wielomianowych.

*Modele wielomianowe* mogą być *uporządkowane* lub *nieuporządkowane*, w zależności od tego w jakiej skali mierzona jest zmienna objaśniana (odpowiednio nominalnej lub porządkowej). Dwa możliwe podejścia do modeli wielomianowych związane są z interpretacją zmiennej objaśnianej.

Przykładowo ankiety testu koniunkturalnego zawierają różne warianty odpowiedzi na zadawane pytania. Wśród barier działalności produkcyjnej (przykład A) wymieniane są m.in.

- niedostateczny popyt krajowy,
- brak siły roboczej,
- niedobory surowców,
- braki energii,
- niejasne prawo gospodarcze.

Natomiast w przypadku oceny bieżącej sytuacji gospodarczej (przykład B) respondent ma do wyboru odpowiedź:

- dobra,
- zadowalająca,
- zła.

Ocena sytuacji gospodarczej mierzona jest w skali porządkowej, od najlepszej do najgorszej. Odpowiedzi można porangować, natomiast nie są znane odległości między nimi. W przypadku barier działalności gospodarczej ich kolejność nie ma żadnego znaczenia, co odpowiada skali nominalnej, nieuporządkowanej. Modele dla zmiennej nominalnej (nieuporządkowanej) nazywane są klasycznymi modelami logitowymi i są bezpośrednim uogólnieniem dwumianowych modeli logitowych.

Różnica w podejściu do obu sytuacji decyzyjnych polega na wykorzystaniu teorii użyteczności. Z punktu widzenia tej teorii podstawowym kryterium wyboru decyzji jest maksymalizacja użyteczności decydenta (oczekiwanej

[5] Kenneth E.Train, *Discrete Choice Methods with Simulation*, University of Cambridge 2003, s. 39.

użyteczności w przypadku działania w warunkach niepewności) przy założeniu, że decydent działa racjonalnie ([13] s. 176 i n., zob. aksjomaty racjonalnego wyboru i twierdzenie o funkcji użyteczności). Wartości funkcji użyteczności wykorzystywane są do uporządkowania zbioru możliwości, nie mają one konkretnej interpretacji ekonomicznej, każdy z decydentów może mieć swoją funkcję użyteczności. Oba przykłady (A i B) dotyczą procesu decyzyjnego jednostki. W przypadku analiz makroekonomicznych trudno rozpatrywać funkcję użyteczności dla gospodarki. Można jednak przyjąć, że rolę funkcji użyteczności spełnia *funkcja symptomatyczna* aktywności gospodarczej. Prawdopodobieństwo wystąpienia danego stanu aktywności gospodarczej rozumiane jest wtedy jako prawdopodobieństwo zdarzenia, że moc symptomatyczna wskaźników do przewidywania tego stanu jest wyższa od mocy symptomatycznej wskaźników do przewidywania każdego innego stanu aktywności gospodarczej.

## 2.1. Uporządkowane wielomianowe modele logitowe

W przypadku uporządkowanych modeli logitowych zakłada się, że istnieje jedna „idealna" funkcja użyteczności. Funkcję użyteczności $U_t$ sytuacji w kraju, z punktu widzenia respondenta $t$ (przykład B), można zdekomponować na część deterministyczną $V_t$, określoną przez obserwowalne wartości wektora $\mathbf{x}_t$ oraz na część stochastyczną, zależną od wskaźników nieobserwowanych, reprezentowaną przez $\varepsilon_t$. Wektor $\mathbf{x}_t$ jest wektorem obserwowanych charakterystyk (zmiennych), na podstawie których respondent dokonuje wyboru. Tak więc

$$U_t = V_t + \varepsilon_t, \tag{15}$$

przy czym

$$V_t = \boldsymbol{\beta}' \mathbf{x}_t. \tag{16}$$

Przedmiotem modelowania jest pewna ukryta zmienna (ang. *latent variable*), której wartości nie są w pełni obserwowane. Ocena sytuacji w kraju zależy od przyjętych przez respondentów kolejnych punktów granicznych. W przykładzie B należałoby oszacować dwa punkty graniczne $k_1$ i $k_2$ ($k_1 < k_2$), przy czym

