POLSKA AKADEMIA NAUK
KOMITET STATYSTYKI I EKONOMETRII

PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

TOM 59
1
2012

WARSZAWA 2012

WYDAWCA

Komitet Statystyki i Ekonometrii Polskiej Akademii Nauk

Wydanie publikacji dofinansowane przez Ministra Nauki i Szkolnictwa Wyższego

Strona WWW „Przegląd Statystyczny”:
http://www.przegladstatystyczny.pan.pl

Nakład: 250 egz.

PAN Warszawska Drukarnia Naukowa
00-056 Warszawa, ul. Śniadeckich 8
Tel./fax:+48 22 628 76 14

Od Redakcji

Szanowni Państwo,

Komitet Statystyki i Ekonometrii PAN, z inicjatywy Przewodniczącego Komitetu prof. dr. hab. Jacka Osiewalskiego, powołał nowy skład Komitetu Redakcyjnego „Przeglądu Statystycznego" oraz nową Radę Redakcyjną kierowaną przez prof. dr. hab. Władysława Milo.
Skład Komitetu Redakcyjnego jest następujący:

Magdalena Osińska – Redaktor Naczelny,

Marek Walesiak – Zastępca Redaktora Naczelnego,

Piotr Fiszeder – Sekretarz Naukowy Redakcji,

Anna Pajor – Redaktor,

Michał Majsterek – Redaktor.

Na ręce dotychczasowego Redaktora Naczelnego Pana Profesora Mirosława Szredera kierujemy słowa podziękowania za wysiłek i zaangażowanie włożone w pracę redakcyjną, która zaowocowała utrzymaniem wysokiego poziomu naukowego czasopisma. W składzie Komitetu Redakcyjnego pracowali także: Magdalena Osińska – Zastępca Redaktora Naczelnego, Marek Walesiak – Zastępca Redaktora Naczelnego oraz Krzysztof Najman – Sekretarz Naukowy Redakcji.

Dziękujemy wszystkim Członkom Rady Redakcyjnej, którzy wspierali pismo swoją obecnością i zaangażowaniem. Sylwetkę zmarłego niedawno Przewodniczącego Rady - Pana Profesora Michała Kolupy przedstawiamy w niniejszym numerze „Przeglądu Statystycznego".

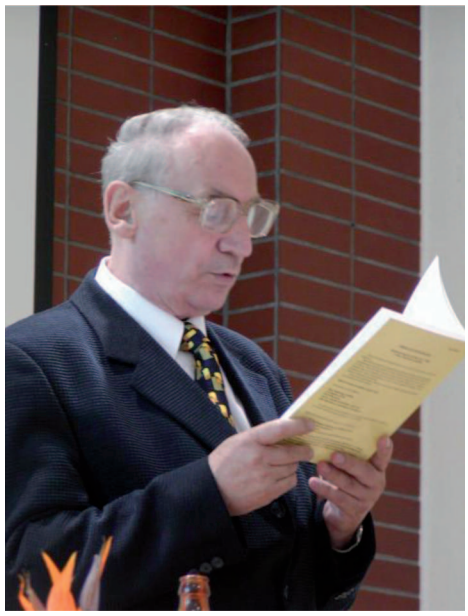
Czasopismo w swojej blisko 60-letniej historii zawsze było najważniejszym w Polsce miejscem wymiany myśli naukowej grona pracowników naukowych, zajmujących się szeroko pojętymi metodami ilościowymi, w tym szczególnie statystyką i ekonometrią. Szanując dotychczasowe osiągnięcia czasopisma, nowy Komitet Redakcyjny będzie czynił starania o podniesienie poziomu naukowego, o unowocześnienie pisma poprzez elektroniczne środki przekazu informacyjnego (m.in. elektroniczny kontakt z autorami i recenzentami, regularną aktualizację oraz rozszerzenie zakresu informacji strony internetowej, wprowadzenie do większej liczby baz czasopism), a także o nadanie czasopismu wyższej rangi w ocenie parametrycznej.

Serdecznie zapraszamy do publikowania swoich osiągnięć naukowych na łamach czasopisma „Przegląd Statystyczny" zgodnie z jego profilem.

Prace zgłaszane do publikacji w „Przeglądzie Statystycznym", spełniające wszystkie wymogi formalne (zob. wymogi zawarte na stronie http://www.przegladstatystyczny.pan.pl/), należy przesyłać pocztą elektroniczną na adres Sekretarza Naukowego Redakcji Piotra Fiszedera: piotr.fiszeder@umk.pl.

Redakcja

PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012



Prof. zw. dr hab. Michał Maria Kolupa
12.06.1936 – 11.02.2012

Zdjęcie pochodzi ze zbiorów autora

Wspomnienie profesora Michała KOLUPY

Cisną się słowa za Janem Kochanowskim – Profesorze, Michale – wielkie pustki uczyniłeś w naszym domu – środowisku polskich ekonometryków, Twoich Uczniów, Współpracowników i Przyjaciół.

Profesora Michała Kolupę poznałem w roku 1964 i od pierwszego spotkania zawiązała się między nami prawdziwa przyjaźń. Profesor był wtedy młodym doktorem a ja młodym magistrem. Jednoczyła nas wspólnota myśli o nauce i świecie.

Profesor Michał Kolupa ukończył w roku 1954 znakomite LVI Liceum Ogólnokształcące w Ursusie a Jego nauczyciel matematyki Józef Chmiel odkrył Go dla matematyki i w konsekwencji podjął studia na Wydziale Matematyki i Fizyki Uniwer-

sytetu Warszawskiego. Magisterium otrzymał w roku 1958 a promotorem Jego pracy magisterskiej był wybitny statystyk matematyczny Profesor Marek Fisz. Pierwszym mentorem Profesora był Profesor Marek Stark ówczesny kierownik Redakcji Matematyki Państwowego Wydawnictwa Naukowego. W tym czasie ewenementem w nauce ekonomii była publikacja książki Oskara Langego – *Wstęp do ekonometrii*. Profesor Marek Stark zachęcił młodego magistra by zajął się ekonometrią, nieznaną wówczas w Polsce. Profesor Kolupa najpierw związał się Instytutem Handlu wewnętrznego i jak mi mówił, odkrył Przegląd Statystyczny i artykuł Zbigniewa Pawłowskiego – *Problem modeli ekonometrycznych o jednym równaniu w badaniach elastyczności popytu w Polsce*.

O Profesorze Z. Pawłowskim Profesor M. Kolupa napisał: „Profesor Z. Pawłowski miał zawsze dla mnie dobre słowo zachęty, otaczał mnie opieką merytoryczną, za co byłem i jestem Mu serdecznie wdzięczny, bo gdyby nie ta pomoc i Jego troska pewno mój start naukowy znacznie by się wydłużył". Profesor M. Kolupa napisał także: "śledziłem dorobek prof. Z. Hellwiga, co pozwoliło mi napisać własne prace traktujące o tematyce zainicjowanej przez prof. Z. Hellwiga a dotyczące koincydencji, efektu katalizy i związku między tymi pojęciami. . . co doprowadziło mnie do tytułu naukowego profesora nadzwyczajnego (1983)". Tak więc Profesor Michał Kolupa napisał iż „wszystko co mnie spotkało w nauce zawdzięczam swoim mistrzom – prof. Z. Pawłowskiemu i prof. Z. Hellwigowi".

Profesor M. Kolupa miał zawsze bardzo silne więzy ze Szkołą Główną Planowania Statystyki (obecnie Szkołą Główną Handlową). Pracę etatową podjął Profesor w SGPiS w roku 1961, aby na tej Uczelni uzyskać stopień doktora w roku 1964 (promotorem był profesor Wiesław Sadowski). W roku 1969 nastąpiła przerwa we współpracy z SGPiS, gdyż Profesor podjął pracę w Wojskowej Akademii Politycznej, aby w roku 1972 powrócić do SGPiS, gdzie pracował już do emerytury, na którą przeszedł w roku 2005. W czasie swej pracy na SGPiS Profesor Kolupa pełnił ważne funkcji – Prorektora, Dziekana, Prodziekana, Kierownika Zakładu Ekonometrii, Dyrektora Instytutu Ekonometrii, Kierownika Zakładu Zarządzania Ryzykiem. Po przejściu na emeryturę Profesor pracował w warszawskich uczelniach niepublicznych i tak: w Wyższej Szkole Zarządzania i Marketingu (1993–1995), Wyższej Szkole Handlu i Prawa (1997–2004), Wyższej Szkole Przedsiębiorczości i Zarządzania (1995–1998) oraz także na Politechnice Radomskiej w latach 1986–1994 oraz 2005–2012).

Pierwszy artykuł Profesora to – *Przyczynek do zagadnienia obciążenia współczynników elastyczności popytu wynikającego z niedostatecznej podaży, Przegląd Statystyczny nr 4/1960.* Pierwsze spotkanie z Profesorem Z. Pawłowskim nastąpiło w 1960 w „Jajku" (klubie pracowniczym) SGPiS. Wielkim bodźcem do pracy naukowej Profesora M. Kolupy było cytowane Jego twierdzenie w pracy Z. Pawłowskiego – *Ekonometryczne metody badania popytu konsumpcyjnego, PWN,1961* (która była podstawą uzyskania przez Z. Pawłowskiego stopnia naukowego docenta (teraz doktora habilitowanego)). Profesor M. Kolupa miał wtedy 25 lat!

W dorobku naukowym Profesora Kolupy jest ponad 250 oryginalnych artykułów i wiele monografii[1]. Na szczególną uwagę zasługują: *Badania popytu w warunkach niedostatecznej podaży, PWE, 1965, Metody estymacji modeli ekonometrycznych, PWE, 1974, Wielokrotne macierze brzegowe i ich ekonometryczne zastosowania* (razem z Jerzym Nowakowskim) Biblioteka Ekonometryczna PWN 1990, *Miary zgodności, metody doboru zmiennych, problemy współliniowości (razem z M. Gruszczyńskim, E. Leniewską, G. Napiórkowskim)* Biblioteka Ekonometryczna PWN 1979, *Macierze brzegowe w badaniach ekonometrycznych, PWE 1982, Macierze brzegowe – teoria, algorytmy zastosowania ogólne i ekonometryczne (razem z E. Gruźlewską) PWE 1991, Teoria par korelacyjnych, WSZiM 1994, Budowa portfela lokat (podstawy teoretyczne, (razem z J. Plebaniak) PWE, 2000, Teoria par korelacyjnych, WSHiP im. R. Łazarskiego (2002).* Profesor Michał Kolupa stworzył własną szkołę naukową opartą o macierze brzegowe.

Szkoła naukowa Profesora Michała Kolupy to – macierze brzegowe, teoria portfelowa oraz teoria ekonometrii.

Profesor Michał Kolupa był wspaniałym kierownikiem zespołów naukowych. W zespołach tych powstawały ważne prace naukowe i dydaktyczne.

Zasługi Profesora Kolupy na polu kształcenia kadr naukowych są olbrzymie. Świadczą o tym także liczby. Profesor wypromował 17 doktorów, sporządził ponad 100 recenzji rozpraw doktorskich, ponad 100 recenzji rozpraw habilitacyjnych, ponad 40 recenzji wniosków o tytuł naukowy profesora.

Autorytet Profesora Michała Kolupy jest niezwykły. Profesor Z. Hellwig tak pisze, że: "Michał Kolupa jest Uczonym, który łączy w spójną całość sześć dziedzin wiedzy – algebrę (głównie liniową), teorię liniowych procesów, ekonomię w jej ujęciu ilościowym i stochastycznym, statystykę matematyczną i metody prognozowania, teorię systemów i modeli, konwencjonalną i nowoczesną naukę o finansach". Profesor Z. Hellwig napisał – „zasługa odkrycia Michała Kolupy należy do Profesora Pawłowskiego. Już na pierwszym seminarium słuchacze mogli ocenić Jego liczne zalety i talenty. Do najważniejszych zaliczono: erudycję matematyczną, świetną pamięć, elokwencję, poczucie humoru, łatwość zajmującego klarownego wykładu i co jest szczególnie ważne u pracownika nauki – żywą wyobraźnię, pomysłowość, odkrywczość, talent narratorski i literacki".

Środowisko polskich matematyków, statystyków i ekonometryków szybko dostrzegło rozległą wiedzę i wielką biegłość warsztatową Profesora Kolupy. Szczególnie wyraziście ujawniało się to przy tablicy – wzory, dowody twierdzeń, przekształcenia spływały spod kredy na czerń tablicy szybko, płynnie i lekko jak gdyby bez udziału prelegenta. Rzadko można spotkać wykładowcę, który bez pomocy rzutnika, notatek, kartek używ-

[1] Prawie pełny wykaz publikacji (do roku 2005) zawarty jest w pracy – *Od badań podstawowych do zastosowań w nowoczesnej gospodarce, 45-lecie ekonometrycznej szkoły Profesora Michała Kolupy, Kolegium Zarządzania i Finansów Szkoła Głowna Handlowa Warszawa 2005.*

wając oszczędnie kredy, pisząc litery i cyfry wyraźnie, szybko, prawie nie używając gąbki i nie pozostawiając śladów ani na katedrze ani na własnym ubraniu.

Profesor brał czynny udział w większości seminariów i konferencji naukowych, najczęściej z referatem oraz ważnym głosem w dyskusji.

Profesor Michał Kolupa był bardzo zaangażowany w prace Komitetu Statystyki i Ekonometrii PAN. W latach 1981-1983 pełnił funkcję wiceprzewodniczącego a w latach 1983-2011 członka Prezydium. W latach 1995-2005 owocnie kierował *Przeglądem Statystycznym* przyczyniając się do jego wysokiego poziomu merytorycznego.

Profesor Michał Kolupa był czułym mężem i Przyjacielem Małżonki Marii, Przyjacielem i dobrym ojcem ukochanej córki, autorytetem i ojcem dla swego zięcia i wspaniałym dziadkiem dla swego wnuka.

Profesor Michał Kolupa był patriotą a postawę tę zawdzięcza ojcu – uczestnikowi III Powstania śląskiego. Odznaczony Krzyżem Oficerskim i Kawalerskim Orderu Odrodzenia Polski oraz wieloma wyróżnieniami i godnościami. Był majorem Wojska Polskiego w stanie spoczynku.

Profesor Michał Kolupa miał doskonałą wiedzę o historii wojskowości i militariach, był zapalonym filatelistą (posiadał prawie pełny zbiór znaczków pocztowych z czasów Generalnej Guberni). Kto miał okazję odbyć z Profesorem spacer na łonie natury, to ze zdumieniem stwierdzał, że Profesor rozpoznaje ptaki po ich śpiewie.

Profesor inspirował adeptów nauki do poszukiwania i rozwiązywania istotnych problemów teoretycznych ekonometrii oraz ich praktycznych rozwiązań. Zawsze służył bezinteresowną pomocą wszystkim, którzy się do niego zwrócili. Zapisał się trwale w historii polskiej nauki, nie tylko poprzez swe publikacje, ale także swoją postawę jako obywatela, wzorowego Uczonego i wiernego Przyjaciela. Będziemy korzystać z Jego dorobku oraz niezłomnej postawy w poszukiwaniu prawdy naukowej[2].

Andrzej Stanisław Barczak

[2] Źródła wykorzystane: M. Kolupa – *Początki mego startu naukowego – wspomnienia i fakty, w pracy Badania ekonometryczne w teorii i praktyce (red. A. Barczak) UE Katowice, 2010 M. Kolupa – Identyfikacja modelu współzależnego na tle twierdzenia Kronekera – Capellego, w pracy Dylematy ekonometrii (red. J. Biolik) UE Katowice, 2009, ponad 500m stron w Internecie, Archiwum Autora.*

PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012

EMIL PANEK

O PEWNEJ INTERPRETACJI SYSTEMU WALRASA Z ZAPASAMI

Przedstawiona w artykule [3] interpretacja modelu równowagi ogólnej typu Walrasa z zapasami budziła od początku niedosyt autora, czego wyrazem były w następstwie m.in. dyskusje na seminariach naukowych Katedry Ekonomii Matematycznej UE w Poznaniu. Do dyskusji tych nawiązuje bezpośrednio niniejsza notatka, w której przedstawiamy inną od zaprezentowanej we wspomnianym artykule interpretację ekonomiczną równań dynamiki cen i zapasów.

W klasycznym modelu L. Walrasa (zob. np [1], [4]) dynamikę cen w n-produktowej gospodarce opisuje (w najprostszej wersji) układ równań różniczkowych

$$\dot{p}(t) = \sigma(f^d(p(t) - f^s(p(t)), \tag{1}$$

w którym $t \in [0, +\infty)$ oznacza zmienną (ciągłą) czasu, $p(t) = (p_1(t), ..., p_n(t))$ jest wektorem cen towarów w momencie t, $f^d(p) = (f_1^d(p), ..., f_n^d(p))$ jest funkcją zagregowanego popytu, a $f^s(p) = (f_1^s(p), ..., f_n^s(p))$ jest funkcją zagregowanej podaży towarów przy cenach p; σ jest dodatnim wskaźnikiem prędkości reakcji cen na zmianę popytu i podaży. Gospodarka osiąga równowagę, gdy ustalą się w niej takie ceny $\bar{p} > 0$, przy których dochodzi do zrównania popytu z podażą: $f^d(\bar{p}) - f^s(\bar{p})$. Wektor $\bar{p}$ nazywamy wektorem cen równowagi rynkowej.

Transakcje kupna-sprzedaży towarów dochodzą do skutku dopiero po osiągnięciu równowagi. W rzeczywistości codzienne transakcje zawierane są permanentnie bez względu na to czy ceny $p(t)$ są cenami równowagi, czy nie i żaden z konsumentów ani producentów specjalnie się nad tym nie zastanawia. Oznacza to, że chwilowo na rynku może powstawać niedobór pewnych towarów (gdy $f_i^d(p(t)) > f_s^s(p(t))$ lub ich nadprodukcja (gdy $f_i^d(p(t)) < f_s^s(p(t))$), co prowadzi do tworzenia zapasów, o których nie ma mowy w klasycznym systemie Walrasa.

W artykule [3] przedstawiony został prosty model gospodarki konkurencyjnej typu L. Walrasa uwzględniający zapasy, które – wbrew temu jednak co na ogół obserwujemy w rzeczywistości – nie mają wpływu na ceny towarów na rynku. Ponadto dochody konsumentów stanowią w nim równowartość produkcji (czystej) sprzedanej, a nie wytworzonej, jak w oryginalnym systemie Walrasa.

Zaprezentowany we wspomnianej pracy model tworzy układ $2n$ równań różniczkowych dynamiki cen (1) i zapasów:

$$\dot{z}_i(t) = \begin{cases} f_i^s(p(t)) - f_i^d(p(t)) - \mu z_i(t), & gdy \quad z_i(t) > 0, \\ \max\{f_i^s(p(t)) - f_i^d(p(t)), 0\}, & gdy \quad z_i(t) = 0, \end{cases} \quad i = 1, ..., n, \tag{2}$$

$z(t) = (z_1(t), ..., z_n(t))$ jest wektorem całkowitych zapasów w gospodarce w momencie t; $\mu \in (0, 1)$ jest stopą naturalnego ubytku towarów. Zapas towaru rośnie, gdy ma miejsce jego nadprodukcja (skorygowana o naturalny ubytek zapasów: gdy $f_i^s(p(t)) > f_i^d(p(t)) + \mu z_i(t)$) i maleje w przypadku niedoboru towaru (gdy $f^d(p(t)) > f^s(p(t)) - \mu z_i(t)$). Dystrybucja towarów pochodzących z zapasów obywa się nieodpłatnie poza rynkiem.

Poniżej proponujemy interpretację ekonomiczną modelu (1)-(2) odmienną od przedstawionej w pracy [3] i jak się wydaje bliższą rzeczywistości.