$$P(\text{sytuacja zła}) = P(U_t < k_1) = P(\varepsilon_t < k_1 - \boldsymbol{\beta}' \mathbf{x}_t) = \frac{e^{k_1 - \boldsymbol{\beta}' \mathbf{x}_t}}{1 + e^{k_1 - \boldsymbol{\beta}' \mathbf{x}_t}},$$

$$P(\text{sytuacja zadowalająca}) = P(k_1 < U_t < k_2) = \frac{e^{k_2 - \boldsymbol{\beta}' \mathbf{x}_t}}{1 + e^{k_2 - \boldsymbol{\beta}' \mathbf{x}_t}} - \frac{e^{k_1 - \boldsymbol{\beta}' \mathbf{x}_t}}{1 + e^{k_1 - \boldsymbol{\beta}' \mathbf{x}_t}},$$

natomiast

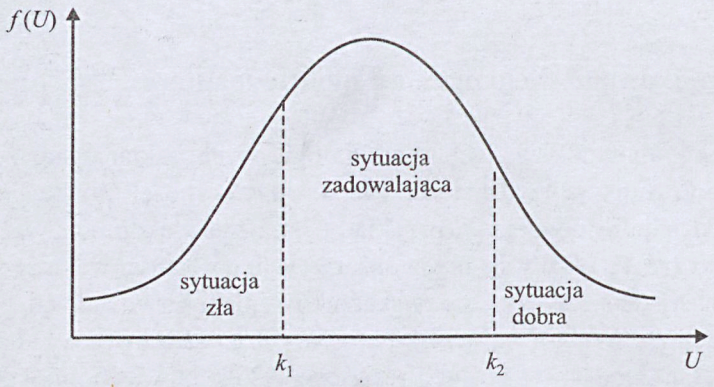
$$P(\text{sytuacja dobra}) = P(U_t > k_2) = 1 - \frac{e^{k_2 - \beta' x_t}}{1 + e^{k_2 - \beta' x_t}}.$$

Prawdopodobieństwo wyboru (wystąpienia) kategorii $j$ ($j = 1, \ldots, J$) w uporządkowanym modelu logitowym zostało przedstwione jako

$$P(y_t = j \mid \mathbf{x}_t, \boldsymbol{\beta}, \boldsymbol{\kappa}) = F(\kappa_r - \boldsymbol{\beta}' \mathbf{x}_t) - F(\kappa_{r-1} - \boldsymbol{\beta}' \mathbf{x}_t), \tag{17}$$

gdzie $\boldsymbol{\kappa}$ oznacza wektor punktów granicznych, natomiast $F$, jak poprzednio, dystrybuantę rozkładu logistycznego.

Rysunek 5 ilustruje sytuację przedstawioną w przykładzie B. Rozkład użyteczności $U_t$ wokół wartości $\boldsymbol{\beta}' \mathbf{x}_t$ jest rozkładem składnika losowego $\varepsilon_t$.




**Rys. 5. Rozkład opinii na temat sytuacji w kraju**

Przedstawione rozważania sugerują wybór modelu ze względu na uporządkowanie stanów aktywności gospodarki. Warto jednak zauważyć, że konsekwencją takiego wyboru jest założenie o jednakowym wpływie wskaźnika wyprzedzającego (np. stopy procentowej) na zmianę aktywności gospodarczej dla każdego z wyróżnionych stanów. Czy słusznie? Uporządkowanie kategorii od najlepszej do najgorszej (lub odwrotnie) w przypadku gospodarki jest procesem o wiele bardziej złożonym niż w przypadku jednostki.

W literaturze autorzy zwracają uwagę, że w sytuacjach typu B można szacować wielomianowy model logitowy dla kategorii nieuporządkowanych (nie byłoby to błędem), ale model uporządkowany jest w pewnych przypadkach praktycznie lepszy i wiąże się z szacowaniem mniejszej liczby parametrów ([15], s. 164).