Jeżeli $p(t)$ jest wektorem cen towarów w momencie t, a $f^s(p(t))$ wektorem ich całkowitej podaży utożsamianej z produkcją wytworzoną, to wartość podaży towarów w gospodarce w momencie t wynosi $\langle p(t), f^s(p(t))\rangle$; symbolem $\langle x, y\rangle$ oznaczamy iloczyn skalarny wektorów $x, y \in R^n$.

Zakładamy, że strumień dochodów w gospodarce $D(p(t))$ jest równy wartości produkcji wytworzonej:

$$D(p(t)) = \langle p(t), f^s(p(t))\rangle.$$

Wektor produkcji sprzedanej w momencie t jest równy

$$S(p(t)) = \min\{f^s(p(t)), f^d(p(t))\} = \left(\min\{f_1^s(p(t)), f_1^d(p(t))\}, ..., \min\{f_n^s(p(t)), f_n^d(p(t))\}\right).$$

Oczywiście, $S(p(t)) \leq f^s(p(t))$. Różnica

$$f^s(p(t)) - S(p(t)) = \max\{f^s(p(t)) - f^d(p(t)), 0\} =$$

$$= (\max\{f_1^s(p(t)) - f_1^d(p(t)), 0\}, ..., \max\{f_n^s(p(t)) - f_n^d(p(t)), 0\}) \tag{3}$$

przedstawia wektor towarów, które w momencie t nie znalazły nabywcy i w związku z tym powiększają zapasy. Wartość niesprzedanej produkcji (wartość nadprodukcji) wynosi zatem

$$\langle p(t), f^s(p(t)) - S(p(t))\rangle = \langle p(t) \max\{f^s(p(t)) - f^d(p(t)), 0\}\rangle.$$

Sprzedana produkcja $S(p(t))$ zaspokaja (częściowo lub całkowicie) popyt w momencie t, którego wartość nie przekracza, oczywiście, dochodów konsumentów:

$$\langle p(t), S(p(t))\rangle \leq \langle p(t), D(p(t))\rangle.$$

Różnica

$$N(p(t)) = \max\{f^d(p(t)) - f^s(p(t)), 0\} =$$

$$= \left(\max\{f_1^d(p(t)) - f_1^s(p(t)), 0\}, ..., \max\{f_n^d(p(t)) - f_n^s(p(t)), 0\}\right) \tag{4}$$

przedstawia wektor niezaspokojonego popytu w momencie t, zatem $\langle p(t), N(p(t))\rangle$ oznacza wartość niezaspokojonego popytu w tym momencie. Jest to część dochodów konsumentów przewidziana na zakup towarów wyprodukowanych w niewystarczającej ilości. W celu zaspokojenia potrzeb konsumentów producenci sięgają po zapasy.

Zgodnie z prawem Walrasa, bez względu na ceny, wartość popytu jest równa wartości podaży:

$$\forall p \geq 0 \left(\langle p, f^d(p) \rangle = \langle p, f^s(p) \rangle \right)$$

i wobec tego

$$\langle p(t), N(p(t)) \rangle = \langle p(t), \max \{ f^d(p(t)) - f^s(p(t)), 0 \} \rangle$$

$$= \langle p(t), \max \{ f^s(p(t)) - f^d(p(t)), 0 \} \rangle,$$

tzn. wartość niezaspokojonego popytu jest równa wartości niezrealizowanej podaży.

Niezrealizowana podaż (nadwyżka produkcji) (3) powiększa, a niezaspokojony popyt pomniejsza zapasy towarów zgodnie z równaniem (2). W momencie t transakcje kupna-sprzedaży są zawierane po cenach $p(t)$ bez względu na to czy towary pochodzą z bieżącej produkcji, czy z zapasów (tzn. cena towaru pochodzącego z zapasów jest taka sama jak cena towaru pochodzącego z bieżącej produkcji).

Przy standardowych założeniach, w gospodarce takiej istnieją ceny równowagi określone jednoznacznie z dokładnością do struktury. Dowód został przytoczony w artykule [3], twierdzenie 2. Jest ona też globalnie asymptotycznie stabilna (zob. twierdzenie 3 w tymże artykule). Gospodarka w równowadze eliminuje ponadto zapasy.

Uniwersytet Ekonomiczny w Poznaniu

LITERATURA

[1] McKenzie L.W. (2002), Classical General Equilibrium Theory, The MIT Press, Cambridge
[2] Panek E., (2003), Ekonomia matematyczna, Wydawnictwo AEP, Poznań
[3] Panek E., (2011), „System Walrasa i zapasy", „Przegląd Statystyczny" nr 3-4, 195-204
[4] Takayama A., (1997), Mathematical Economics, Cambridge University Press

O PEWNEJ INTERPRETACJI SYSTEMU WALRASA Z ZAPASAMI

S t r e s z c z e n i e

Nawiązując do artykułu [3] zaproponowano odmienną jak się wydaje bliższą rzeczywistości ekonomicznej, interpretację modelu Walrasa z zapasami.

Słowa kluczowe: system Walrasa, popyt, podaż, równowaga konkurencyjna

A CERTAIN INTERPRETATION OF THE WALRAS'S SYSTEM AND STOCKS

A b s t r a c t

According to article [3], it has been proposed another interpretation of Walras's model and stocks, as it seems to be closer to economic reality.

Key words: Walras's economy, demand, supply, general equilibrium

PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012

MAREK WALESIAK

KLASYFIKACJA SPEKTRALNA A SKALE POMIARU ZMIENNYCH[1]

1. WPROWADZENIE

Analiza skupień bazująca na dekompozycji spektralnej (*spectral clustering*) rozwija się w literaturze poświęconej wielowymiarowej analizie danych od końca XX w. Nazwa metody „klasyfikacja spektralna" wywodzi się stąd, że w jednym z jej podstawowych kroków wyznacza się spektrum (widmo) macierzy Laplace'a. W matematyce zbiór wartości własnych macierzy nazywa się spektrum (widmem) macierzy (zob. np. [7], s. 182). Podstawowy algorytm klasyfikacji spektralnej dla danych metrycznych zaproponowano w pracy [8]. Inne algorytmy klasyfikacji spektralnej scharakteryzowano m.in. w pracach [10] i [14].

W artykule scharakteryzowano metodę klasyfikacji spektralnej z punktu widzenia skal pomiaru zmiennych. Rozpatrzono jej zastosowanie w klasyfikacji danych nominalnych, porządkowych, przedziałowych oraz ilorazowych. W tym celu w procedurze tej metody przy wyznaczaniu macierzy podobieństwa (*affinity matrix*) zastosowano funkcję (1) z miarami odległości właściwymi dla danych mierzonych na różnych skalach pomiaru. Dzięki takiemu podejściu dla danych niemetrycznych (nominalnych i porządkowych) możliwe jest pośrednie wzmocnienie skali pomiaru zmiennych.

Zaproponowana metoda klasyfikacji spektralnej może być z powodzeniem stosowana we wszystkich zagadnieniach klasyfikacyjnych, w tym dotyczących pomiaru, analizy i wizualizacji preferencji.

2. TYPY SKAL POMIAROWYCH I ICH CHARAKTERYSTYKA

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru, wprowadzone przez Stevensa w pracy [13]. Skale pomiaru są uporządkowane od najsłabszej do najmocniejszej: nominalna, porządkowa, przedziałowa, ilorazowa. Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową do niemetrycznych.

Podstawowe własności skal pomiaru przedstawia tab. 1.

[1] Praca naukowa finansowana ze środków na naukę w latach 2009-2012 jako projekt badawczy nr N N111 446037 nt. „Pomiar, analiza i wizualizacja preferencji ujawnionych i wyrażonych z wykorzystaniem metod wielowymiarowej analizy statystycznej i programu R".

Tabela 1.

Podstawowe własności skal pomiaru

Typ skali	Dozwolone przekształcenia matematyczne	Dopuszczalne relacje	Dopuszczalne operacje arytmetyczne
Nominalna	$z = f(x), f(x)-$ dowolne przekształcenie wzajemnie jednoznaczne	równości ($x_A = x_B$), różności ($x_A \neq x_B$)	zliczanie zdarzeń (liczba relacji równości, różności)
Porządkowa	$z = f(x), f(x)-$ dowolna ściśle monotonicznie rosnąca funkcja	powyższe oraz większości ($x_A > x_B$) i mniejszości ($x_A < x_B$)	zliczanie zdarzeń (liczba relacji równości, różności, większości, mniejszości)
Przedziałowa	$z = bx + a \quad (b > 0), z \in R$ dla wszystkich x zawartych w R, wartość zerowa na tej skali jest zwykle przyjmowana arbitralnie lub na podstawie konwencji*	powyższe oraz równości różnic i przedziałów ($x_A - x_B = x_C - x_D$)	powyższe oraz dodawanie i odejmowanie
Ilorazowa	$z = bx \quad (b > 0), z \in R_+$ dla wszystkich x zawartych w R_+, naturalnym początkiem skali ilorazowej jest wartość zerowa (zero lewostronnie ogranicza zakres skali)	powyższe oraz równości ilorazów ($\frac{x_A}{x_B} = \frac{x_C}{x_D}$)	powyższe oraz mnożenie i dzielenie

* Por. [1], s. 240.
Źródło: [18], s. 15.

Jedna z podstawowych reguł teorii pomiaru mówi, że jedynie rezultaty pomiaru w skali mocniejszej mogą być transformowane na liczby należące do skali słabszej (por. np. [11], s. 17; [12], s. 19; [16], s. 40; [23]; [24]). Bezpośrednia transformacja skal pomiaru zmiennych polegająca na ich wzmacnianiu nie jest możliwa, ponieważ z mniejszej ilości informacji nie można uzyskać większej jej ilości. W klasyfikacji spektralnej możliwe jest pośrednie wzmocnienie skali pomiaru zmiennych. Pierwotna macierz danych, w której zmienne mierzone są na skali nominalnej lub porządkowej zostaje przekształcona w macierz danych, w której zmienne mierzone są na skali przedziałowej.

3. KLASYFIKACJA SPEKTRALNA DLA RÓŻNYCH SKAL POMIARU ZMIENNYCH

Rys. 1 przedstawia trzy pierwsze etapy klasyfikacji spektralnej (występujące także w klasycznej analizie skupień), obejmujące ustalenie zbioru obiektów i zmiennych (po zgromadzeniu danych konstruuje się macierz danych, a w przypadku danych metrycznych w następnym kroku znormalizowaną macierz danych), dobór zmiennych oraz wybór miary odległości. Szczegółową charakterystykę tych etapów zaprezentowano m.in. w pracach [17] i [19].

Rysunek 1. Trzy pierwsze etapy w klasyfikacji spektralnej oraz w klasycznej analizie skupień

Źródło: Opracowanie własne.

Dalsza procedura klasyfikacji spektralnej obejmuje następujące kroki[2]:

4. Obliczenie symetrycznej macierzy podobieństw $\mathbf{A} = [A_{ik}]_{nxn}$ (*affinity matrix*) między obiektami, dla której $A_{ii} = 0$ oraz

$$A_{ik} = \exp(-\sigma \cdot d_{ik}) \text{ dla } \quad i \neq k, \tag{1}$$

gdzie: σ – parametr skali,

d_{ik} – miary odległości dla różnych skal pomiaru ujęte w tab. 2,

$i, k = 1, \ldots, n$ – numery obiektów.

W kroku tym można zastosować do obliczenia elementów macierzy podobieństw A_{ik} ($i \neq k$) estymatory jądrowe (zob. [6], s. 13-14; [9]): jądro gaussowskie, jądro wielomianowe, jądro liniowe, jądro w postaci tangensa hiperbolicznego, jądro Bessela, jądro Laplace'a, jądro ANOVA, jądro łańcuchowe (dla danych tekstowych).

W oryginalnym algorytmie klasyfikacji spektralnej dla danych metrycznych w pracy [8] zastosowano jądro gaussowskie:

$$A_{ik} = \exp(-\frac{d_{ik}^2}{2\sigma^2}) \text{ dla } \quad i \neq k, \tag{2}$$

[2] Jest to algorytm zaproponowany w pracy [8] (por. [20], [21]). W artykule dokonano jego modyfikacji w kroku 4 przy obliczaniu macierzy podobieństw (*affinity matrix*).

Tabela 2.

Miary odległości dla różnych skal pomiaru*

Nazwa miary odległości	Skala pomiaru zmiennych	Funkcja (pakiet) programu R
Minkowskiego: miejska (Manhattan); euklidesowa; Czebyszewa (maximum)	metryczne	`dist (stats)`
Canberra	ilorazowe	`dist (stats)`
Braya-Curtisa	ilorazowe	`dist.BC (clusterSim)`
GDM1	metryczne	`dist.GDM (clusterSim)`
GDM2	porządkowe	`dist.GDM (clusterSim)`
Sokala i Michenera	nominalne	`dist.SM (clusterSim)`

* Formuły odległości zawiera m.in. praca [18].
Źródło: opracowanie własne.

gdzie: $d_{ik} = \sqrt{\sum_{j=1}^{m} (z_{ij} - z_{kj})^2}$,

$z_{ij}, (z_{kj})$– znormalizowana wartość j-tej zmiennej dla i-tego (k-tego) obiektu.

5. Konstrukcja znormalizowanej macierzy Laplace'a $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ($\mathbf{D}$ – diagonalna macierz wag, w której na głównej przekątnej znajdują się sumy każdego wiersza z macierzy $\mathbf{A} = [A_{ik}]$, a poza główną przekątną są zera). W rzeczywistości znormalizowana macierz Laplace'a przyjmuje postać: $\mathbf{I} - \mathbf{L}$. W algorytmie dla uproszczenia analizy pomija się macierz jednostkową $\mathbf{I}$ (zob. [8]). Własności tej macierzy przedstawiono m.in. w pracy [15], s. 5-6.

6. Obliczenie wartości własnych i odpowiadających im wektorów własnych (o długości równej jeden) dla macierzy $\mathbf{L}$. Uporządkowanie wektorów własnych według malejących wartości własnych. Pierwsze u wektorów własnych (u – liczba klas) tworzy macierz $\mathbf{E} = \left[e_{ij}\right]$ o wymiarach $n \times u$.

Podobnie jak w przypadku klasycznym analizy skupień zachodzi potrzeba ustalenia optymalnej liczby klas. Odpowiedni algorytm zaproponował Girolami w pracy [4].

Macierz podobieństw (*affinity matrix*) $\mathbf{A} = [A_{ik}]$ (dla $\sigma = 1$) poddawana jest dekompozycji $\mathbf{A} = \mathbf{U}\ \mathbf{\Lambda}\ \mathbf{U}^T$, gdzie $\mathbf{U}$ jest macierzą wektorów własnych macierzy $\mathbf{A}$ składającą się z wektorów $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$, a $\mathbf{\Lambda}$ jest macierzą diagonalną zawierającą wartości własne $\lambda_1, \lambda_2, \ldots, \lambda_n$.

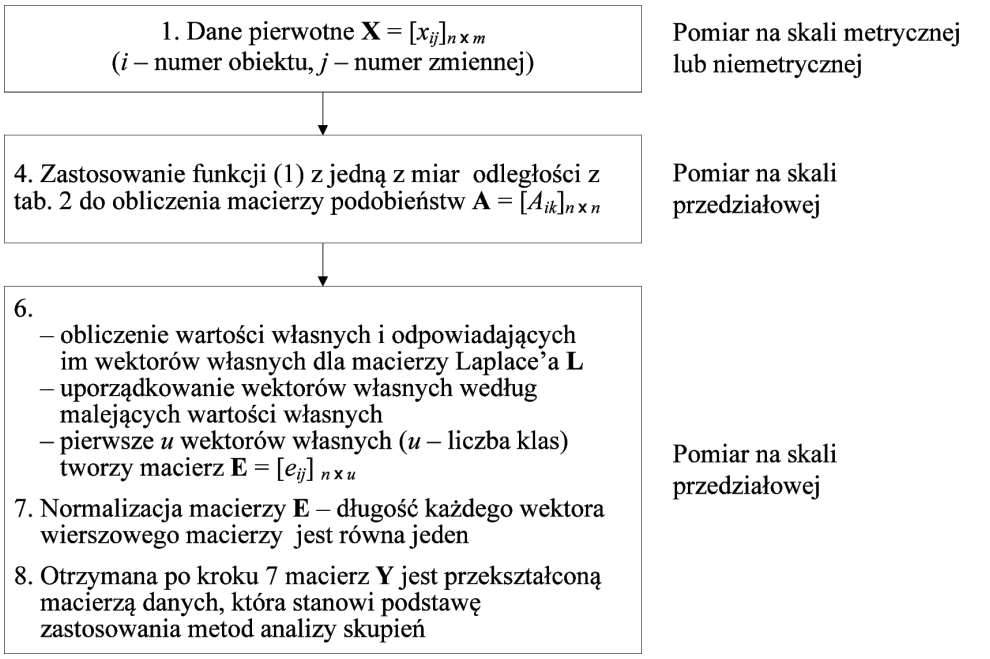
Obliczany jest wektor $\mathbf{K} = (k_1, k_2, \ldots, k_n)$, gdzie $k_i = \lambda_i \left\{\mathbf{1}_n^T \mathbf{u}_i\right\}^2$ ($\mathbf{1}_n^T$ – wektor o wymiarach $1 \times n$ zawierający wartości $1/n$). Wektor $\mathbf{K}$ jest porządkowany malejąco, a liczba jego dominujących elementów (wyznaczona np. poprzez kryterium osypiska) wyznacza optymalną liczbę skupień u, na którą algorytm klasyfikacji spektralnej powinien podzielić zbiór badanych obiektów.

7. Przeprowadza się normalizację macierzy $\mathbf{E}$ zgodnie ze wzorem $y_{ij} = e_{ij} \Big/ \sqrt{\sum_{j=1}^{u} e_{ij}^2}$ ($i = 1, \ldots, n$ – numer obiektu, $j = 1, \ldots, u$ – numer zmiennej, u – liczba klas). Dzięki

tej normalizacji długość każdego wektora wierszowego macierzy $\mathbf{Y} = \left[y_{ij}\right]$ jest równa jeden.

8. Macierz **Y** stanowi punkt wyjścia zastosowania klasycznych metod analizy skupień (proponuje się tutaj wykorzystanie metody *k*-średnich).

Rys. 2 pokazuje wybrane kroki postępowania w klasyfikacji spektralnej i odpowiadające im skale pomiaru.



Rysunek 2. Wybrane kroki postępowania w klasyfikacji spektralnej i odpowiadające im skale pomiaru

Źródło: Opracowanie własne.

Jeśli dane pierwotne $\mathbf{X} = [x_{ij}]$ mierzone są na skali niemetrycznej (porządkowej, nominalnej) w wyniku zastosowania funkcji (1) z jedną z odległości właściwych dla tych skal pomiaru (zob. tab. 2) podobieństwa w macierzy $\mathbf{A} = [A_{ik}]$ mierzone są na skali przedziałowej. Ostatecznie w kroku 7 otrzymuje się metryczną macierz danych **Y** o wymiarach $n \times u$. Pozwala ona na zastosowanie dowolnych metod analizy skupień (w tym metod bazujących bezpośrednio na macierzy danych, np. metodę *k*-średnich).

4. PARAMETR σ W KLASYFIKACJI SPEKTRALNEJ

Parametr σ ma fundamentalne znaczenie w klasyfikacji spektralnej. W literaturze zaproponowano wiele heurystycznych sposobów wyznaczania wartości tego parametru (zob. np. [3]; [9]; [25]). W metodach heurystycznych wyznacza się wartość σ na podstawie pewnych statystyk opisowych macierzy odległości $[d_{ik}]$. Lepszy sposób wyznaczania parametru σ zaproponował Karatzoglou w pracy [6]. Poszukuje się takiej

wartości parametru σ, która minimalizuje zmienność wewnątrzklasową przy zadanej liczbie klas u. Jest to heurystyczna metoda poszukiwania minimum lokalnego. Zbliżony koncepcyjnie algorytm znajdowania optymalnego parametru σ zaproponowano w pracy [20]:

Krok 0. Wybierana jest próba bootstrapowa $\mathbf{X}'$ składającą się z n' obiektów opisanych wszystkimi m zmiennymi (wartość n' jest najczęściej tak dobierana, aby $\frac{1}{2}n \leq n' \leq \frac{3}{4}n$). Początkowy przedział przeszukiwania optymalnej wartości parametru σ ustalany jest jako $S_0 = [0;\ D]$ (gdzie D oznacza sumę odległości w dolnym trójkącie macierzy odległości a dla kwadratu odległości euklidesowej – pierwiastek z sumy odległości w dolnym trójkącie macierzy odległości).