## 2.2. Nieuporządkowane wielomianowe modele logitowe

Podstawy teoretyczne dla rozwoju badań nad modelami dyskretnymi stworzył D.L. McFadden, który na gruncie teorii użyteczności wyjaśnił, co oznacza przyjęcie założenia o typie rozkładu zmiennej $\xi_t$.

Załóżmy tym razem, że z każdą możliwością, stanem aktywności $j\,(j=1,\ldots,J)$ związana jest inna funkcja użyteczności

$$U_{tj}=V_{tj}+\varepsilon_{tj}, \quad \text{gdzie} \quad V_{tj}=\boldsymbol{\beta}'_{(j)}\mathbf{x}_{tj}. \tag{18}$$

Symbol $\boldsymbol{\beta}_{(j)}$ oznacza wektor parametrów kategorii $j$ (nie należy go mylić z oznaczeniem pojedynczego parametru $\beta_j$). Dla uproszczenia zapisu, w dalszych rozważaniach pominięto indeks $t$, odnoszący wybór kategorii $j$ do określonego respondenta $t$. Odtąd określenie wielomianowe modele logitowe, jeśli tego nie zaznaczono inaczej, będzie odnosiło się do kategorii nieuporządkowanych.

Zgodnie z teorią racjonalnych wyborów zakładamy (przykład A), że respondent $t$ wybiera kategorię $j$ tylko w przypadku, gdy użyteczność tej kategorii jest wyższa od użyteczności wszystkich innych kategorii. Stąd

$$\begin{aligned} P_j&=P(U_j\geqslant U_m)=P(V_j+\varepsilon_j\geqslant V_m+\varepsilon_m)= \\ &=P(\varepsilon_m-\varepsilon_j\leqslant V_j-V_m)\ \forall m\neq j. \end{aligned} \tag{19}$$

Prawdopodobieństwo wyboru kategorii $j$ jest wyrażone przez dystrybuantę rozkładu różnicy dwóch zmiennych losowych.

Rozkład logistyczny dla zmiennej losowej $\xi$ otrzymamy przy założeniu, że zmienne losowe $\varepsilon_j$ i $\varepsilon_m$ $(j\neq m)$ są niezależne i mają rozkład wartości ekstremalnych I typu, lub – w szczególności – Weibulla ([7], s. 61 i n.).

Przy założeniu że składniki losowe mają rozkład normalny, ich różnica ma też rozkład normalny i otrzymujemy model probitowy[6]. Dla modelu probitowego nie jest wymagane założenie o niezależności składników losowych.

Przyjmijmy, że $F(\varepsilon_1,\ldots,\varepsilon_J)$ określa dystrybuantę wielowymiarowej zmiennej losowej $(\varepsilon_1,\ldots,\varepsilon_J)$. Jeżeli $F$ jest funkcją ciągłą i ma pochodne cząstkowe $F_j$ względem $\varepsilon_j$, to zależność (19) można zapisać jako

$$P_j=\int_{-\infty}^{\infty} F_j(\varepsilon+V_j-V_1,\ldots,\varepsilon+V_j-V_J)d\varepsilon. \tag{20}$$

[6] Zob. *Handbook of econometrics* (vol. 3) eds Z. Griliches, M.D. Intriligator, chapter 27, P.J. Dhrymes, *Limited dependent variables*, s. 1573, Elsevier Science Publishers, BV, 1986.

Wyrażenie podcałkowe oznacza rozkład brzegowy zmiennej $\varepsilon_j$. Dla dwóch niezależnych zmiennych losowych $\varepsilon_1$ i $\varepsilon_2$ o identycznych rozkładach otrzymujemy

$$P_1 = \int_{-\infty}^{\infty} F_1(\varepsilon + V_1 - V_1,\ \varepsilon + V_1 - V_2)d\varepsilon = \int_{-\infty}^{\infty} F(\varepsilon + V_1 - V_2) \cdot F'(\varepsilon)d\varepsilon.$$

W przykładzie A wybór opisywany byłby przez $J=5$ różnych funkcji użyteczności. Model logitowy otrzymamy przy założeniu, że wszystkie zmienne losowe $\varepsilon_1, \ldots, \varepsilon_5$ mają identyczne niezależne rozkłady wartości ekstremalnych o funkcji gęstości

$$f(\varepsilon) = \mu e^{-\mu(\varepsilon-\eta)} \exp(-e^{-\mu(\varepsilon-\eta)}), \qquad \mu > 0, \tag{21}$$

i dystrybuancie

$$F(\varepsilon) = \exp(-e^{-\mu(\varepsilon-\eta)}), \tag{22}$$

gdzie $\eta$ jest parametrem pozycji, $\mu$ natomiast dodatnim parametrem skali. Wartość oczekiwana zmiennej $\varepsilon$ jest równa

$$E(\varepsilon) = \eta + \frac{\gamma}{\mu}, \tag{23}$$

przy czym

$$\gamma = \lim_{k \to \infty} \sum_{i=1}^{k} \frac{1}{i} - \ln(k) \cong 0{,}5772 \tag{24}$$

jest tzw. stałą Eulera. Wariancja zmiennej $\varepsilon$ jest równa

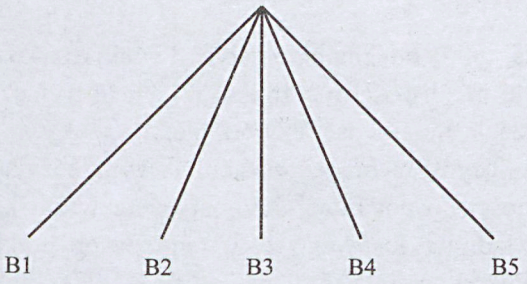
$$\mathrm{Var}(\varepsilon) = \frac{\pi^2}{6\mu^2}. \tag{25}$$

Jeśli założymy, że składniki losowe $\varepsilon_1, \ldots, \varepsilon_J$ mają identyczne i niezależne rozkłady, to wszystkie błędy będą miały tę samą wariancję. W wielomianowym modelu logitowym, najczęściej arbitralnie przyjmuje się, że $\mu = 1$, $\eta = 0$.

Prawdopodobieństwo wyboru $i$-tej bariery działalności produkcyjnej, z pięciu różnych możliwości (przykład A) można zapisać jako:

$$P_i = \frac{\exp(V_i)}{\sum_{j=1}^{5} V_j}. \tag{26}$$

Schemat wyboru jednego z pięciu wariantów pokazano na rysunku 6.




**Rys. 6. Schemat wyboru w wielomianowym modelu logitowym**

Omawiany model składa się z pięciu powiązanych ze sobą równań. Identyfikacja układu równań wymaga, aby $\beta_{(1)}=0$. Oznacza to, że pierwsza możliwość została uznana za bazową, a pozostałe cztery wektory parametrów $\beta_{(j)}$ mierzą stosunek szans zaistnienia możliwości $j$ w grupie zdarzeń, zawierających wyłącznie daną możliwość oraz możliwość bazową. Szacowane są parametry modelu:

$$\ln\frac{P_j}{P_1}=V_j-V_1. \tag{27}$$

Wartości teoretyczne prawdopodobieństw wyznaczane są na podstawie równań:

$$P_1=\frac{1}{1+\sum_{j=2}^{5}\exp(V_j)}, \tag{28}$$

$$P_i=\frac{\exp(V_i)}{1+\sum_{j=2}^{5}\exp(V_j)} \quad \text{dla} \quad i=2,\ldots,5. \tag{29}$$

Podstawowym ograniczeniem standardowych wielomianowych modeli logitowych jest aksjomat niezależności od warunków pobocznych (ang. *independent of irrelevant alternatives*, IIA)[7]. Własność wynika z relacji prawdopodobieństw

[7] W przypadku modeli probitowych zakładamy, że składniki losowe mają łączny rozkład normalny: $\varepsilon_t=(\varepsilon_{t1},\ldots,\varepsilon_{tJ})=N(0,\boldsymbol{\Omega})$. Macierz kowariancji $\boldsymbol{\Omega}$ pozwala uwzględnić zarówno heterogeniczność, jak i korelację składników losowych. Jeśli istnieje taka potrzeba, to można rozszerzyć macierz kowariancji tak, aby ująć korelację składnika losowego w czasie. Pozostaje jednak problem symulacji parametrów rozkładu i konieczność nałożenia warunków na macierz $\boldsymbol{\Omega}$.