Krok 1. Przedział S_k (gdzie k oznacza numer iteracji; na początku $S_k = S_0$) dzielony jest na przedziały jednakowej długości: $p_r^k = [\underline{p_r^k};\ \overline{p_r^k}]$, $r = 1, \ldots, R$ (R – liczba przedziałów w każdej iteracji; domyślnie $R = 10$).

Krok 2. Dla każdego przedziału p_r^k obliczamy jego środek: $\sigma_r^k = \frac{\underline{p_r^k} + \overline{p_r^k}}{2}$. Dla wszystkich wartości σ_r^k przeprowadzana jest klasyfikacja spektralna zbioru $\mathbf{X}'$ na ustaloną liczbę klas u.

Krok 3. Wybierane jest takie σ_r^k, dla którego zmienność wewnątrzklasowa jest minimalna.

Krok 4. Z przedziałem zawierającym wybraną wartość σ_r^k w kroku 3 przechodzi się do kroku 1 i kontynuuje procedurę do osiągnięcia zadanej liczby iteracji (domyślnie: 3).

5. OPROGRAMOWANIE W ŚRODOWISKU R

Klasyfikację spektralną zgodną z algorytmem zmodyfikowanym w artykule przeprowadza się z wykorzystaniem funkcji `speccl` pakietu `clusterSim` (zob. [22]):

```
speccl(data,nc,distance="GDM1",sigma="automatic",
sigma.interval="default",mod.sample=0.75,R=10,iterations=3)
```

Argumenty:

`data`	macierz danych
`nc`	liczba klas
`distance`	miary odległości z tabeli 2 (`"sEuclidean"` – kwadrat odległości euklidesowej, `"euclidean"` – odległość euklidesowa, `"manhattan"` – odległość miejska, `"maximum"` – odległość Czebyszewa, `"canberra"` – odległość Canberra, `"BC"` – odległość Braya-Curtisa, `"GDM1"` – odległość GDM dla danych metrycznych, `"GDM2"` – odległość GDM dla danych porządkowych, `"SM"` – odległość Sokala-Michenera dla danych nominalnych)

`sigma`	parametr skali: `sigma="automatic"` – parametr ustalany automatycznie zgodnie z algorytmem z punktu 4 `sigma=200` – parametr podany przez użytkownika, np. 200
`sigma.interval`	przedział przeszukiwania parametru sigma: `sigma.interval="default"` – przedział wartości od zera do sumy odległości w dolnym trójkącie macierzy odległości (dla kwadratu odległości euklidesowej – do pierwiastka z sumy odległości w dolnym trójkącie macierzy odległości) `sigma.interval=1000` – przedział wartości od zera do wartości podanej przez użytkownika, np. 1000
`mod.sample`	proporcja danych stosowanych do estymacji parametru sigma
`R`	liczba przedziałów w każdej iteracji
`iterations`	maksymalna liczba iteracji

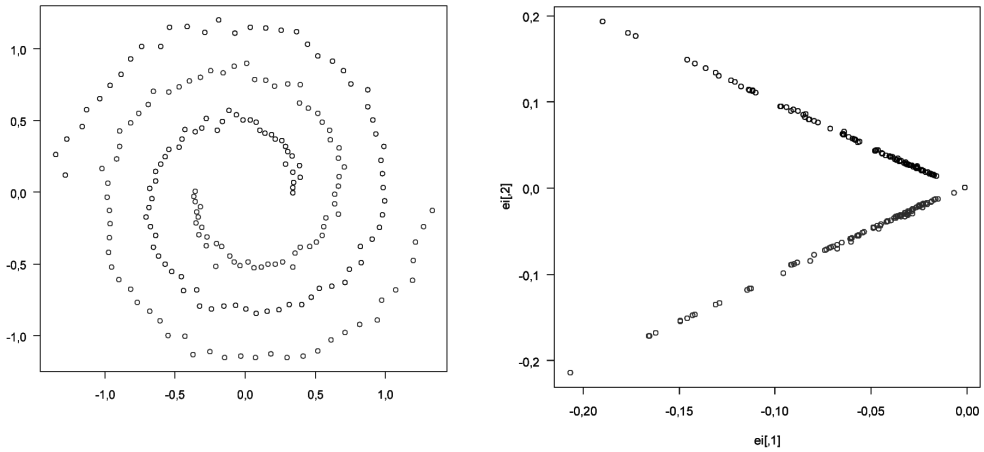
Graficzną prezentację wybranych kroków klasyfikacji spektralnej dla danych metrycznych przedstawiających strukturę dwóch klas zobrazowano na rys. 3. Do wygenerowania zbioru danych metrycznych wykorzystano funkcję `mlbench.spirals` pakietu `mlbench` (zob. rys. 3a). Do klasyfikacji zbioru obiektów zastosowano metodę klasyfikacji spektralnej wyznaczając w kroku 4 macierz podobieństw zgodnie ze wzorem (1) z odległością GDM1. Rys. 3b i 3c prezentują odpowiednio obiekty z macierzy $\mathbf{E}$ o wymiarach 200×2 (krok 6) oraz obiekty ze znormalizowanej macierz $\mathbf{Y} = \left[y_{ij}\right]$ o wymiarach 200×2 (krok 7).

Graficzną prezentację wybranych kroków klasyfikacji spektralnej dla danych porządkowych przedstawiających strukturę trzech klas zobrazowano na rys. 4. Do wygenerowania zbioru danych porządkowych[3] wykorzystano funkcję `cluster.Gen` pakietu `clusterSim` (zob. rys. 4a). Do klasyfikacji zbioru obiektów zastosowano metodę klasyfikacji spektralnej wyznaczając w kroku 4 macierz podobieństw zgodnie ze wzorem (1) z odległością `GDM2`. Rys. 4b i 4c prezentują odpowiednio obiekty z macierzy $\mathbf{E}$ o wymiarach 150×3 (krok 6) oraz obiekty ze znormalizowanej macierz $\mathbf{Y} = \left[y_{ij}\right]$ o wymiarach 150×3 (krok 7).

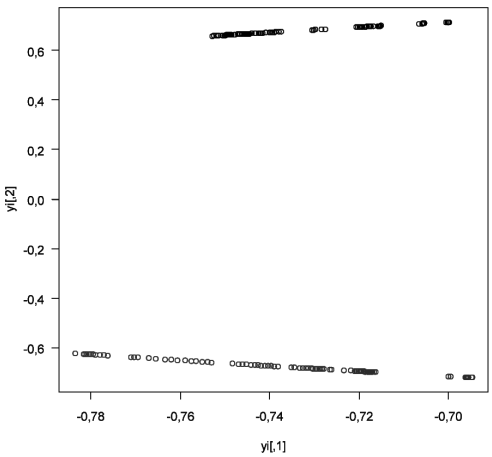
[3] Przy tworzeniu wykresu rozrzutu dla danych porządkowych trzeba wziąć pod uwagę częstość występowania identycznych par kategorii. W funkcji `plotCategorial` znajduje to wyraz w długości promienia koła.

a) graficzna prezentacja zbioru danych

b) zbiór danych w przestrzeni dwóch wektorów własnych macierzy Laplace'a



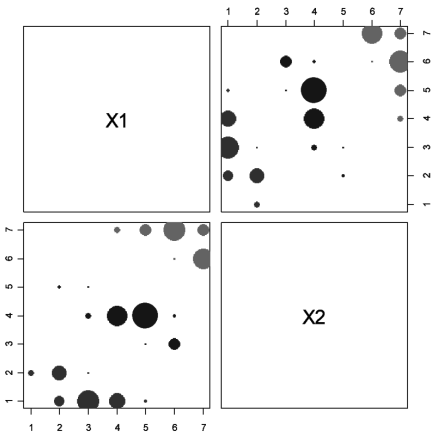
c) zbiór danych w przestrzeni dwóch wektorów własnych macierzy Laplace'a po normalizacji



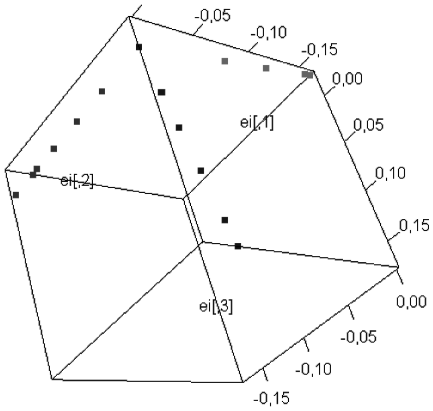
Rysunek 3. Wybrane etapy klasyfikacji spektralnej dla przykładowego zbioru danych metrycznych wygenerowanego z wykorzystaniem funkcji `mlbench.spirals` pakietu `mlbench`

Źródło: Opracowanie własne.

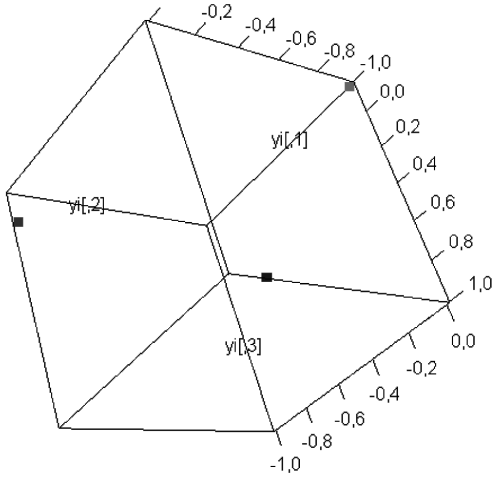
a) graficzna prezentacja zbioru danych



b) zbiór danych w przestrzeni trzech wektorów własnych macierzy Laplace'a



c) zbiór danych w przestrzeni trzech wektorów własnych macierzy Laplace'a po normalizacji



Rysunek 4. Wybrane etapy klasyfikacji spektralnej dla przykładowego zbioru danych porządkowych wygenerowanego z wykorzystaniem funkcji `clusterGen` pakietu `clusterSim`

Źródło: Opracowanie własne.

6. ANALIZA PORÓWNAWCZA METOD KLASYFIKACJI SPEKTRALNEJ Z METODAMI ANALIZY SKUPIEŃ DLA DANYCH O ZNANEJ STRUKTURZE KLAS

Analizę porównawczą metod klasyfikacji spektralnej z metodami analizy skupień, z uwzględnieniem różnych miar odległości, dla danych o znanej strukturze klas przeprowadzono dla trzech typów danych.

W eksperymencie pierwszym i trzecim wykorzystano odpowiednio dane metryczne oraz porządkowe o znanej strukturze klas obiektów wygenerowane z wykorzystaniem funkcji `cluster.Gen` pakietu clusterSim na podstawie modeli zawartych w tab. 3.

Tabela 3.

Charakterystyka modeli w analizie symulacyjnej

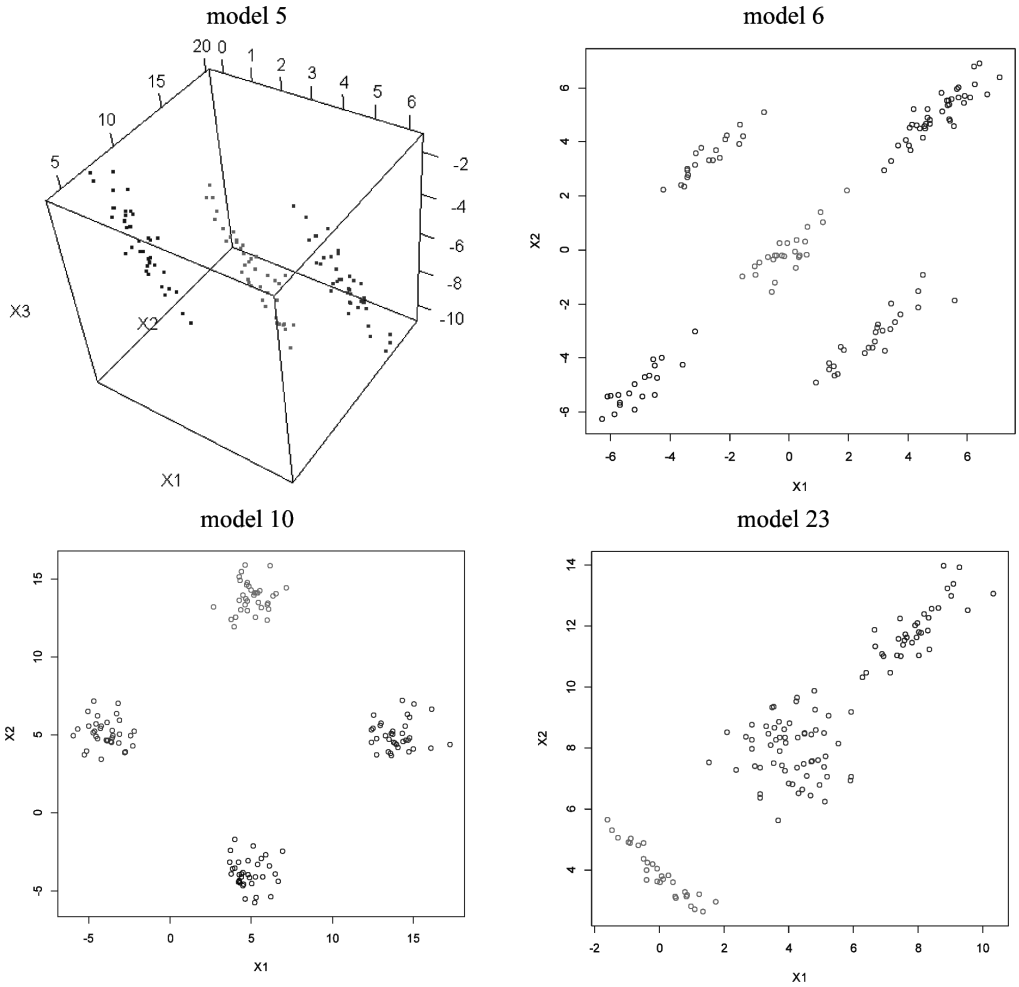
Nr modelu	*m*	*nk**	*u*	*lo*	środki ciężkości klas	Macierz kowariancji	*ks*
5	3	7	3	40	(1,5; 6, – 3), (3; 12; –6) (4,5; 18; –9)	$\sigma_{jj} = 1\ (1 \le j \le 3)$ $\sigma_{12} = \sigma_{13} = -0,9, \quad \sigma_{23} = 0,9$	1
6	2	5, 7	5	40, 20, 25, 25, 20	(5; 5), (–3; 3), (3; –3), (0; 0), (–5; –5)	$\sigma_{jj} = 1, \quad \sigma_{jl} = 0,9$	2
10	2	6, 8	4	35	(–4; 5), (5; 14), (14; 5), (5; –4)	$\sigma_{jj} = 1, \quad \sigma_{jl} = 0$	3
23	2	5	3	30, 60, 35	(0; 4), (4; 8), (8; 12)	$\Sigma_1 = \begin{bmatrix} 1 & -0,9 \\ -0,9 & 1 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 1 & 0,9 \\ 0,9 & 1 \end{bmatrix}$	4

* tylko dla danych porządkowych;
m – liczba zmiennych, *nk* – liczba kategorii (jedna liczba oznacza stałą liczbę kategorii); *u* – liczba klas; *lo* – liczba obiektów w klasach (jedna liczba oznacza klasy równoliczne); *ks* – kształt skupień: a) skupienia dobrze separowalne (1 – skupienia wydłużone, 3 – skupienia normalne), skupienia słabo separowalne (2 – skupienia wydłużone, 4 – skupienia zróżnicowane dla klas).
Źródło: [21].

Na rys. 5 i 6 przedstawiono graficzną prezentację przykładowych zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` dla danych metrycznych (rys. 5) i danych porządkowych (rys. 6).

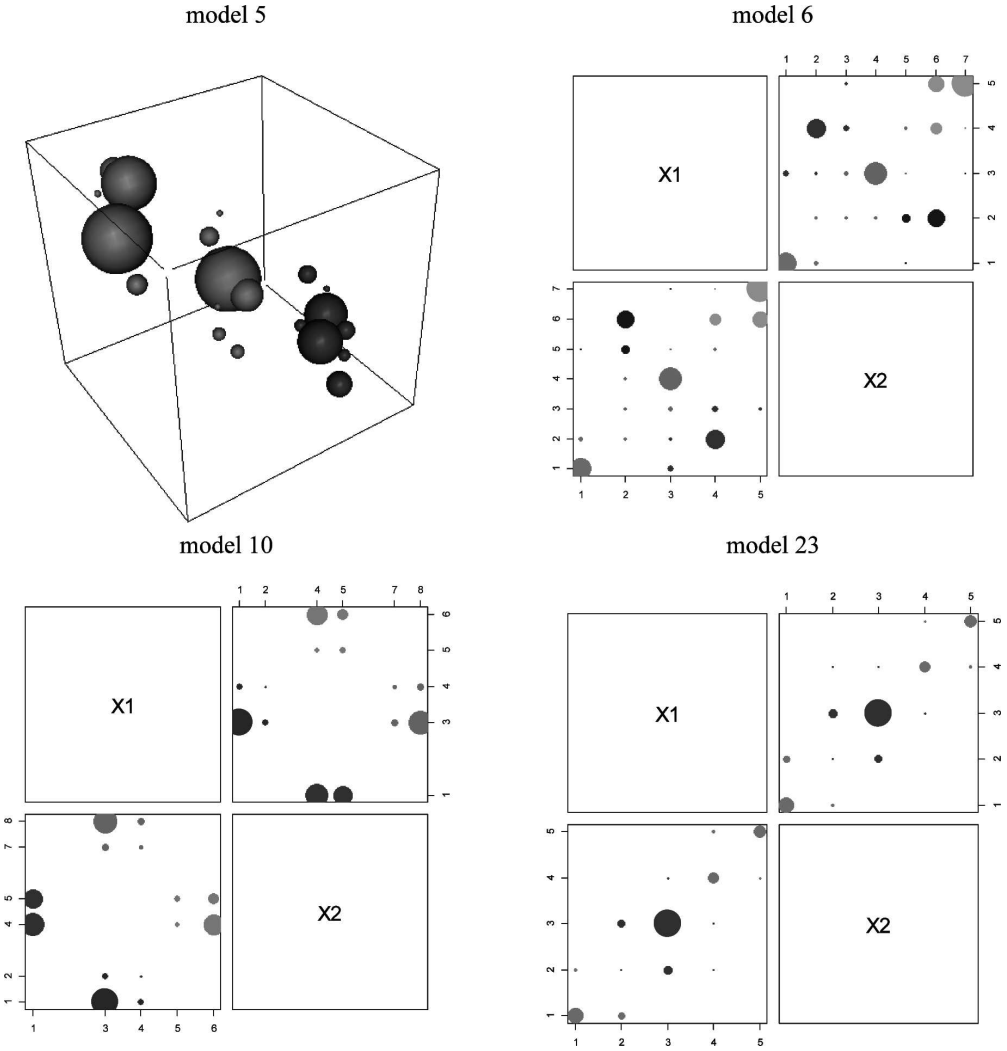
W eksperymencie drugim zbiory danych zawierające 360 obiektów (zob. rys. 7) wygenerowano z wykorzystaniem funkcji pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` [20] i `banana` [2].

Dla modeli w każdym eksperymencie wygenerowano 40 zbiorów danych, przeprowadzono procedurę klasyfikacyjną i porównano otrzymane rezultaty klasyfikacji ze znaną strukturą klas przy pomocy skorygowanego indeksu Randa (zob. [5]).