$$\frac{P_j}{P_k}=\frac{\exp(V_j)}{\exp(V_k)}, \tag{30}$$

i oznacza zależność tylko od charakterystyk dwóch rozważanych kategorii. Pojawienie się nowej możliwości nie zmieni relacji (30), niezależnie od tego, czy istnieje inna, trzecia możliwość, czy też nie.

W modelu testującym wybrane wskaźniki wyprzedzające w celu prognozowania aktywności gospodarczej automatycznie wszystkie inne czynniki włączane są do składnika losowego i w ten sposób stają się przyczyną skorelowania składników losowych poprzez różne kategorie. Stąd z jednej strony standardowy model logitowy jest słabym odzwierciedleniem rzeczywistych warunków, z drugiej jednak umożliwia ujęcie zróżnicowania wpływu czynników w różnych fazach aktywności.

Intensywne prace nad modelami logitowymi wynikały z potrzeby dostosowania ich do realnych warunków. Złagodzenie krępującego założenia IIA (wzór (30)) jest obecnie możliwe poprzez zastosowanie różnych postaci uogólnionych modeli wartości ekstremalnych (GEV), w tym również zagnieżdżonych modeli logitowych.

## 2.3. Zagnieżdżone modele logitowe

Zagnieżdżony model logitowy, po raz pierwszy zaproponowany przez Ben-Akiva w 1973 roku, jest uogólnieniem wielomianowego modelu logitowego, uwzględniającym korelację między kategoriami ([4], chapter 2). Zakładamy, że istnieje $M$ wyczerpujących i wzajemnie wykluczających się podzbiorów możliwości wyboru $C_m$ $(m=1,\ldots,M)$, to znaczy kategorie $(j=1,\ldots,J)$ można podzielić między tzw. gniazda tak, aby:

- własność IIA obowiązywała dla możliwości w każdym gnieździe,
- własność IIA nie obowiązywała dla możliwości z dwóch różnych gniazd.

Zakładamy, że użyteczność danej kategorii $j$ $(j\in C_m)$ można zapisać jako:

$$U_j=V_j+\varepsilon_j=\tilde{V}_j+\tilde{\varepsilon}_j+\tilde{V}_{C_m}+\tilde{\varepsilon}_{C_m}, \tag{31}$$

gdzie:

$\tilde{V}_{C_m}$ – zależy od atrybutów gniazda $C_m$, część użyteczności wspólna dla wszystkich kategorii gniazda,

$\tilde{V}_j$ – zależy od atrybutów danej kategorii $j\in C_m$.

Składniki losowe $\tilde{\varepsilon}_j$ i $\tilde{\varepsilon}_{C_m}$ są niezależne, przy czym $\tilde{\varepsilon}_j$ dla $j \in C_m$ – tak jak w wielomianowym modelu logitowym – ma rozkład (21) z parametrem skali $\mu_m$. Rozkład składnika losowego $\tilde{\varepsilon}_{C_m}$ jest taki jak zmiennej $\max_{j \in C_m} U_j$ z parametrem skali $\mu$.

Uchylamy założenie o niezależności składników losowych w gniazdach, tzn. dla dowolnych możliwości $i, j \in C_m$, współczynnik korelacji

$$r(\varepsilon_i, \varepsilon_j) = 1 - \frac{\mu^2}{\mu_m^2}, \tag{32}$$

natomiast jeśli $i \in C_m$, $j \in C_k$ $(m \neq k)$, to $r(\varepsilon_i, \varepsilon_j) = 0$.

Z definicji (32) wynika, że korelacja jest nieujemna i dlatego

$$0 \leqslant \frac{\mu}{\mu_m} \leqslant 1. \tag{33}$$

Jeśli $\mu/\mu_m = 1$, to korelacja jest zerowa.