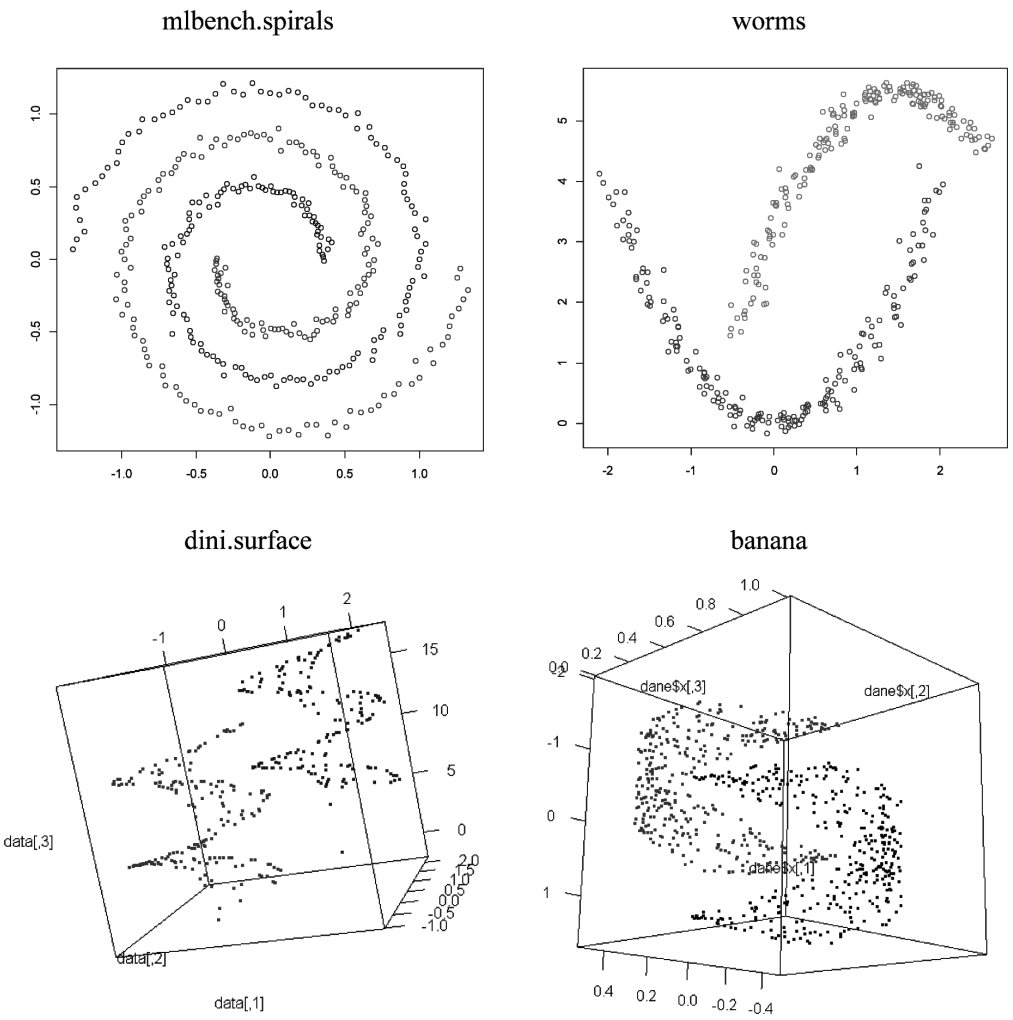
Rysunek 5. Graficzna prezentacja przykładowych zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` (dane metryczne)

Źródło: opracowanie własne z wykorzystaniem programu **R**.



Rysunek 6. Graficzna prezentacja przykładowych zbiorów danych utworzonych z wykorzystaniem funkcji `cluster.Gen` pakietu `clusterSim` (dane porządkowe)

Źródło: opracowanie własne z wykorzystaniem programu **R**.



Rysunek 7. Przykładowe zbiory danych utworzone z wykorzystaniem funkcji pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `banana`

Źródło: opracowanie własne z wykorzystaniem programu **R**.

Dla danych metrycznych (eksperyment 1 i 2) uwzględniono następujące metody klasyfikacji (z odległościami: (1) – kwadrat odległości euklidesowej, (2) – odległość euklidesowa, (3) – odległość miejska, (4) – odległość GDM1): 1. speccl – klasyfikacja spektralna; 2. pam – metoda *k*-medoidów; 3. complete – metoda kompletnego połączenia; 4. average – metoda średniej klasowej; 5. ward – metoda Warda; 6. centroid – metoda środka ciężkości; 7. diana – hierarchiczna metoda deglomeracyjna. Ponadto dla celów porównawczych zastosowano dodatkowo metodę *k*-średnich, która bazuje bezpośrednio na macierzy danych.

Dla danych porządkowych (eksperyment 3) uwzględniono w analizie metody klasyfikacji o numerach 1-7 z odległością GDM2.

Tab. 4 prezentuje uporządkowanie analizowanych metod klasyfikacji (z 4 odległościami) według średnich wartości skorygowanego indeksu Randa policzonego z 40 symulacji dla danych metrycznych wygenerowanych w pakiecie clusterSim.

Tabela 4.
Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych metrycznych wygenerowanych w pakiecie clusterSim

Poz.	Metoda	średnia*	Kształt skupień				Liczba zmiennych zakłócających		
			1	2	3	4	0	1	2
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	ward(3)	0,753	1,000	0,983	1,000	0,960	0,985	0,701	0,574
2	average(3)	0,746	0,989	0,974	1,000	0,954	0,979	0,695	0,563
3	pam(3)	0,736	0,999	0,992	1,000	0,967	0,989	0,701	0,517
4	speccl(2)	0,723	1,000	0,953	1,000	0,919	0,968	0,780	0,420
5	diana(3)	0,676	0,934	0,758	0,997	0,791	0,870	0,646	0,513
6	speccl(4)	0,670	0,899	0,851	0,898	0,867	0,879	0,695	0,435
7	speccl(1)	0,658	0,900	0,945	0,925	0,848	0,904	0,767	0,303
8	speccl(3)	0,643	0,950	0,948	0,975	0,884	0,939	0,692	0,298
9	complete(3)	0,632	0,751	0,733	1,000	0,921	0,851	0,573	0,471
10	average(4)	0,601	0,989	0,982	1,000	0,943	0,978	0,496	0,328
11	average(2)	0,599	1,000	0,975	1,000	0,957	0,983	0,495	0,320
12	ward(4)	0,594	1,000	0,978	1,000	0,960	0,984	0,478	0,320
13	pam(4)	0,591	1,000	0,975	1,000	0,924	0,974	0,482	0,317
14	ward(2)	0,590	1,000	0,979	1,000	0,960	0,985	0,471	0,316
15	pam(2)	0,582	1,000	0,992	1,000	0,966	0,989	0,455	0,302
16	centroid(3)	0,581	0,878	0,906	1,000	0,923	0,927	0,570	0,245
16	pam(1)	0,581	1,000	0,992	1,000	0,967	0,989	0,442	0,312
18	average(1)	0,577	1,000	0,973	1,000	0,962	0,984	0,463	0,285
19	ward(1)	0,575	1,000	0,979	1,000	0,964	0,986	0,446	0,293
20	centroid(4)	0,559	0,989	0,980	1,000	0,946	0,979	0,442	0,256

cd. Tabela 4.

21	diana(2)	0,534	0,988	0,757	0,983	0,846	0,894	0,438	0,270
22	centroid(1)	0,513	0,989	0,964	1,000	0,959	0,978	0,371	0,190
23	kmeans	0,502	0,819	0,839	0,898	0,967	0,881	0,406	0,219
24	diana(4)	0,491	0,966	0,762	0,992	0,582	0,826	0,404	0,242
25	diana(1)	0,457	0,988	0,735	0,992	0,678	0,848	0,345	0,179
26	complete(4)	0,447	0,894	0,912	1,000	0,886	0,923	0,281	0,135
27	complete(2)	0,437	0,960	0,869	1,000	0,931	0,940	0,253	0,119
27	complete(1)	0,437	0,960	0,869	1,000	0,931	0,940	0,253	0,119
29	centroid(2)	0,427	0,989	0,942	1,000	0,926	0,964	0,298	0,019

* (k8+k9+k10)/3, gdzie k8 = (k4+k5+k6+k7)/4
Liczba w nawiasie przy nazwach metod klasyfikacji: (1) – kwadrat odległości euklidesowej, (2) – odległość euklidesowa, (3) – odległość miejska, (4) – odległość GDM1.

Źródło: obliczenia własne z wykorzystaniem programu R[4].

W przypadku typowych zbiorów danych metrycznych metody klasyfikacji spektralnej sprawdzają się dobrze w odkrywaniu rzeczywistej struktury klas (pozycje: 4, 6, 7 i 8 w zestawieniu). W przeprowadzonym eksperymencie najlepiej strukturę klas odkrywały metody klasyczne (ward, average i pam) z odległością miejską. Wśród metod klasyfikacji spektralnej dominuje klasyfikacja spektralna z odległością euklidesową. Nieco gorsze rezultaty otrzymuje się z wykorzystaniem klasyfikacji spektralnej z odległością GDM1 (poz. 6 w zestawieniu).

Tab. 5 prezentuje uporządkowanie analizowanych metod klasyfikacji (z 4 odległościami) według średnich wartości skorygowanego indeksu Randa policzonego z 40 symulacji dla nietypowych danych metrycznych wygenerowanych z wykorzystaniem pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `banana`.

Dla nietypowych zbiorów danych metody klasyfikacji spektralnej zdecydowanie lepiej od klasycznych metod analizy skupień odkrywają prawidłową strukturę klas. Klasyfikacja spektralna z odległością GDM1 daje rezultaty lepsze od metod klasyfikacji spektralnej z pozostałymi odległościami.

Tab. 6 prezentuje uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa policzonego z 40 symulacji dla danych porządkowych wygenerowanych w pakiecie `clusterSim`.

W przypadku zbiorów danych porządkowych bez zmiennych zakłócających najlepsza jest metoda Warda. Metoda klasyfikacji spektralnej z odległością GDM2 daje gorsze rezultaty od klasycznych metod analizy skupień (za wyjątkiem metody diana). Należy jednak pamiętać, że zbiory tego typu bardzo rzadko występują w rzeczywistych problemach klasyfikacyjnych. Uwzględnienie zmiennych zakłócających pokazuje wyraźną przewagę metody klasyfikacji spektralnej z odległością GDM2.

[4] Skrypty do analiz symulacyjnych z punktu 6 są autorstwa dra Andrzeja Dudka.

Tabela 5.
Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych metrycznych otrzymanych z pakietów `mlbench` (`mlbench.spirals`), `geozoo` (`dini.surface`) oraz zbiorów `worms` i `banana`

Poz.	Metoda	średnia*	Zbiory danych			
			spirals	worms	dini	banana
1	*2*	*3*	*4*	*5*	*6*	*7*
1	speccl(4)	0,837	0,961	0,929	0,916	0,544
2	speccl(3)	0,822	0,901	0,959	0,694	0,736
3	speccl(2)	0,779	0,938	0,985	0,563	0,631
4	speccl(1)	0,741	1,000	0,889	0,407	0,671
5	pam(3)	0,259	0,006	0,438	0,274	0,316
6	average(3)	0,251	0,030	0,468	0,221	0,284
7	ward(3)	0,249	0,034	0,474	0,214	0,276
8	pam(2)	0,220	0,025	0,503	0,184	0,169
9	complete(3)	0,216	0,022	0,440	0,206	0,195
10	pam(4)	0,214	0,016	0,519	0,172	0,147
10	pam(1)	0,214	0,026	0,517	0,175	0,138
12	ward(1)	0,213	0,036	0,499	0,170	0,148
13	average(2)	0,211	0,039	0,517	0,152	0,135
13	diana(2)	0,211	0,040	0,528	0,155	0,120
15	diana(4)	0,210	0,037	0,516	0,158	0,129
16	diana(3)	0,209	-0,001	0,486	0,181	0,172
16	complete(2)	0,209	0,033	0,488	0,141	0,172
16	complete(1)	0,209	0,033	0,488	0,141	0,172
19	ward(2)	0,208	0,053	0,471	0,144	0,165
20	average(4)	0,205	0,029	0,471	0,143	0,177
20	kmeans	0,205	0,032	0,519	0,159	0,111
22	diana(1)	0,204	0,032	0,515	0,159	0,112
22	average(1)	0,204	0,034	0,503	0,150	0,130
24	ward(4)	0,202	0,046	0,487	0,142	0,132
25	centroid(1)	0,197	0,022	0,520	0,141	0,107
26	centroid(4)	0,194	0,038	0,478	0,145	0,116
27	complete(4)	0,193	0,041	0,464	0,140	0,126
28	centroid(3)	0,170	0,006	0,460	0,134	0,079
29	centroid(2)	0,167	0,020	0,487	0,083	0,078

* (k4+k5+k6+k7)/4
Liczba w nawiasie przy nazwach metod klasyfikacji: (1) – kwadrat odległości euklidesowej, (2) – odległość euklidesowa, (3) – odległość miejska, (4) – odległość GDM1.
Źródło: obliczenia własne z wykorzystaniem programu **R**.

Tabela 6.
Uporządkowanie analizowanych metod klasyfikacji według średnich wartości skorygowanego indeksu Randa dla danych porządkowych wygenerowanych w pakiecie `clusterSim`

Poz.	Metoda	średnia*	Kształt skupień				Liczba zmiennych zakłócających		
			1	2	3	4	0	1	2
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	speccl(5)	0,696	0,998	0,951	0,798	0,777	0,881	0,709	0,497
2	average(5)	0,602	1,000	0,968	1,000	0,962	0,982	0,495	0,327
3	pam(5)	0,593	1,000	0,971	1,000	0,934	0,976	0,483	0,321
4	ward(5)	0,591	1,000	0,971	1,000	0,973	0,986	0,471	0,317
5	centroid(5)	0,560	1,000	0,962	1,000	0,965	0,982	0,451	0,248
6	diana(5)	0,493	0,959	0,753	0,998	0,595	0,826	0,388	0,266
7	complete(5)	0,444	0,882	0,885	1,000	0,851	0,904	0,279	0,149

* (k8+k9+k10)/3, gdzie: k8 = (k4+k5+k6+k7)/4
Liczba (5) w nawiasie przy nazwach metod klasyfikacji oznacza odległość GDM2.
Źródło: obliczenia własne z wykorzystaniem programu R.

7. PODSUMOWANIE

W artykule zaproponowano modyfikację metody klasyfikacji spektralnej umożliwiającą jej zastosowanie w klasyfikacji danych prezentowanych na różnych skalach pomiaru. W procedurze klasyfikacji spektralnej, zaproponowanej przez autorów Ng, Jordan i Weiss [8], wprowadzono modyfikację polegającą na zastosowaniu funkcji (1) z miarami odległości właściwymi dla danych mierzonych na różnych skalach pomiaru. Dodatkowo dzięki takiemu podejściu pośrednio wzmacnia się skale pomiaru zmiennych. Dane niemetryczne zostają przekształcone w dane przedziałowe. Umożliwia to zastosowanie w klasyfikacji zbioru obiektów m.in. metody k-średnich. Scharakteryzowano funkcję `speccl` pakietu `clusterSim` umożliwiającą klasyfikację spektralną zgodną z algorytmem zmodyfikowanym w artykule.

W tym miejscu wskazać trzeba na ograniczenia związane z klasyfikacją spektralną. Efektywne wykorzystanie metod klasyfikacji spektralnej jest uzależnione od prawidłowego doboru parametru skali σ. W części 4 zaprezentowano heurystyczną metodę poszukiwania minimum lokalnego.

W części 6 poświęconej analizie porównawczej metod klasyfikacji spektralnej z metodami analizy skupień dla danych o znanej strukturze klas przeprowadzono analizy symulacyjne dla danych metrycznych oraz porządkowych. Nie uwzględniono badań symulacyjnych dotyczących takiego porównania dla danych nominalnych. Wynikało to z braku metod generowania danych nominalnych o znanej strukturze klas.

Uniwersytet Ekonomiczny we Wrocławiu

LITERATURA

[1] Ackoff R.L. (1969), *Decyzje optymalne w badaniach stosowanych*, PWN, Warszawa.
[2] Dudek A. (2012), *A comparison of the performance of clustering methods using spectral approach*, W: J. Pociecha, R. Decker (red.), *Data analysis methods and its applications*, Wydawnictwo C.H. Beck, Warszawa, 143-156.
[3] Fischer I., Poland J. (2004), *New methods for spectral clustering*, Technical Report No. IDSIA-12-04, Dalle Molle Institute for Artificial Intelligence, Manno-Lugano, Switzerland.
[4] Girolami M. (2002), *Mercer kernel-based clustering in feature space*, „IEEE Transactions on Neural Networks", vol. 13, no. 3, 780-784.
[5] Hubert L., Arabie P. (1985), *Comparing partitions*, „Journal of Classification", no. 1, 193-218.
[6] Karatzoglou A. (2006), *Kernel methods. Software, algorithms and applications*, Rozprawa doktorska, Uniwersytet Techniczny we Wiedniu.
[7] Kolupa M. (1976), *Elementarny wykład algebry liniowej dla ekonomistów*, Państwowe Wydawnictwo Naukowe, Warszawa.
[8] Ng A., Jordan M., Weiss Y. (2002), *On spectral clustering: analysis and an algorithm*, W: T. Dietterich, S. Becker, Z. Ghahramani (red.), Advances in Neural Information Processing Systems 14, Cambridge, MIT Press, 849-856.
[9] Poland J., Zeugmann T. (2006), *Clustering the Google distance with eigenvectors and semidefinite programming*, Knowledge Media Technologies, First International Core-to-Core Workshop, Dagstuhl, July 23-27, Germany.
[10] Shortreed S. (2006), *Learning in spectral clustering*, Rozprawa doktorska, University of Washington.
[11] Steczkowski J., Zeliaś A. (1981), *Statystyczne metody analizy cech jakościowych*, PWE, Warszawa.
[12] Steczkowski J., Zeliaś A. (1997), *Metody statystyczne w badaniach cech jakościowych*, Wydawnictwo AE, Kraków.
[13] Stevens S.S. (1946), *On the theory of scales of measurement*, „Science", Vol. 103, No. 2684, 677-680.
[14] Verma D., Meila M. (2003), *A comparison of spectral clustering algorithms*. Technical report UW-CSE-03-05-01, University of Washington.
[15] von Luxburg U. (2007), *A tutorial on spectral clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149.
[16] Walesiak M. (1990), *Syntetyczne badania porównawcze w świetle teorii pomiaru*, „Przegląd Statystyczny", z. 1-2, 37-46.
[17] Walesiak M. (2005), *Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów*, W: A. Zeliaś (red.), „Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych", Wydawnictwo AE, Kraków, 185-203.
[18] Walesiak M. (2006), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*. Wydanie drugie rozszerzone. Wydawnictwo AE, Wrocław.
[19] Walesiak M. (2009), *Analiza skupień*, W: M. Walesiak, E. Gatnar (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa, 407-433.
[20] Walesiak M., Dudek A. (2009), *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe UE we Wrocławiu nr 84, 9-19.
[21] Walesiak M., Dudek A. (2010), *Klasyfikacja spektralna z wykorzystaniem odległości GDM*, W: K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 17, Prace Naukowe UE we Wrocławiu nr 107, 161-171.
[22] Walesiak M., Dudek A. (2011), *clusterSim package*, URL http://www.R-project.org.
[23] Wiśniewski J.W. (1986), *Korelacja i regresja w badaniach zjawisk jakościowych na tle teorii pomiaru*, „Przegląd Statystyczny", z. 3, 239-248.
[24] Wiśniewski J.W. (1987), *Teoria pomiaru a teoria błędów w badaniach statystycznych*, „Wiadomości Statystyczne", nr 11, 18-20.

[25] Zelnik-Manor L., Perona P. (2004), *Self-tuning spectral clustering*, W: Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04), http://books.nips.cc/nips17.html.

KLASYFIKACJA SPEKTRALNA A SKALE POMIARU ZMIENNYCH

Streszczenie

W artykule zaproponowano modyfikację metody klasyfikacji spektralnej (zob. Ng, Jordan i Weiss [2002]) umożliwiającą jej zastosowanie w klasyfikacji danych nominalnych, porządkowych, przedziałowych oraz ilorazowych. W tym celu w procedurze tej metody przy wyznaczaniu macierzy podobieństwa (*affinity matrix*) zastosowano funkcję $A_{ik} = \exp(-\sigma \cdot d_{ik})$ (σ – parametr skali) z miarami odległości d_{ik} właściwymi dla danych mierzonych na różnych skalach pomiaru. Takie podejście umożliwia ponadto pośrednie wzmocnienie skali pomiaru zmiennych dla danych niemetrycznych.

Zaproponowana metoda klasyfikacji spektralnej może być z powodzeniem stosowana we wszystkich zagadnieniach klasyfikacyjnych, w tym dotyczących pomiaru, analizy i wizualizacji preferencji.

Słowa kluczowe: klasyfikacja spektralna, miary odległości, skale pomiaru

SPECTRAL CLUSTERING AND MEASUREMENT SCALES OF VARIABLES

Abstract

In article the proposal of modification of spectral clustering method for nominal, ordinal, interval and ratio data, based on procedure of Ng, Jordan and Weiss [2002], is presented. In construction of affinity matrix we implement function $A_{ik} = \exp(-\sigma \cdot d_{ik})$ (σ – scale parameter) with distance measures d_{ik} appropriate for different scales of measurement. This approach gives possibility of conversion nonmetric data (nominal, ordinal) into interval data.

The proposed method of spectral clustering can be successfully used in all classification problems, including the measurement, analysis and visualization of preferences.

Key words: spectral clustering, distance measures, scales of variables

PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012

HENRYK GURGUL, TOMASZ WÓJTOWICZ

HIPOTEZA BRODY W ŚWIETLE WYNIKÓW BADAŃ EMPIRYCZNYCH

1. METODA I TABLICE INPUT-OUTPUT

Pierwszą ideę metody input-output można znaleźć (za [7]) w dziele Francois Quesnaya *Tableau Economique* (1758). W tej książce autor przedstawił gospodarkę wymienną w postaci ruchu okrężnego. Ideę tę rozbudował następnie Leon Walras. Jednak nowoczesną analizę input-output (zwaną również w literaturze polskiej analizą nakładów i wyników albo analizą przepływów międzygałęziowych) stworzył dopiero W. Leontief. Jej podstawy zostały wyłożone w roku 1941 w pracy [7], gdzie zostały przedstawione wzajemne powiązania w gospodarce Stanów Zjednoczonych w okresie 1919-1939. Po opublikowaniu tej pracy rozpoczął się burzliwy, wszechstronny rozwój teorii input-output.

Metodologia input-output była i jest nadal intensywnie rozwijana i modyfikowana. Obecnie opiera się ona na tzw. Systemie Rachunków Narodowych (SNA'93, ang. skrót *System of National Accounts*). Pierwszą wersję SNA opracowaną przez laureatów Nagrody Nobla Simona Kuznetsa i Richarda Stone'a opublikowano w 1952 roku. Wersja SNA z roku 1993 (zalecana przez ONZ) stanowi uporządkowany wewnętrznie zbiór spójnych, logicznych i zintegrowanych rachunków makroekonomicznych, bilansów i tablic opracowanych według obowiązujących norm i reguł statystycznych. Odmianą SNA'93 jest system ESA'95 (*European System of Accounts*) bardziej nastawiony na warunki i dane wymagane przez Unię Europejską. System ten jest dopasowany do pojęć i klasyfikacji stosowanych w wielu innych statystykach społecznych i gospodarczych Unii Europejskiej. Dzięki temu może być podstawą odniesienia dla statystyki społecznej i gospodarczej krajów członkowskich. Klasyfikacje i założenia polskiego systemu rachunków narodowych są zgodne zarówno z założeniami zalecanymi przez ONZ jak i przez Unię Europejską. Polski System Rachunków Narodowych uwzględnia jednak pewną specyfikę polskiego systemu organizacyjnego i prawnego.

Istotną rolę w systemie ESA'95 odgrywają tablice wykorzystania (ang. *use tables*) w układzie produkt/działalności oraz tablice podaży (ang. *supply tables*) wiążące wytwarzanie produktów z określonymi działalnościami. W ESA'95 nie występuje jako takie pojęcie gałęzi. Jednakże w rozważaniach teoretycznych i empirycznych używa się tego pojęcia jako agregatu obejmującego produkty czyli produkcję przedsiębiorstw lub zakładów o określonym rodzaju działalności. Symetryczne tablice przepływów międzygałęziowych są tablicami typu produkt/produkt lub gałąź/gałąź i łączą tablice wyko-

rzystania i tablice podaży w jedną tablicę z identyczną klasyfikacją produktów lub gałęzi w zastosowaniu do kolumn i wierszy. Tablice przepływów międzygałęziowych w układzie produktowym w przypadku polskiej gospodarki są zestawiane zgodnie z PKWiU (Polska Klasyfikacja Wyrobów i Usług). Natomiast rodzaje działalności uwzględniane w tablicach wykorzystania i podaży klasyfikowane są zgodnie z Polską Klasyfikacją Działalności (skrót PKD2004, ang. skrót NACE). W części empirycznej tego artykułu wykorzystano tablice przepływów międzygałęziowych w układzie produkt/produkt.

W pewnym uproszczeniu tablica przepływów międzygałęziowych w swojej najogólniejszej postaci może być rozumiana jako ilustracja *ex post* ruchu okrężnego strumieni dóbr i usług pomiędzy jednostkami produkcyjnymi połączonymi w gałęzie. Strumienie te są wyrażone przeważnie w ujęciu wartościowym w walucie krajowej i dotyczą najczęściej jednego roku. Proces produkcyjny będący centralnym punktem tablic input-output może być rozumiany jako "przekształcenie" nakładów (input) w produkty (output). Nakłady dzielą się na nakłady bieżące (inaczej bezpośrednie) oraz wartość dodaną. Popyt na produkty dzieli się na pośredni oraz końcowy.

Tablica IO opisująca wzajemne powiązania gałęziowe wewnątrz całego kraju lub regionu opiera się na następujących teoretycznych założeniach:

a) cała gospodarka lub region są podzielone na gałęzie,
b) każda z gałęzi stosując określoną technologię produkuje jednorodne dobro,
c) wszystkie procesy produkcyjne są efektywne przy stałej skali produkcji,
d) w procesie produkcyjnym są stosowane czynniki produkcji, których zasoby są ograniczone.

W praktyce realnie funkcjonujące gospodarki nigdy ściśle nie spełniają podanych warunków, a zwłaszcza b) i c).

Tablica przepływów międzygałęziowych może być przedstawiona schematycznie za pomocą macierzy blokowych podanych w tabeli 1.

Tabela 1.

Macierze blokowe tablicy IO

Macierz zużycia produkcyjnego I	Macierz popytu końcowego II
Macierz wartości dodanej III	Macierz podziału wtórnego IV

Źródło: opracowanie własne.

Centralną rolę w tablicy odgrywa I ćwiartka (macierz I) obrazująca strumienie dóbr i usług przepływające między gałęziami produkcyjnymi. Suma elementów danego wiersza podaje łączną wielkość produkcji pozostającej w tej gałęzi oraz przepływającej do pozostałych gałęzi. Suma elementów dowolnej kolumny tej macierzy obrazuje

łączne zużycie produkcyjne własnych produktów oraz produktów pozostałych gałęzi. Macierz I jest z reguły kwadratowa, choć dla określonych celów konstruuje się też macierze prostokątne. Macierz popytu końcowego (macierz II) obrazuje w wierszach zużycie produktów gałęzi w podziale na poszczególne kategorie popytu końcowego. Kolumny tej macierzy pokazują gałęziowy skład odpowiednich kategorii popytu końcowego. Sumy poszczególnych wierszy obrazują zużycie końcowe produktów gałęzi, sumy kolumn ukazują globalne wielkości poszczególnych kategorii popytu końcowego. W praktyce wyróżnia się najczęściej następujące kategorie popytu końcowego: spożycie indywidualne, spożycie instytucjonalne i agend rządowych, przyrost zapasów i rezerw, nakłady na środki trwałe i kapitalne remonty oraz eksport. Macierz popytu końcowego jest prostokątna. Wyjaśnijmy jeszcze, że nakłady inwestycyjne nie oznaczają tu inwestycji poczynionych w poszczególnych gałęziach, a jedynie zrealizowany popyt na dobra inwestycyjne dostarczone przez produkujące te dobra gałęzie.

W III ćwiartce w różnych wersjach tablic przeważnie wykazuje się pozycję podatki od produktów pomniejszone o dotacje do produktów, koszty związane z zatrudnieniem, podatki od producentów, dotacje dla producentów oraz nadwyżkę operacyjną brutto. Tę ostatnią pozycję oblicza się jako różnicę pomiędzy wartością dodaną brutto a kosztami związanymi z zatrudnieniem i podatkami od producentów minus dotacje dla producentów. Nadwyżkę operacyjną netto uzyskuje się po odjęciu amortyzacji.

Dane z zakresu przepływów międzygałęziowych (np. popyt końcowy, produkcję globalną) podaje się najczęściej w cenach bazowych. Ceny bazowe otrzymuje się poprzez odjęcie od każdego przepływu w cenach nabywcy podatków od produktów pomniejszonych o dotacje do produktów oraz marż handlowych i transportowych. Dane dotyczące przepływów wyrobów i usług w bilansie przepływów międzygałęziowych w cenach bazowych dla produkcji krajowej otrzymuje się przez odjęcie od przepływów wyrobów i usług w cenach bazowych wyrobów i usług pochodzących z importu.

Podział wartości globalnej w poszczególnych gałęziach można wyrazić za pomocą tożsamości

$$\sum_k X_{ik} + Y_i = X_i \tag{1}$$

lub zakładając, że $a_{ik} X_k = X_{ik}$ otrzymuje się równość

$$\mathbf{Ax} + \mathbf{y} = \mathbf{x}, \tag{2}$$

gdzie macierz **A** jest macierzą nakładów bezpośrednich (lub bieżących) albo macierzą materiałochłonności bezpośredniej. W niektórych publikacjach, zwłaszcza niemieckojęzycznych jest też nazywana macierzą produkcji.

Wzór (2) definiuje tzw. statyczny model Leontiefa. Ma on liczne zastosowania, w tym i niestandardowe (por. np. [6], [10], [12]). Jak widać kluczową rolę w statycznym modelu Leontiefa odgrywa nieujemna macierz nakładów **A** bazująca na danych z I ćwiartki tablic przepływów międzygałęziowych. Jest ona także zasadniczą składową modelu dynamicznego Leontiefa stanowiącego uogólnienie modelu (2). Dla modeli dynamicznych IO definiuje się tzw. rozwiązanie zbilansowane, tzn. taką ścieżkę wzrostu

sektorów gospodarki (tzw. ścieżka równowagi), na której nie zmieniają się ani proporcje produkcji sektorów ani ich tempo wzrostu (z roku na rok jest takie samo). Okazuje się, że w świetle hipotezy Samuelsona ten typ wzrostu gospodarki jest optymalny.

Naturalnym jest pytanie, od czego zależy tempo powrotu gospodarki, po wystąpieniu ewentualnego zaburzenia, do tak rozumianego stanu równowagi. Pewne wnioski w tym zakresie pojawiły się w literaturze ekonomicznej w kontekście dyskusji wokół tzw. hipotezy Brody, której omówieniu jest poświęcony następny podrozdział.

2. MACIERZE INPUT-OUTPUT

A. Brody [3] napisał, że wyznaczenie punktu lub ścieżki równowagi w przypadku dużych systemów Leontiefa lub von Neumanna wymaga mniejszej liczby iteracji niż w przypadku małych systemów, czyli systemów silnie zagregowanych. Z tej obserwacji dotyczącej obliczeń wywiódł przypuszczenie, że zdezagregowane systemy ekonomiczne szybciej osiągają proporcje Frobeniusa niż systemy silnie zagregowane. W obliczeniach, o których wspomniał autor chodziło głównie o iteracyjne uzyskanie rozwiązania zrównoważonego o określonej dokładności.

Jak wiadomo dla dowolnej macierzy **A** o dodatnich elementach zachodzi następująca równość

$$v_m = A^m v = \lambda_1^m x_1 + \lambda_2^m x_2 + \ldots + \lambda_n^m x_n, \tag{3}$$

gdzie $\mathbf{v} = \mathbf{x}_1 + \ldots + \mathbf{x}_n$ oraz λ_i i $\mathbf{x}_i$ oznaczają odpowiednio wartość własną i związany z nią wektor własny macierzy **A**. Wartości własne są ponumerowane według ich malejących modułów. Ze wzrostem m lewa strona będzie się stawała coraz bardziej podobna do tej składowej prawej strony, która jest związana z największą (dominującą) wartością własną, zwaną też wartością własną Frobeniusa lub liczbą Frobeniusa. Jest to liczba rzeczywista i dodatnia, co wynika z twierdzenia Frobeniusa.

Łatwo zauważyć, że liczba iteracji niezbędna do "dopasowania" $\mathbf{A}^m\mathbf{v}$ do $\lambda_1^m x_1$ jest zdeterminowana przez wielkość drugiej co do wielkości wartości własnej macierzy **A**. Jeśli rząd macierzy **A** jest równy 1, tzn. jeśli macierz ta składa się na przykład z takich samych liczb różnych od zera, to liczba Frobeniusa jest równa 1, a wszystkie pozostałe wartości własne takiej macierzy są równe 0. Dlatego iloczyn

$$\mathbf{v}_{k+1} = \mathbf{A}\mathbf{v}_k \tag{4}$$

prowadzi do proporcji równowagi już po pierwszym mnożeniu macierzy **A** oraz $\mathbf{v}_k$. W tym przypadku zbieżność jest najszybszą z możliwych, gdyż proporcje równowagi są osiągane w pojedynczym kroku.

Macierz **A** utożsamiana jest z macierzą materiałochłonności bezpośredniej (nakładów bezpośrednich), która jest macierzą kwadratową o elementach nieujemnych. Jej wymiary zależą od stopnia agregacji, tzn. od ilości gałęzi gospodarki uwzględnionych w tablicy IO, która jest podstawą macierzy **A.**

Posługując się intuicją i eksperymentami komputerowymi Brody sformułował następującą hipotezę ([3], str. 255):

"Im większe wymiary ma (losowa) kwadratowa macierz **A***, tzn. im więcej elementów (gałęzi lub branż) zawiera ta macierz, tym mniejszy jest stosunek jej drugiej wartości własnej do dominującej wartości własnej Frobeniusa i w konsekwencji tym szybsza jest jej zbieżność do stanu równowagi, czyli wejście na zbilansowaną ścieżkę wzrostu (ścieżka Frobeniusa)".*

W celu uzasadnienia swojej hipotezy Brody zdefiniował macierz losową o elementach niezależnych stochastycznie i mających rozkład jednostajny na przedziale (0, 1). Łatwo zauważyć, że dla współczynników tak zdefiniowanej losowej macierzy nakładów parametry opisowe wynoszą odpowiednio: wartość oczekiwana $E(a_{ij}) = \frac{1}{2}$, odchylenie standardowe $D(a_{ij}) = \frac{1}{2\sqrt{3}}$. Widać, że $E(a_{\cdot j}) = \frac{n}{2}$, gdzie $a_{\cdot j}$ oznacza sumę j-tej kolumny macierzy **A**. Dzieląc wyrazy każdej kolumny tej macierzy przez jej wartość oczekiwaną $n/2$ otrzymuje się macierz $\mathbf{A}^*$ o postaci

$$A^* = A \cdot \begin{bmatrix} \frac{2}{n} & & \frac{2}{n} \\ & \ddots & \\ \frac{2}{n} & & \frac{2}{n} \end{bmatrix}.$$

Z założenia wynika, że współczynniki a^*_{ij} macierzy losowej $\mathbf{A}^*$ są niezależne stochastycznie i mają rozkład jednostajny na przedziale $\left(0, \frac{2}{n}\right)$. Wartości oczekiwane tych współczynników wynoszą $E(a^*_{ij}) = \frac{1}{n}$, zaś odchylenia standardowe $D(a^*_{ij}) = 1/(n\sqrt{3})$ dla każdych $i, j = 1,\ldots, n$. Zachodzi też $E(a^*_{\cdot j}) = 1$ oraz $D(a^*_{\cdot j}) = 1/(\sqrt{3n})$, gdzie $a^*_{\cdot j}$ oznacza sumę elementów w j–tej kolumnie macierzy $\mathbf{A}^*$. Dlatego – jak przekonuje Brody – iloraz $\lambda_2(\mathbf{A}^*)/\lambda_1(\mathbf{A}^*)$ drugiej wartości własnej i pierwiastka Frobeniusa (ten iloraz jest oczywiście taki sam również dla macierzy **A**) będzie oscylował wokół liczby $1/(\sqrt{3n})$. Łatwo zauważyć, że rząd $(E(\mathbf{A}^*))=1$ oraz spektrum macierzy $E(\mathbf{A}^*)$ stanowią elementy zbioru $\{1,0,\ldots,0\}$ dla każdego n. Wraz z rosnącym n realizacje losowych współczynników a^*_{ij} stają się coraz bliższe ich wartości oczekiwanej $(1/n)$. Dlatego spektrum macierzy $\mathbf{A}^*$ zmierza do spektrum $E(\mathbf{A}^*)$. Dla macierzy $\mathbf{A}^*$ zachodzi więc własność: im większe n, tym bliższa zera jest wartość $\lambda_2(A^*)$ (lub dla macierzy **A**: im większe n, tym bliższy 0 będzie iloraz $\lambda_2(\mathbf{A})/\lambda_1(\mathbf{A})$). Ten efekt potwierdziły wyniki zeprezentowanej w artykule symulacji.

Wkrótce po ukazaniu się pracy Brody dwaj autorzy Molnar i Simonovits [9] próbowali rozwiązać pokrewny problem. Zamiast macierzy losowej, o której pisał Brody zaproponowali ujęcie deterministyczne. Przeskalowali oni macierz **A** w ten sposób, aby miała pojedynczą (równą 1) dominującą wartość własną (wartość własna Frobeniusa).

Wymienieni autorzy rozważali macierze o nieujemnych elementach, których sumy w kolumnach wynoszą jeden. Takie macierze nazywa się w teorii prawdopodobieńst-

wa macierzami stochastycznymi. Ich wartość własna Frobeniusa wynosi 1. Rozważania autorów dotyczyły macierzy stochastycznych spełniających dodatkowo warunek, że elementy tych macierzy są "bliskie" wartościom $1/n$.

Typowym przykładem ekonomicznym macierzy stochastycznej **A** jest macierz nakładów IO reprezentująca tzw. "*self-replacing system*" (statyczny zamknięty model Leontiefa – system reprodukcji prostej, tzn. gospodarka produkuje tylko tyle dóbr aby pokryć nakłady i nie więcej). Taki system może się sam odtwarzać, ale nie dostarcza żadnej nadwyżki ("produkcja dla produkcji", brak nadwyżek na rynek).

Formalne sformułowanie Molnara i Simonovitsa [9] warunku prawdziwości hipotezy Brody brzmi:

"Jeśli każdy element macierzy **A** odchyla się od 1/n nie więcej niż $\tau/n^{1+\varepsilon}$, to moduł drugiej wartości własnej jest nie większy niż τ/n^{ε}, gdzie τ oraz ε są pewnymi liczbami rzeczywistymi".