Każde gniazdo $C_m$ związane jest z tzw. pseudoużytecznością, którą można zapisać jako kompozycję użyteczności

$$V'_{C_m} = \tilde{V}_{C_m} + \frac{1}{\mu_m} \ln \sum_{j \in C_m} e^{\mu_m \tilde{V}_j}, \tag{34}$$

gdzie

$$\ln \sum_{j \in C_m} e^{\mu_m \tilde{V}_j} = I_m = E(\max_{j \in C_m} U_j) = E[\max_{j \in C_m} (\tilde{V}_j + \varepsilon_j)], \tag{35}$$

jest wartością oczekiwaną maksymalnej użyteczności z gniazda $C_m$ (ang. *inclusive value* lub *log-term value*).

Przyjmując, że parametr skali $\mu = 1$ (tzw. normalizacja z góry) oraz

$$\frac{1}{\mu_m} = \lambda_m, \tag{36}$$

parametr $\lambda_m$ można interpretować jako miarę stopnia niezależności składników losowych funkcji użyteczności zakwalifikowanych do gniazda $C_m$, natomiast $1 - \lambda_m$ jako miarę korelacji tych składników. Jeśli $\lambda_m = 1$, to składniki losowe są niezależne i otrzymuje się standardowy model logitowy. Im wartość $\lambda_m$ jest bliższa zeru, tym składniki losowe są bardziej skorelowane. W badaniach

empirycznych można przyjmować założenie, że parametr $\lambda_m$ jest różny lub taki sam dla różnych gniazd. Jeśli oszacowanie $\lambda_m \in [0, 1]$ dla $m = 1, ..., M$, to model jest zgodny z zasadą maksymalizacji użyteczności ([15], s. 85). W przypadku gdy $\lambda_m > 1$, należy podejrzewać, że model został źle wyspecyfikowany lub jest zgodny z zasadą maksymalnej użyteczności tylko dla pewnego zakresu wartości zmiennych objaśniających, ale nie dla wszystkich. Ujemną wartość $\lambda_m$ również należy interpretować jako niezgodność z zasadą maksymalizacji użyteczności (wzrost użyteczności zmniejsza prawdopodobieństwo wyboru).

W przypadku *dwupoziomowego zagnieżdżonego modelu logitowego*[8] prawdopodobieństwo wystąpienia kategorii $j$ ($j$-tego stanu aktywności) $P_j$ można zapisać jako iloczyn prawdopodobieństwa (brzegowego poprzez możliwości $P_{C_m}$) wyboru podzbioru $C_m$, do którego należy kategoria (stan) $j$, oraz prawdopodobieństwa (warunkowego $P_{j|C_m}$) wystąpienia stanu $j$ z tego podzbioru

$$P_j = P_{j|C_m} P_{C_m}, \tag{37}$$

gdzie

$$P_{C_m} = \frac{\exp(\tilde{V}_{C_m} + \lambda_m I_m)}{\sum_{l=1}^{M} \exp(\tilde{V}_{C_l} + \lambda_l I_l)}, \tag{38}$$

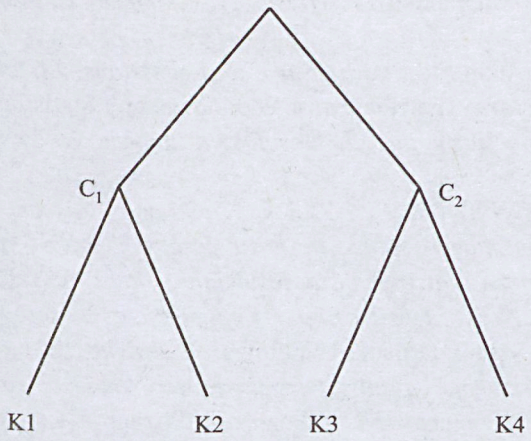
$$P_{j|C_m} = \frac{\exp\left(\frac{\tilde{V}_j}{\lambda_m}\right)}{\sum_{j \in C_m} \exp\left(\frac{\tilde{V}_j}{\lambda_m}\right)}, \tag{39}$$

oraz

$$I_m = \ln \sum_{j \in C_m} \exp\left(\frac{\tilde{V}_j}{\lambda_m}\right). \tag{40}$$