Wynik ten uzyskany na drodze analitycznej potwierdza hipotezę Brody, jednakże tylko w pewnym w szczególnym przypadku, a mianowicie, gdy macierz **A** jest "bliska" macierzy **E**, składającej się tylko z elementów $1/n$. Zauważmy, że jeśli w przykładzie Brody przeskalowałoby się macierz **A** do $\mathbf{A}^*$, to odchylenia realizacji zmiennych losowych a_{ij}^* od liczb $1/n$ zmniejszałyby się z rosnącym n i znikałyby w nieskończoności. Tak więc wynik Brody ma ścisły związek z cytowanym twierdzeniem Molnara i Simonovitsa.

3. DRUGA WARTOŚĆ WŁASNA MACIERZY NAKŁADÓW BIEŻĄCYCH W ZAMKNIĘTYM MODELU IO REPRODUKCJI PROSTEJ (*SELF-REPLACING SYSTEM*)

Zaraz na początku swojego artykułu Brody napisał, że "*estymator drugiej wartości własnej czysto losowej macierzy współczynników nakładów pokazuje, że jego wartość oczekiwana maleje monotonicznie do zera wraz ze wzrostem wymiarów tej macierzy*".

Jednak w dalszej części swojej pracy Brody przyznał, że "*A special distribution and/or a special structure of the matrix may still permit exceptions*" czyli, że dla specjalnych postaci rozkładów prawdopodobieństwa współczynników macierzy nakładów czy dla specjalnych postaci samych macierzy nakładów możliwe są odstępstwa i wyjątki od jego hipotezy.

Białas i Gurgul [1] zbadali hipotezę Brody w przypadku, gdy macierz **A** jest macierzą nakładów bezpośrednich, dla której sumy w kolumnach wynoszą 1. Macierz ta występuje w statycznym, zamkniętym modelu Leontiefa. W modelu tym pierwiastek Frobeniusa λ_1 (nieujemnej i nierozkładalnej) macierzy **A** wynosi 1. Dlatego dla innych wartości własnych musi zachodzi zależność:

$$\lim_{m\to\infty} \lambda_i^m = 0, \text{ for } \quad i = 2, \ldots, n,$$

gdzie λ_i oznacza i-ty pierwiastek charakterystyczny macierzy **A**.

Jak wiadomo macierz $\mathbf{P}=[p_{ij}]\in \mathrm{IR}^{n\times n}$ nazywa się dodatnią i stochastyczną, gdy $p_{ij}>0$ $i,\ j=1,\ldots,\ n$ oraz $\sum_{i=1}^{n} p_{ij}=1,\ j=1,\ 2,\ \ldots,\ n$. Dla takiej szczególnej macierzy (dodatniej i stochastycznej) hipoteza Brody została sformułowana przez autorów w postaci:

Jeśli $\mathbf{A}=[a_{ij}]\in \mathrm{IR}^{n\times n}$ jest macierzą dodatnią i stochastyczną oraz $\lambda_n=\max_{1<i\leq n}|\lambda_i(\mathbf{A})|$, to

$$\lambda_1(A)=1 \text{ i } \quad \lim_{n\to\infty}\lambda_n=0.$$

Aby zbadać prawdziwość powyższej hipotezy Białas i Gurgul przytoczyli za książką Marcusa i Minca, [8] następujące dwie własności macierzy stochastycznych:

Własność 1. Wartości własne macierzy stochastycznych $\mathbf{P}=[p_{ij}]\in \mathrm{IR}^{n\times n}$ spełniają warunki:

$$|\lambda_i(P)|\leq 1, \quad i=1,\ldots,n \text{ oraz } \quad \lambda_1(P)=1.$$

Własność 2. Jeśli $\mathbf{P}=[p_{ij}]\in \mathrm{IR}^{n\times n}$ jest macierzą stochastyczną oraz $\omega=\min_{1\leq i\leq n} p_{ii}$, to

$$|\lambda_i(P)-\omega|\leq 1-\omega \text{ dla } \quad i=1,2,\ldots,n.$$

Poprzez kontrprzykład Białas i Gurgul pokazali następnie, że hipoteza Brody nie jest prawdziwa dla dowolnej stochastycznej macierzy nakładów.

Autorzy rozważyli macierz stochastyczną o elementach dodatnich wyrażoną wzorem

$$\bar{P}=[\bar{p}_{ij}], \tag{5}$$

gdzie:

$$\bar{p}_{11}=\frac{3}{4},$$

$\frac{3}{4}<\bar{p}_{ii}$ dla $i=2,\ 3,\ \ldots,\ n$,

$\bar{p}_{ij}>0$ dla $i,\ j=1,\ 2,\ \ldots,\ n,\ i\neq j$,

$\sum_{i=1}^{n}\bar{p}_{ij}=1$ dla $j=1,\ 2,\ \ldots,\ n$.

Łatwo zauważyć, że dla macierzy (5) zachodzi

$$\omega=\min_{1\leq i\leq n}\bar{p}_{ii}=\frac{3}{4}.$$

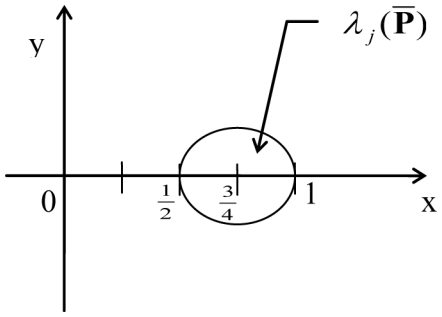
Biorąc pod uwagę *Własność 2* autorzy otrzymali, że dla dowolnego $j=1,...,n$

$$\left|\lambda_j(\bar{P})-\frac{3}{4}\right|\leq\frac{1}{4}. \tag{6}$$

Przyjmując $\lambda_j(\bar{P})=x+iy$, gdzie $i=\sqrt{-1}$. Powyższa równość może być zapisana w postaci:

$$\left(x-\frac{3}{4}\right)^2+y^2\leq\frac{1}{16}. \tag{7}$$

A to oznacza, że wszystkie wartości własne leżą w kole o środku w punkcie $(3/4, 0)$ i promieniu 1/4.



Rysunek 1. Koło, w którym zawierają się wszystkie wartości własne (wzór (7))

Autorzy zauważyli, że $\lambda_1(\bar{P}) = 1$ oraz $|\lambda_j(\bar{P})| \geq \frac{1}{2}$ dla $j = 2, \ldots, n$. Dlatego $\lambda_n = \max_{1<i\leq n}|\lambda_i(\bar{P})| \geq \frac{1}{2}$ oraz $\lim_{n\to\infty} \lambda_n \neq 0$, co dowiodło, że hipoteza Brody sformułowana dla tego szczególnego przypadku nie jest prawdziwa.

Poza tym Białas i Gurgul udowodnili, że dla dowolnej liczby rzeczywistej $\alpha \in (0, 1)$ istnieje dodatnia i stochastyczna macierz $\mathbf{P} = [p_{ij}] \in \mathrm{IR}^{n\times n}$ taka, że

$$\alpha \leq |\lambda_i(P)| \leq 1, \quad i = 1, 2, \ldots, n.$$

Ta teza wynikła stąd, że macierz

$$\mathbf{P} = [p_{ij}] \in \mathrm{IR}^{n\times n}, \tag{8}$$

gdzie: $p_{11} = \frac{1}{2}(\alpha + 1)$, $\frac{1}{2}(\alpha + 1) < p_{ii} < 1$, $i = 2, 3, \ldots, n$, $0 < p_{ij} < 1$, $i, j = 1, 2, \ldots, n$, $i \neq j$ oraz $\sum_{i=1}^{n} p_{ij} = 1$, $j = 1, 2, \ldots, n$ jest macierzą dodatnią i stochastyczną. Dlatego $|\lambda_i(P)| \leq 1$, $i = 1, 2, \ldots, n$. Stosując *Własność 2* w odniesieniu do macierzy (8) otrzymano $\omega = \min_{1\leq i\leq n} p_{ii} = \frac{1}{2}(\alpha+1)$. Tak więc $|\lambda_i(P)-\omega| \leq 1-\omega$ oraz $2\omega-1 \leq |\lambda_i(P)| \leq 1$. Przyjmując $\omega = \frac{1}{2}(\alpha + 1)$ autorzy otrzymali $\alpha \leq |\lambda_i(P)| \leq 1$, $i = 1, 2, \ldots, n$.

Podsumowując, Białas i Gurgul sformułowali wniosek, że druga wartość własna macierzy nakładów bieżących, w ogólnym przypadku nie maleje ze wzrostem wymiarów tej macierzy. Co więcej, dla każdej liczby rzeczywistej $\alpha \in (0, 1)$ można zbudować macierz dodatnią i stochastyczną o wymiarach $n \times n$, dla której moduły wszystkich wartości własnych będą większe lub równe α (niezależnie od wymiaru n).

4. UJĘCIE PROBABILISTYCZNE HIPOTEZY BRODY

Wyniki przedstawione wyżej oznaczają, że hipoteza mówiąca, że wraz ze wzrostem wymiaru macierzy nakładów opisującej gospodarkę wzrasta jej tempo powrotu do stanu równowagi nie jest prawdziwa dla każdego rodzaju macierzy. W szczególności nie jest prawdziwa dla pewnej grupy macierzy stochastycznych. Gdyby jednak spojrzeć na to zagadnienie od strony probabilistycznej, to okaże się, że wraz ze wzrostem n zmniejsza się istotnie prawdopodobieństwo wylosowania spośród wszystkich macierzy stochastycznych wymiaru $n \times n$ macierzy stochastycznej P posiadającej wymienione w pracy Białasa i Gurgula własności. To z kolei oznacza, że opisywana hipoteza może być rozważana jedynie na gruncie probabilistycznym.

Sytuacji, gdy elementy a_{ij} macierzy **A** są zmiennymi losowymi dotyczy treść artykułu Bidarda i Schattemana [2], w którym autorzy dowodzą następującego twierdzenia.

Twierdzenie 1 ([2], twierdzenie 1)

Niech elementy a_{ij} dodatniej macierzy kwadratowej **A** będą niezależnymi zmiennymi losowymi o tym samym rozkładzie o wartości oczekiwanej μ, posiadającymi skończone momenty do rzędu k, $k>2$. Wówczas iloraz $|\lambda_2|/|\lambda_1|$ dwóch największych wartości własnych macierzy **A** zmierza stochastycznie do 0, gdy n zmierza do nieskończoności. Dokładniej:

$$\lim_{n\to\infty} \text{Prob}\left\{\left|\lambda_2\right|/\left|\lambda_1\right| \le n^{-0,23}\right\} = 1 \tag{9}$$

Z tego twierdzenia wynika prawdziwość opisywanej hipotezy o szybkości powrotu do stanu równowagi dla macierzy o losowych i niezależnych elementach. Ponadto, Bidard i Schatteman pokazują tempo w jakim rozważany iloraz zmierza do zera. Z zaprezentowanego twierdzenia wynika, że iloraz modułów dwóch największych wartości własnych losowej macierzy **A** zmierza do zera w tempie $n^{-0,23}$. Jest to jednak mniej niż postulowane w pracy Brody [3] tempo $n^{-0,5}$. Zauważają to również sami autorzy sugerując, że w rzeczywistości tempo zbieżności do zera jest takie jak podpowiada intuicja, jednak zastosowane przez nich metody dowodu nie pozwalają na lepsze oszacowanie.

Problem ten rozstrzyga G.-S. Sun [11] przywołując twierdzenie uzyskane przez Goldberga i in. [4], które z jednej strony stanowi potwierdzenie hipotezy Brody dla przypadku macierzy losowej, a z drugiej daje odpowiednie oszacowanie tempa zmierzania modułu drugiej, wartości własnej do zera:

Twierdzenie 2 ([4] twierdzenie 1.1)

Niech $\mathbf{B} = [b_{ij}] \in \text{IR}^{n\times n}$ będzie macierzą losową, której elementy są niezależnymi zmiennymi losowymi o tym samym rozkładzie takimi, że $E(b_{ij}) = 1/n$ i $var(b_{ij}) = c/n^2$ (gdzie c jest pewną dodatnią stałą). Wówczas dla dowolnych $\varepsilon, p \in (0, 1)$ istnieje stała $N(\varepsilon, p)$ taka, że dla dowolnego $n > N(\varepsilon, p)$:

i) $\text{Prob}\,(|\lambda_1 - 1| < \varepsilon) \geq p$,
ii) $\text{Prob}\,(|\lambda_2| < \varepsilon) \geq p$.

Ponieważ w powyższym twierdzeniu p może być dowolnie bliskie 1, a ε może być dowolnie bliskie 0, to warunek *ii*) tezy oznacza, że druga co do modułu wartość własna takiej macierzy losowej zmierza stochastycznie do zera, podczas gdy największa wartość własna dąży do 1. Co więcej, Goldberg i in. wykazują, że dla dostatecznie dużych n, dla dowolnego $p \in (0, 1)$ istnieje stała $\alpha(p)$ taka, że

$$\text{Prob}\left(|\lambda_2| < \alpha(p)/\sqrt{n}\right) \geq p. \tag{10}$$

Zależność (10) oznacza, że modułdrugiej wartości własnej losowej macierzy **B** zmierza stochastycznie do zera w tempie $1/\sqrt{n}$ sugerowanym przez Brody.

Macierz losowa występująca w tezie twierdzenia 2 jest dosyć specyficzna, tzn. wartości oczekiwane jej elementów maleją wraz ze wzrostem wymiaru, podczas gdy, oryginalnie, hipoteza została sformułowana dla macierzy, której wartości oczekiwane wszystkich elementów są stałe.

Z jednej strony na macierz **B** można patrzeć jako na zaburzoną macierz stochastyczną, bo wartość oczekiwana sumy poszczególnych wierszy jest równa 1. Z drugiej strony tezę twierdzenia 2 można uogólnić na przypadek macierzy losowej, której elementy mają dowolną wspólną wartość oczekiwaną.

Jeżeli $\mathbf{A}_n = [a_{ij}] \in \text{IR}^{n\times n}$ jest macierzą losową, której elementy a_{ij} są niezależnymi zmiennymi losowymi o tym samym rozkładzie o wartości oczekiwanej μ i skończonej wariancji σ^2 to wówczas możemy zdefiniować macierz $\mathbf{B}_n = [b_{ij}]$ następująco: $b_{ij} = \dfrac{a_{ij}}{n\mu}$. Wówczas pomiędzy wartościami własnymi macierzy $\mathbf{A}_n$ i $\mathbf{B}_n$ zachodzi następująca zależność:

$$\lambda(\mathbf{A}_n) = n\mu\lambda(\mathbf{B}_n).$$

Z drugiej jednak strony mamy: $E(b_{ij}) = 1/n$ oraz $var(b_{ij}) = \dfrac{\sigma^2}{n^2\mu^2}$, czyli macierz $\mathbf{B}_n$ spełnia założenia wspomnianego powyżej twierdzenia 4.2. Oznacza to, że $\lambda_1(\mathbf{B}_n) \to 1$ i $\lambda_2(\mathbf{B}_n) \to 0$ stochastycznie. Z tego wynika również, że $\dfrac{|\lambda_2(\mathbf{A}_n)|}{|\lambda_1(\mathbf{A}_n)|} = \dfrac{|\lambda_2(\mathbf{B}_n)|}{|\lambda_1(\mathbf{B}_n)|}$ czyli iloraz modułów drugiej i pierwszej wartości własnej macierzy $\mathbf{A}_n$ zmierza stochastycznie do zera.

Tak więc, twierdzenie 2 potwierdza prawdziwość przypuszczenia Brody, że wraz ze wzrostem wymiaru macierzy losowej Leontiefa opisywana przez nią gospodarka szybciej powraca do stanu równowagi po wystąpieniu ewentualnego zaburzenia. Oczywiście dalej pozostaje pytanie czy rzeczywistą gospodarkę można opisać za pomocą macierzy, której wyrazy są niezależnymi zmiennymi losowymi.

Dalsze badania własności macierzy losowych pozwoliły uogólnić wyniki z twierdzenia 2 poprzez osłabienie założenia o niezależności wyrazów macierzy **B**.

Twierdzenie 3 ([5], twierdzenie 1.1)

Niech $0 < \delta < 1$ i $0 < p < 1$. Niech macierz $\mathbf{B} = [b_{ij}] \in \mathrm{IR}^{n\times n}$ będzie postaci: $b_{ij} = a_{ij} + 1/n$, gdzie:

i) elementy a_{ij} są zmiennymi losowymi, przy czym wiersze macierzy **A** są niezależnymi n – wymiarowymi wektorami losowymi,

ii) $E(a_{ij}) = 0$ dla $i, j = 1, ..., n$

iii) $var(a_{ij}) \leq c/n^2$ dla $i, j = 1, ..., n$

iv) $|cov(a_{ik}, a_{il})| \leq c/n^3$ dla $i, k, l = 1, ..., n$, $k \neq l$.

Wówczas istnieje stała $N(\delta, p)$ taka, że dla dowolnego $n > N(\delta, p)$ zachodzi

$$\mathrm{Prob}\left(\max_{i=2,...,n} |\lambda_i| \leq \delta\right) \geq p \tag{11}$$

Teza twierdzenia 3 oznacza, że wszystkie wartości własne z wyjątkiem największej mają z prawdopodobieństwem co najmniej p moduły mniejsze niż δ. Z dowolności wyboru δ i p oznacza to, że moduły tych wartości własnych zmierzają stochastycznie do 0. W szczególności, więc zmierza do 0 również druga wartość własna macierzy **B**.

W porównaniu do wcześniejszego, twierdzenie 3 dopuszcza występowanie zależności pomiędzy elementami jednego wiersza, przy czym wraz ze wzrostem wymiaru macierzy siła tej zależności słabnie.

5. TEORIA A RZECZYWISTOŚĆ

Hipoteza postawiona przez Brody, że wraz ze wzrostem liczby gałęzi w gospodarce szybciej powraca ona na ścieżkę równowagi wygląda bardzo obiecująco tak z teoretycznego, jak i z praktycznego punktu widzenia. Oznacza bowiem, że postępująca globalizacja, wzrost powiązań pomiędzy gospodarkami na świecie i tworzenie się wielkiej światowej sieci wzajemnych połączeń powinno zwiększać stabilność i umożliwiać szybszy powrót gospodarek na ścieżkę równowagi po wystąpieniu ewentualnego szoku (zaburzenia). Doświadczenia ostatnich lat nie potwierdzają jednak tego przypuszczenia. Wręcz przeciwnie, ogólnoświatowa gospodarka, poprzez powstałe powiązania stała się bardziej podatna na występujące zakłócenia. Rodzi się więc pytanie, czy rozważana hipoteza znajduje potwierdzenie w danych rzeczywistych, tzn. czy w przypadku analizy rzeczywistych macierzy nakładów bezpośrednich można stwierdzić, że wraz ze wzrostem liczby uwzględnionych gałęzi produkcji maleje iloraz modułów dwóch największych wartości własnych. Aby to sprawdzić zostały przeanalizowane dane dotyczące przepływów międzygałęziowych w 2005 roku państw członkowskich Unii Europejskiej. Badanie zostało przeprowadzone na podstawie tablic input-output opublikowanych przez Eurostat. Dane te zostały uzupełnione o tablice przepływów międzygałęziowych Wielkiej Brytanii opublikowane przez Office of National Statistics.

Zaletą rozważania danych opublikowanych przez Eurostat i ONS jest ich jednorodność, ponieważ wszystkie zostały opracowane w oparciu o jednolity system rachunków

narodowych zgodny z ESA'95. Rozważane tablice zawierają bilans przepływów międzygałęziowych w bieżących cenach bazowych dla produkcji krajowej w układzie 59x59 działów wyrażony w jednostkach monetarnych danego kraju[1]. Produkty zostały wyodrębnione zgodnie z klasyfikacją produktów i usług CPA 2002. Stosując ten sam system klasyfikacji dokonano agregacji danych tworząc symetryczne tablice przepływów międzygałęziowych o układzie 30×30 i 16×16 działów. Pozwoliło to zbadać empirycznie czy liczba wyodrębnionych gałęzi danej gospodarki (czyli wymiar tablicy input-output albo inaczej stopień ich agregacji) ma istotne znaczenie dla szybkości powrotu gospodarki na ścieżkę równowagi. Na podstawie tablic input-output obliczone zostały macierze współczynników materiałochłonności bezpośredniej, bo to o ich wartościach własnych mówi hipoteza Brody.

W tabeli 2 zaprezentowane zostały ilorazy modułów dwóch największych wartości własnych poszczególnych macierzy współczynników materiałochłonności. Nie potwierdzają one hipotezy Brody, że wraz ze wzrostem wymiaru macierzy **A** iloraz modułów dwóch największych wartości własnych maleje do zera, a tym samym gospodarka opisywana za pomocą takiej macierzy szybciej powraca na ścieżkę równowagi. W zdecydowanej większości przypadków widoczne jest coś wręcz przeciwnego – zwiększanie liczby rozważanych gałęzi produkcji powoduje wzrost wartości tego ilorazu. Jest tak w przypadku 17 spośród 22 rozważanych gospodarek. Ponadto tylko w przypadku Hiszpanii wartość ilorazu $|\lambda_2|/|\lambda_1|$ jest najmniejsza dla agregacji 59×59, podczas gdy dla 18 krajów iloraz ten przyjmuje wówczas wartość największą. Wzrost wartości rozważanego ilorazu wraz ze zwiększaniem się wymiaru tablicy input-output dobrze widoczny jest również w przypadku zarówno średniej, jak i mediany wyników uzyskanych dla poszczególnych krajów.

Dalsze zwiększanie wymiaru tablicy IO nie prowadzi do zmniejszenia wartości ilorazu. W przypadku danych dotyczących Wielkiej Brytanii, dla tablicy w układzie 108x108 działów, wartość ilorazu $|\lambda_2|/|\lambda_1|$ jest równa 0,698, czyli wciąż jest to więcej niż w agregacji 16×16 i 30×30.

Brak sugerowanej zależności pomiędzy ilorazem modułów wartości własnych a wymiarem macierzy nakładów bezpośrednich nie wydaje się czymś zaskakującym. Bowiem w rzeczywistości hipoteza w formie zaproponowanej przez Brody oznaczałaby, że wnioskowanie o szybkości powrotu do stanu równowagi danej gospodarki zależałoby istotnie od stopnia agregacji wykorzystanych danych. W zależności od liczby rozważanych gałęzi raz tempo to mogłoby być uznane za niskie, natomiast przy większej liczbie uwzględnionych gałęzi uzyskane wyniki wskazywałyby na możliwość szybkiego powrotu do stanu równowagi. Co więcej, prawdziwość tej hipotezy oznaczałaby, że stosując podział na dostatecznie dużą liczbę gałęzi, praktycznie każdą gospodarkę można by było uznać za bardzo szybko powracającą do stanu równowagi.

[1] Pierwotnie dane dotyczące gospodarki Wielkiej Brytanii uwzględniały podział na 108 grup produktów i usług, jednak dla celów porównawczych dokonano ich agregacji do 59 grup zgodnie z CPA 2002.

Tabela 2.

Ilorazy modułów dwóch największych wartości własnych macierzy współczynników nakładów bezpośrednich wybranych państw Unii Europejskiej

	Wymiar macierzy		
	n =16	n =30	n =59
Austria	0,76	0,76	0,84
Belgia	0,64	0,65	0,75
Czechy	0,55	0,64	0,96
Dania	0,49	0,77	0,73
Estonia	0,61	0,61	0,73
Finlandia	0,55	0,70	0,80
Francja	0,57	0,62	0,67
Grecja	0,42	0,65	0,77
Holandia	0,64	0,70	0,76
Hiszpania	0,73	0,74	0,71
Irlandia	0,75	0,82	0,88
Litwa	0,53	0,74	0,69
Niemcy	0,71	0,72	0,75
Polska	0,45	0,70	0,72
Portugalia	0,83	0,83	0,85
Rumunia	0,71	0,72	0,88
Słowacja	0,75	0,74	0,76
Słowenia	0,49	0,70	0,82
Szwecja	0,47	0,53	0,62
Węgry	0,52	0,70	0,71
Wielka Brytania	0,62	0,64	0,82
Włochy	0,53	0,58	0,56
średnia	0,61	0,69	0,76
Mediana	0,59	0,70	0,75

Źródło: obliczenia własne.

Jednak na kwestię szybkości powrotu gospodarki na ścieżkę równowagi można spojrzeć również z innej strony. Być może nie jest istotna liczba rozważanych gałęzi produkcji, czyli stopień agregacji tych samych danych, lecz znaczenie ma wielkość i różnorodność gałęzi samej gospodarki. Przypadek Unii Europejskiej, w której wys-

tępują silne powiązania gospodarek poszczególnych krajów członkowskich, umożliwia analizę również tego aspektu hipotezy Brody. Dodatkowo więc zbadane zostały macierze materiałochłonności bezpośredniej oparte na tablicach przepływów międzygałęziowych w roku 2005 strefy Euro oraz całej Unii Europejskiej. Wyniki tych obliczeń zostały zebrane w tabeli 3.

Tabela 3.
Ilorazy modułów dwóch największych wartości własnych macierzy współczynników nakładów strefy Euro i całej Unii Europejskiej

	Wymiar macierzy		
	n =16	n =30	n =59
Strefa Euro	0,56	0,59	0,60
Unia Europejska	0,49	0,57	0,52

Źródło: obliczenia własne.

Iloraz $|\lambda_2|/|\lambda_1|$ obliczony dla całej unii Europejskiej jest dla wszystkich rozważanych agregacji mniejszy niż dla strefy Euro. Co więcej, dla n =59 jest on niższy niż w przypadku jakiegokolwiek badanego państwa członkowskiego. Podobnie jest dla bardziej zagregowanych tablic: dla n =30 iloraz $|\lambda_2|/|\lambda_1|$ osiąga mniejszą wartość tylko w przypadku Szwecji, natomiast dla n =16 jest tak w przypadku Danii, Grecji, Polski i Szwecji. Te wyniki zdają się potwierdzać hipotezę, że większa gospodarka szybciej powraca do stanu równowagi. Tak więc liczba rozważanych gałęzi produkcji czyli stopień agregacji danych dotyczących przepływów międzygałęziowych nie ma istotnego znaczenia dla liczby iteracji niezbędnych do wyznaczenia wektora Frobeniusa macierzy **A**, a pośrednio szybkości powrotu gospodarki do stanu równowagi, lecz jej wielkość.

6. WNIOSKI

Dyskusja zapoczątkowana przez artykuł Brody [3] trwa już kilkanaście lat. Z przeprowadzonych tu rozważań wynika, że sformułowana w tym artykule hipoteza może być prawdziwa tylko w pewnych szczególnych wypadkach dotyczących macierzy losowych o specjalnych własnościach. Macierze nakładów bezpośrednich nie mają jednak tych własności. Ani kolumny, ani wiersze nie są ani losowe, ani niezależne. Każda gospodarka jest „systemem naczyń połączonych". Poszczególne kategorie jednych sektorów zależą od pozostałych sektorów. To przenosi się na powiązania pomiędzy kolumnami czy wierszami macierzy nakładów. Wyniki empiryczne zaprezentowane w tym artykule mogą być podstawą do przypuszczenia, że nie stopień agregacji decyduje o szybkości osiągnięcia stanu równowagi danej gospodarki, ale jej wielkość (mierzona np. wielkością PKB), struktura oraz siła wzajemnych powiązań pomiędzy poszczególnymi gałęziami.

AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

LITERATURA

[1] Białas S., Gurgul H., (1998), On a Hypothesis about the Second Eigenvalue of the Leontief Matrix, *Economic Systems Research-Journal of International Input-Output Association*, vol.10, no.3, 285-289.
[2] Bidard C., Schatteman T., (2001), The spectrum of random matrices, *Economic Systems Research-Journal of International Input-Output Association*, vol. 13, 289-298.
[3] Brody A., (1997), The Second Eigenvalue of the Leontief Matrix, *Economic Systems Research-Journal of International Input-Output Association*, vol. 9, no. 3, 253-258.
[4] Goldberg G., Okuney P., Neumann M., Schneider H., (2000), Distribution of subdominant eigenvalues of random matrices, *Methodol. Comput. Appl. Probab.*, 2, 137-151.
[5] Goldberg G., Neumann M., (2003), Distribution of Subdominant Eigenvalues of Matrices with Random Rows, *SIAM Journal on Matrix Analysis and Applications*, vol. 24, no. 3, 747-761.
[6] Gurgul H., Majdosz P., (2006), Interfund Linkages Analysis: The Case of the Polish Pension Fund Sector, *Economic Systems Research-Journal of International Input-Output Association*, vol. 18, no.1, 1-27.
[7] Leontief W. W., (1941), *The Structure of American Economy 1919-1939*; *An Empirical Application of Equilibrium Analysis*, New York.
[8] Marcus M., Minc H., (1964), *A Survey of Matrix Theory and Matrix Inequalites*, (Boston, Allyn and Bacon).
[9] Molnar G., Simonovits A., (1998), A Note on the Subdominant Eigenvalue of a Large Stochastic Matrix, *Economic Systems Research-Journal of International Input-Output Association*, vol.10, no.1, 79-82.
[10] Plich M., (2003), Budowa i zastosowanie wielosektorowych modeli ekonomiczno-ekologicznych, Wydawnictwo Uł (rozprawa habilitacyjna).
[11] Sun G.-Z., (2008), The First Two Eigenvalues of Large Random Matrices and Brody's Hypothesis on the Stability of Large Input-Output Systems, *Economic Systems Research-Journal of International Input-Output Association*, vol. 20, no. 4, 429-432.
[12] Tomaszewicz Ł., (1994), Metody analizy input-output, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1994.

HIPOTEZA BRODY W ŚWIETLE WYNIKÓW BADAŃ EMPIRYCZNYCH

S t r e s z c z e n i e

W tym artykule została przedstawiona dyskusja hipotezy Brody sformułowanej w 1997 roku, a dotyczącej szybkości zbieżności do stanu równowagi (proporcji produkcji wyznaczonych przez wektor własny Frobeniusa) dużych systemów input-output lub systemów von Neumanna w zależności od ich wymiarów. Brody napisał, iż im większa macierz nakładów **A**, to znaczy im więcej zawiera ona elementów (sektorów i branż) tym szybciej system, w którym występuje ta macierz jest zbieżny do stanu równowagi. Od tego czasu toczy się dyskusja nad matematycznymi aspektami tej hipotezy. W tym artykule został zamieszczony przegląd prac nawiązujących do rozważanej hipotezy. Część z nich dostarcza dowodów na prawdziwość hipotezy, część zaś dowodzi, że hipoteza na ogół nie jest prawdziwa. W artykule zostały wyróżnione dwa kierunki badań stochastyczny i deterministyczny. Tylko przy bardzo wyszukanych założeniach odnośnie losowości macierzy A zachodzi hipoteza Brody. Jednakowoż w ogólnym przypadku, jak pokazaliśmy w tym artykule, hipoteza Brody nie jest prawdziwa. Jest to wniosek z obliczeń dla różnych agregacji (16, 30 i 59) wykonanych dla całej Unii Europejskiej, strefy Euro i poszczególnych krajów członkowskich.

Zwiększenie stopnia agregacji nie spowodowało na ogół wzrostu stosunku drugiej do pierwszej wartości własnej macierzy **A**. Najmniejszy stosunek obu wartości własnych miał miejsce w przypadku całej Unii Europejskiej, potem strefy Euro i dalej poszczególnych państw. Stąd wysnuto wniosek, że szybkość zbieżności danej gospodarki do stanu równowagi nie zależy od liczby wyróżnionych w niej sektorów, lecz od jej wielkości i być może stopnia współzależności tych sektorów (duże gospodarki szybciej zmierzają do stanu równowagi niż mniejsze gospodarki).

Słowa kluczowe: modele input-output, równowaga, wartości własne, hipoteza Brody

BRODY'S HYPOTHESIS VERSUS RESULTS OF EMPIRICAL RESEARCH

Abstract

This paper presents discussion about the importance of degree of aggregation in input-output systems for speed of convergence to the equilibrium. The basis is the hypothesis stated by Brody (1997) that the greater the dimension of flow coefficients matrix **A** is, the faster the economy convergences to the equilibrium because the ratio between modulus of the subdominant and dominant eigenvalue declines to zero. Since then, several papers have been published to verify mathematical aspects of the hypothesis. The development of random matrices theory provided theorems that confirm the hypothesis in the case of i.i.d. elements of flow coefficients matrix. However, in the case of deterministic input-output table, there were constructed counterexamples where the ratio between subdominant and dominant eigenvalue does not decline when the size of the matrix increases to infinity and the degree of aggregation does not influence the speed of convergence to the equilibrium.

This paper verifies how the hypothesis fits to empirical data. Analysis of different aggregation of input-output tables for European Union, Euro zone and several European countries shows that increasing the number of branches in IO table does not reduce the ratio of modulus of subdominant and dominant eigenvalues of the flow coefficients matrix A. Hence, the speed of convergence to the equilibrium does not depend on the number of sectors in IO table but rather on the size of economy and relationships between its sectors. (i.e. large economies converge faster to equilibrium than small economies).

Key words: input-output models, equlibrium, eigenvalues, Brody's conjecture

PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012

MARCIN ŁUPIŃSKI

SHORT-TERM FORECASTING AND COMPOSITE INDICATORS CONSTRUCTION WITH HELP OF DYNAMIC FACTOR MODELS HANDLING MIXED FREQUENCIES DATA WITH RAGGED EDGES

This paper aims at presenting practical applications of latent variable extraction method based on second generation dynamic factor models, which use modified Kalman filter combined with Maximum Likehood Method and can be applied for time series with mixed frequencies (mainly monthly and quarterly) and unbalanced beginning and the end of the data sample (ragged edges). These applications embrace nowcasting, short-term forecasting of Polish GDP and construction of composite coincident indicator of economic activity in Poland. Presented approach adopts the idea of short-term forecasting used by Camacho and Perez-Quirioz in Banco de Espana and concept of Arouba, Diebold and Scotti index compiled in the FRB of Philadelphia. According to the author's knowledge, it is the first such adaptation for Central and Eastern Europe country. Quality of the nowcasts/forecasts obtained with these models are compared with standard methods used for short-term forecasting with series of statistical tests in the pseudo real-time forecasting exercise. Moreover described method is applied for construction of composite coincident indicator of economic activity in Polish economy. This newly-created coincident indicator is compared with first generation coincident indicator, based on standard dynamic factor model (Stock and Watson) approach, which has been computed by the author for Polish economy since 2006.

Keywords: short-term forecasting, coincident indicators, factor models, mixed frequencies, ragged edges

1. INTRODUCTION

Since construction of the first set of composite economic indicators by Burns and Mitchell in the first half of 20th century these indicators have proved crucial in determining and explaining current and future economic situation of particular economies and international organizations. Decision makers and agents observe them to have clear synthetic picture of the state of the economy and to derive its future perspectives. Broad group of the most popular coincident indicators prepared by international institutions such as The Conference Board or OECD is based on heuristic methods examining common properties of time series with help of simple measures like cross-correlation or coherence. Simultaneously there is constantly growing number of institutions (CEPR, Banco de Espana) which use dynamic factor econometric framework proposed by Stock and Watson [8] (S&W, 1989) or its generalized version prepared by Forni, Hallin, Lippi and Reichlin [7] (FHLR, 2001). S&W coincident indicator is estimated in the time domain on a small set of selected time series, on the other hand FHLR model operates on the broad cross-section of economic data (up to few thousands of

series, which gives possibility of exploitation of the model asymptotic features) and use spectral representation during estimation process. Both of them are based on strict mathematical structure of the model, which removes bias of arbitrary judgment of economic conditions by experts, the main disadvantage of heuristic methods. However similarly to expert approach S&W and FHLR frameworks assume that time series, which are coincident indicators' components, should have the same length and the same frequency. This feature generates serious problems for analysts trying to use constructed coincident indicators for their practical application like nowcasting and forecasting because they are not able to enclose recent information due to delays of some time series publication (e.g. for Polish GDP this delay lasts up to 60 days after the end of reference period). Moreover, according to the paper of Boivin and Ng [3] (2006) and surveys performed by Lupinski [15] (2006) generalized dynamic factor models do not outperform their smaller counterparts, although they have quite complex structures which are not so easy to understand by analyst, who are not interested with their technical interior.

Dealing with dynamic factor models' drawbacks directed some researchers like Shumway and Stoffer [19] (1982) towards some accompanying algorithms (in their case Expectation Maximization /EM/ algorithm) to replace missing observations with their optimal extrapolations. Unfortunately incorporating these algorithms into base econometric state space models generated often additional problems like those connected with identification of matrices linking latent components with observed variables. Alternative approach was proposed by Mariano and Murasawa [10] (2003) who directly incorporated missing data and mixed frequency handling into their econometric framework. They used properties of state space representation and Maximum Likehood Estimation (MLE) with missing observations replaced with draws from selected probability distribution (most often from Gaussian one). Mariano and Murasawa model inherited all advantages of dynamic factor models, beside that it gave transparent way of coincident indicator estimation based on well-established mathematical background. Their framework was applied in practice by Camacho and Perez-Quirios [6] (2009) to build coincident indicator used by them to compute nowcasts and forecasts of Euro Area GDP and validate their power in the pseudo real-time exercise. Similar indicator for American Economy was proposed by Arouba, Diebold and Scotti [1] (2009). The small scale dynamic factor model with missing data and mixed frequencies handling proposed by Mariano and Murasawa is applied in this paper for estimation of new coincident indicator of Polish GDP and evaluation of its statistical features. Achieved results are compared with Stock and Watson style coincident indicator that has been computed by the author for Polish economy since 2006. It is worth to mention that before presented survey was started, complete estimation framework was built in MATLAB, which was used to semi-automatize procedure of coincident indicator building and its nowcasting and forecasting power validation. It is also first survey known to the author, which uses real-time database of Polish time series applied to do historical analysis of nowcasting and forecasting performance with coincident indicators applied.

This article is organized as follows. The next section outlines common mathematical background used by dynamic factor models. In the third section second generation dynamic models (with ragged edges and mixed frequencies handling) are presented. Fourth section is devoted to description of economic data and procedures used in the coincident indicator estimation process. Description of recursive procedure of coincident indicators estimation and validation is enclosed in the fifth section. Sixth section was prepared for presentation of achieved results and their interpretation. Last section concludes the survey.

2. DYNAMIC FACTOR MODELS – MATHEMATICAL BACKGROUND

In this section mathematical background of econometric models used in the article is presented. Investigation of these model's structures and mechanisms applied for their estimation is crucial for understanding their econometric characteristics as well as their advantages and drawbacks. For example issues connected with model's identification implicate problems while modeling variables with strong stochastic trend, which are the subject of interest in this article. At the beginning general idea of dynamic models with unobserved components is presented.

Let's assume that there is a time sequence (of length T) of vector's elements (of size n)

$$\tilde{Y} = [\tilde{Y}_1, \tilde{Y}_2, ..., \tilde{Y}_T] \tag{1}$$

which groups set of main macroeconomic variables describing current state of the economy. These variables' observations can be perceived as realizations of multivariate stochastic process Y, which is modelled with proposed framework. Let's further assume that this stochastic process is influenced by unobserved (latent) variables. The goal of the framework is to estimate values of these variables. However because of their latent character it is necessary to go beyond traditional econometric framework. One of the most popular ways of dealing with this issue is to write dependencies determined by the model in the state space representations. This representation consists of two blocks of equations:

- Measurement block describing relations between unobserved components and sequences of stochastic processes' realizations (elements of the matrix Y).
- Transition block, which determines dynamics of latent variables in the form of difference equations' set.

Simple notation of dynamic model with r unobserved components and exogenous economic variables (X, of size kxT) written in the state space form is presented below:

$$\begin{aligned} y_t &= Ax_t + Hh_t + w_t \\ h_t &= Fh_{t-1} + v_t \end{aligned} \tag{2}$$

where A, H, F are (nxk), (nxr) and (rxr) matrices respectively, w_t and v_t represent i.i.d errors drawn from Gaussian distributions with covariance matrices R and Q respectively ($E[w_t, v_\tau] = 0$ for each t, τ =1,2,...,T).