Zagnieżdżone modele logitowe mają strukturę drzewa. Schematyczny wybór można przedstawić tak jak na rysunku 7. Jest on zgodny z przeprowadzoną wcześniej klasyfikacją obserwacji do różnych faz aktywności gospodarczej (rysunek 1). Wprowadzenie do modelu parametru $\lambda_m$ pozwala

[8] Przedstawiony zagnieżdżony model logitowy można wyprowadzić na podstawie uogólnionej dystrybuanty wartości ekstremalnych, zob. [15], s. 99.




**Rys. 7. Schemat wyboru w zagnieżdżonym modelu logitowym**

zweryfikować hipotezę o niezależności składników losowych. Hierarchiczna struktura modelu umożliwia testowanie dwóch grup wskaźników, na przykład indeksów rocznych i miesięcznych, wskaźników krajowych i zagranicznych.

Przedstawione teoretyczne rozważania umożliwiają analizę procesu prognozowania aktywności gospodarczej na gruncie przyjmowanych założeń oraz przedstawiają możliwości zastosowania modeli wyboru dyskretnego nie tylko w mikro-, ale i w makroekonomii.

## Literatura

[1] Aczel A.D., *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa 2000

[2] Barczyk R., Kowalczyk Z., *Metody badania koniunktury gospodarczej*, Wydawnictwo Naukowe PWN, Warszawa-Poznań 1993

[3] Bartosiewicz S., *Estymacja modeli ekonometrycznych*, PWE, Warszawa 1990

[4] Ben-Akiva M., Bierlaire M., *Discrete Choice Methods and Their Applications*, w: *Handbook of Transportation Science*, ed. R. Hall, International Series in Operations Research and Management Science vol. 23, 1999

[5] Chow G.C., *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa 1995

[6] Dhrymes P.J., *Limited Dependent Variables*, w: *Handbook of Econometrics* (vol. 3) ed. Z. Griliches, M.D. Intriligator, Elsevier Science Publishers, BV, 1986

[7] Domencich T., McFadden D., *Urban Travel Demand*, North-Holland, Amsterdam 1975
[8] Estrella A., Mishkin F.S., *Predicting U.S. Recessions: Financial Variables as Leading Indicators*, The Review of Economics and Statistics 1998, 80, 45-61
[9] Gatnar E., *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa 1998
[10] Grabiński T., Wydymus S., Zeliaś A., *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa 1989
[11] Greń J., *Statystyka matematyczna. Modele i zadania*, PWN, Warszawa 1982
[12] Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Szkoła Główna Handlowa, Warszawa 2001
[13] Jurek W., *Konstrukcja i analiza portfela papierów wartościowych o zmiennym dochodzie*, Wydawnictwo AE w Poznaniu, Poznań 2001
[14] McFadden D.L., *Econometric Analysis of Qualitative Response Models*, w: *Handbook of Econometrics* (vol. 3) eds Z. Griliches, M.D. Intriligator, Elsevier Science Publishers, BV, 1986
[15] Train K.E., *Discrete Choice Methods with Simulation*, University of Cambridge 2003

## SOME REMARKS ON THE SUBJECT OF CHOICE DISCRETE MODEL TO FORECASTING OF POLISH ECONOMIC ACTIVITY

**Summary**

This paper studies the possibilities of the applications qualitative (discrete choice) models in forecasting of economic activity. The brief review of different logit models such as simple binary model, multinomial ordered and unordered model and nested logit model are introduced on the basis of their theoretical assumptions. For many cases economists may be interested in estimate not only the probability of recession, but also probability of economic situations after or before recession. For that reason this paper proposes definition a dependent variable in the multinomial model. The nested logit model is an extension of the multinomial logit model designed to capture some correlation among alternatives (reflect by parameter $\lambda$). The result is that the structure nested logit model is the same as in the method of definition dependent variable. On the ground of the investigation of the theoretical and practical aspects of qualitative models it is shown how to choose the suitable model to forecasting of economic activity.