For example popular Box and Jenkins's ARIMA framework augmented with unobserved component, used by Stock and Watson [18] (1998) to decompose US GDP into permanent and transitory part:

$$\begin{cases} y_t = y_{1,t} + y_{2,t} \\ y_{1,t} = \delta + y_{1,t-1} + w_t \\ y_{2,t} = \phi_1 y_{2,t-1} + \phi_2 y_{2,t-2} + v_t \end{cases} \tag{3}$$

where $w_t \backsim i.i.d\, N(0, \sigma_{w,t}), v_t \backsim i.i.d\, N(0, \sigma_{v,t})$, $E[w_t, v_\tau] = 0$ for t, τ =1,2,...,T, can be noted in the state space form as follows:

$$\begin{gathered} y_t = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{2,t-1} \end{bmatrix} \\ \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{2,t-1} \end{bmatrix} = \begin{bmatrix} \delta \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & \phi_1 & \phi_2 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} w_t \\ v_t \\ 0 \end{bmatrix} \end{gathered} \tag{4}$$

Let's go back to the general model described with the set of equations (2). The goal of an econometrician is to estimate values of unobserved variables grouped in the vector h_t in all periods t = 1,2,...,T, having at disposal appropriate set of information (data). The estimation procedure can be started once values of all model's parameters (elements of the matrices A, H, F, R, Q) are known. Unfortunately they are not available at the beginning of the modeling process. However mentioned econometrician can determine them using Maximum Likehood Estimation (MLE) method. Description of the MLE's application is presented further into this paper, at this stage assumption is made that all parameters of the model are available. The procedure used for estimating optimal values of unobserved variables in dynamic models with known initial parameters was proposed originally by Kalman in 1960. His recursive procedure yields estimates with lowest Mean Squared Error (MSE).

During each stage of Kalman's procedure (indexed with time subscript t) conditional expectations of unobserved components observations vector are formulated:

$$h_{t|t-1} = E[h_t \mid z_{t-1}] \tag{5}$$

having at disposal information set $z_{t-1} = (y_{t-1}, y_{t-2}, ..., y_1, x_t, x_{t-1}, x_{t-2}, ..., , x_1)$ / x_t is exogenous variable therefore it does not affect values of $h_{t+s}, s = 0, 1, ..., T - t$ /. Conditional expectation of h_t are associated with matrix of estimation squared errors:

$$P_{t|t-1} = E[(h_t - h_{t|t-1})(h_t - h_{t|t-1})'] \tag{6}$$

At the beginning of the estimation process no value of h_t is available. Initial unobserved component vector ($h_{0|0}$) is derived from transition block of set (2) after applying

unconditional expectation operator to both sides of this expression and assuming that h_t process is stationary which means that its unconditional expectation is constant across time ($E[h_t] = E[h_{t-1}]$)

$$(F - I)E[h_t] = 0 \Rightarrow h_{0|0} = 0 \tag{7}$$

Associated initial value of estimation error is the solution of equation describing h_t's covariance

$$E[h_t h_t'] = E[(Fh_{t-1} + v_{t-1})(h_{t-1}'F' + v_{t-1}')] = FE[h_{t-1}h_{t-1}']F' + E[v_{t-1}v_{t-1}'] \tag{8}$$

with applied stationary covariance matrix condition ($E[h_t h_t'] = E[h_{t-1}h_{t-1}'] = \Sigma$):

$$P_{1|0} = vec(\Sigma) = (I - F \otimes F)^{-1} vec(Q) \tag{9}$$

where $\otimes$ is Kronecker product of two matrices and *vec* vectorization operator.

After obtaining initial values of h_t and P_t during each iteration of the Kalman's procedure one step ahead conditional prediction of h_t is computed. The following expression:

$$\hat{h}_{t|t-1} = E[h_t \mid z_{t-1}] = F\hat{h}_{t-1|t-1} \tag{10}$$

is used there to evaluate error associated with this prediction. Part of the derivation shown in (8) can be used to gain:

$$P_{t|t-1} = FP_{t-1|t-1}F' + Q \tag{11}$$

Predicted value of unobserved component can be entered into measurement equation with conditional expectation operator applied to its both sides

$$\hat{y}_{t|t-1} = E[y_t \mid z_{t-1}] = Ax_t + H\hat{h}_{t|t-1} \tag{12}$$

to compute y_t's forecast. Observed variables' prediction error ($\xi_{t|t-1} = \tilde{y}_{t|t-1} - Ax_t - H\hat{h}_{t|t-1}$) has in this case the following covariance matrix:

$$\Xi = E[(\tilde{y}_t - \hat{y}_{t|t})(\tilde{y}_t - \hat{y}_{t|t})'] = HP_{t|t-1}H' + R \tag{13}$$

This part of the Kalman's procedure is called prediction phase.

The last equation allows to incorporate present information about behavior of observed variables in the forecast of h_t:

$$\begin{aligned} h_{t|t} = E[h_t \mid z_t] &= \hat{h}_{t|t-1} + E[(h_t - \hat{h}_{t|t-1})(\tilde{y}_t - \hat{y}_{t|t-1})']E[(\tilde{y}_t - \hat{y}_{t|t-1})(\tilde{y}_t - \hat{y}_{t|t-1})']^{-1}(\tilde{y}_t - \hat{y}_{t|t-1}) = \\ &= \hat{h}_{t|t-1} + P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1}H(\tilde{y}_{t|t-1} - Ax_t - H\hat{h}_{t|t-1}) \end{aligned} \tag{14}$$

Estimation error's covariance matrix can be updated as well:

$$P_{t|t} = P_{t|t-1} + P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1}HP_{t|t-1} \tag{15}$$

Expression enclosed in the last two terms $(P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1}HP_{t|t-1})$ is of special interest because it allows to separate the effect of update process on forecast of h_t and P_t. It is called Kalman gain. Computing it completes update phase of Kalman filter and thus ends one iteration of the whole filtering procedure.

To recapitulate, one particular step of Kalman filter procedure, summarized with six equations, is presented below:

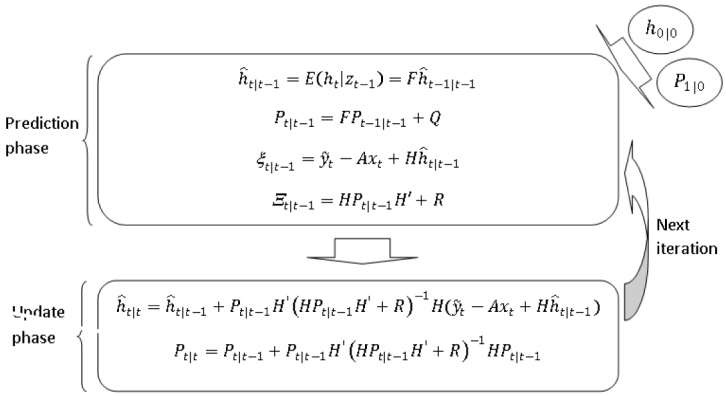


Figure 1.

Kalman filter offers optimal estimation of unobserved variables, taking into account that all structural parameters of the model are identified. As it was mentioned before parameters are unknown at the beginning of the estimation process. Hence MLE framework is used to get their estimates. In the second part of this section MLE framework customized to Kalman filter needs is presented.

Any econometric model specifies joint density function for set of input data $(\tilde{y})$ given m model parameters (θ):

$$p(\tilde{y} \mid \theta) \tag{16}$$

Because this parameters' set is not known, joint density function for parameters given available data (likehood function) is formulated:

$$L(\theta \mid \tilde{y}) \tag{17}$$

and then used to find values for which this function reaches its maximum:

$$\hat{\theta}_{MLE} = argmax(L(\theta \mid \tilde{y})) \tag{18}$$

For time series with independent observations it is easy to identify likehood function:

$$L(\theta \mid \tilde{y}) = \prod_{t=1}^{T} p(\tilde{y}_t \mid \theta) \tag{19}$$

where $p(\tilde{y}_t \mid \theta)$ is marginal density function for observation t.

In the case of dependent observations situation can get complicated as factors of (19) are replaced by conditional densities (conditional on information gathered in sequence $\tilde{y}_{t-1}, \tilde{y}_{t-2}, ..., \tilde{y}_1$)

$$L(\theta \mid \tilde{y}) = \prod_{t=2}^{T} p(\tilde{y}_t \mid \tilde{y}_{t-1}, \theta) p(\tilde{y}_1 \mid \theta) \tag{20}$$

which are often not so easy to identify. However in the case of models for which it can be assumed, that observations of input data are drawn from normal distribution. This problem can be solved by writing likehood function as a multivariate normal distribution and then applying triangular factorization to its covariance matrix. This allows to write it as a function of model prediction errors (ξ_t) and their covariances (Ξ_t):

$$L(\theta \mid \tilde{y}) = \prod_{t=1}^{T} \frac{1}{\sqrt{(2\pi)^n det(\Xi_t)}} exp(-\frac{1}{2} \xi_t' \Xi_t^{-1} \xi_t) \tag{21}$$

which can be easily obtained from models estimated with Kalman filter because ξ_t and Ξ_t are computed as "byproducts" of this procedure. In practice natural logarithm of likehood function (which gives the same results because natural log is strictly monotonic function) is very often maximized instead of its original version. Computing logarithm of the function simplifies computations and allows to specify directly asymptotic covariance matrix of the parameters estimator $\hat{\theta}_{MLE}$. This matrix is determined by applying Cramer-Rao inequality, which states that difference between parameters estimator covariance matrix and inversion of information matrix $I(\theta)$ is positive semidefinite.

$$Cov(\theta) - I(\theta)^{-1} \tag{22}$$

Hence the inverse of information matrix is lower bound of covariance matrix. Information matrix is defined as negative expected value of Hessian of loglikehood function at $\hat{\theta}_{MLE}$:

$$I(\theta) = -E\left[\frac{\partial^2 lnL(\theta \mid \tilde{y})}{\partial\theta\partial\theta'}\right] \tag{23}$$

Maximum log-likehood estimator of vector has asymptotic normal distribution:

$$\sqrt{T}(\hat{\theta}_{MLE} - \theta) \rightarrow N(0, \hat{H}^{-1}) \tag{24}$$

where $\hat{H} = lim\frac{1}{T} I(\theta)$, so it is consistent and asymptotically efficient. Hence, to compute covariance matrix of MLE estimator the following formula can be used:

$$\left[-\frac{\partial^2 lnL(\theta \mid \tilde{y})}{\partial\theta\partial\theta'}\right]^{-1} \tag{25}$$

at $\theta = \hat{\theta}_{MLE}$

Very often some additional constraints on chosen elements of likehood function parameters' vector are set. For example when stationary series is modeled with AR(2) process, it is necessary to assume that roots of lag polynomial $1 - L\psi_1 - L^2\psi_2 = 0$ associated with this process lie inside the unit circle, which yields:

$$s_1 = \frac{1}{1 + |\psi_1|},\ s_2 = \frac{1}{1 + |\psi_2|} \tag{26}$$

$$\psi_1' = s_1 + s_2,\ \psi_2' = -s_1 s_2 \tag{27}$$

Imposing this kind of constrains can be applied by computing appropriate continuous function $g(\cdot)$ subject to parameters achieved in the particular step of MLE procedure ($\theta = g(\psi)$). In such case constrained MLE optimization of θ is equivalent to unconstrained optimization of ψ:

$$lnL(\theta \mid \tilde{y}) = lnL(g(\psi) \mid \tilde{y}) = lnL(\psi|\tilde{y}) \tag{28}$$

Although the formula of likehood function doesn't change, in the case of constrained optimization, computing MLE estimator error covariance matrix for $\hat{\theta} = g(\hat{\psi}_{MLE})$ requires quadratic form of the matrix (25) with gradient of the constraint function $g(\cdot)$:

$$\left[\frac{\partial g(\hat{\psi}_{MLE})}{\partial \hat{\psi}_{MLE}}\right]\left[-\frac{\partial^2 lnL(\hat{\psi}_{MLE} \mid \tilde{y})}{\partial \hat{\psi}_{MLE} \partial \hat{\psi}'_{MLE}}\right]^{-1}\left[\frac{\partial g(\hat{\psi}_{MLE})}{\partial \hat{\psi}_{MLE}}\right]' \tag{29}$$

Combination of model's parameters MLE procedure and Kalman filer estimation of unobserved variable is presented in the following figure.

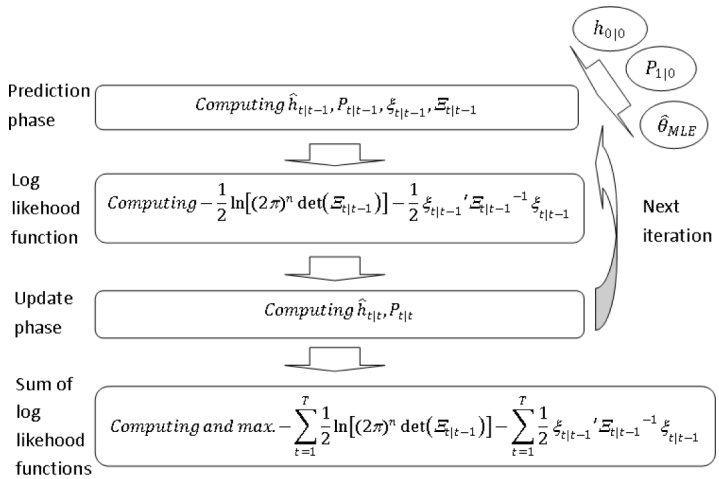


Figure 2.

3. APPLYING SECOND GENERATION DYNAMIC FACTOR MODEL FOR GDP COINCIDENT INDICATOR CONSTRUCTION, GDP NOWCAST AND SHORT-TERM FORECAST

Before survey described in this paper was initiated, the National Bank of Poland (NBP) computed a set of conincident indicators based on three schemes: The Conference Board/OECD, Stock and Watson and Forni Hallin Lippi and Reichlin. All above mentioned frameworks offer mechanisms, which allow to capture common dynamics of GDP and other key macroeconomic variables which influence current state of Polish economy. The first one is the simplest – based on heuristics. Second one introduces simple dynamic factor models which offers transparent dependencies between observed series and latent components so they could be easily understood by decision makers and analysts. The third one is the most sophisticated and demands frequency domain methods for estimation and large datasets to exploit its asymptotic features. However, as it was mentioned before, in empirical exercise it doesn't perform better than its simpler competitors. Having this situation in mind the NBP turned its attention to small dynamic factor models with possibility of frequency mixing and handling ragged edges which are relatively easy to estimate and give easily interpretable results.

The NBP's base model prepared in accordance with Stock and Watson approach excludes possibility of mixing frequencies and endogenous handling of input time series' missing observations. Moreover it was designed to model growth of explanatory variables (due to stationarity requirements) and assumes lack of correlation between common component (interpreted as coincident indicator) and idiosyncratic factors (factors associated with particular input time series). Structure of described model shows the block of equations below:

$$\begin{aligned} & y_t = \delta + \gamma h_t + i_t \\ & \Phi(L) h_t = v_{h,t} \\ & \Theta(L) i_t = v_{i,t} \\ & \begin{bmatrix} v_{h,t} \\ v_{i,t} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \end{bmatrix}, \begin{bmatrix} \sigma_h & 0 \\ 0 & \Sigma_i \end{bmatrix} \right) \end{aligned} \tag{30}$$

where h_t is common latent component, i_t vector of idiosyncratic components, γ vector of coefficients quantifying dependencies between latent variable(s) and input time series, $\Phi(L)$, $\Theta(L)$ lag polynomials of common and idiosyncratic components accordingly, σ_h common component's variance and Σ_i covariance matrix of idiosyncratic component.

Second generation dynamic model inherits many features of its ascendant, but it allows to introduce mixed frequencies and deal with missing observation within and at the beginning and the end of the sample (ragged edges). First extension is based on a simple approximation of arithmetic average by geometric one and should be only applied to flow data. For monthly and quarterly data (observed month on month and

quarter on quarter respectively) dependency between dlog of quarterly and monthly data is derived in the following way:

$$\begin{aligned}
&\tilde{Y}_t = 3\frac{\tilde{X}_t + \tilde{X}_{t-1} + \tilde{X}_{t-2}}{3} \backsimeq 3(\tilde{X}_t\tilde{X}_{t-1}\tilde{X}_{t-2})^{1/3} \\
&ln(\tilde{Y}_t) = ln(3) + \frac{1}{3}ln(\tilde{X})_t + \frac{1}{3}ln(\tilde{X}_{t-1}) + \frac{1}{3}ln(\tilde{X}_{t-2}) \\
&ln(\tilde{Y}_t) - ln(\tilde{Y}_{t-3}) = \frac{1}{3}(ln\tilde{X}_t - \frac{1}{3}ln\tilde{X}_{t-3}) + \frac{1}{3}(ln\tilde{X}_{t-1} - \frac{1}{3}ln\tilde{X}_{t-4}) + ... \\
&\tilde{y}_t^{\triangle q} = \frac{1}{3}\tilde{x}_t^{\triangle m} + \frac{2}{3}\tilde{x}_{t-1}^{\triangle m} + \tilde{x}_{t-2}^{\triangle m} + \frac{2}{3}\tilde{x}_{t-3}^{\triangle m} + \frac{1}{3}\tilde{x}_{t-4}^{\triangle m}
\end{aligned} \tag{31}$$

where $\tilde{y}_t^{\triangle q} = \tilde{Y}_t - \tilde{Y}_{t-3}$, $\tilde{x}_t^{\triangle m} = \tilde{X}_t - \tilde{X}_{t-1}$.

Error associated with proposed transformation is depicted below:

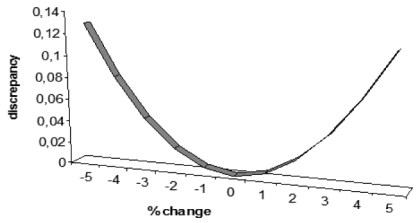


Figure 3.

Schema of the model with mixed frequencies and all data available in the sample has the structure of:

$$\begin{aligned}
&\begin{bmatrix} y_{1,t} \\ y_{j,t} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_j \end{bmatrix} + \begin{bmatrix} \gamma_1(\frac{1}{3}h_t + \frac{2}{3}h_{t-1} + h_{t-2} + \frac{2}{3}h_{t-3} + \frac{1}{3}h_{t-4}) \\ \gamma_j h_t \end{bmatrix} + \\
&+ \begin{bmatrix} \frac{1}{3}i_{1,t} + \frac{2}{3}i_{1,t-1} + i_{1,t-2} + \frac{2}{3}i_{1,t-3} + \frac{1}{3}i_{1,t-4} \\ i_{j,t} \end{bmatrix} \\
&\Phi(L)h_t = v_{h,t} \\
&\Theta(L)i_{j,t} = v_{i,j,t} \\
&\begin{bmatrix} v_{h,t} \\ v_{i,j,t} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \end{bmatrix}, \begin{bmatrix} \sigma_h & 0 \\ 0 & \Sigma_i \end{bmatrix} \right)
\end{aligned} \tag{32}$$

The simplest state space specification of the model with dynamics of h_t and $i_{j,1}$ represented with AR(1) processes (which in fact turns out to be quite robust during

coincident index estimation) has the following form:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ y_{4,t} \\ y_{5,t} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \end{bmatrix} + \begin{bmatrix} \frac{1}{3}\gamma_1 & \frac{2}{3}\gamma_1 & \gamma_1 & \frac{2}{3}\gamma_1 & \frac{1}{3}\gamma_1 & \frac{1}{3} & \frac{2}{3} & 1 & \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \gamma_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \gamma_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \gamma_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \gamma_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} h_{t-1} \\ h_{t-2} \\ h_{t-3} \\ h_{t-4} \\ h_{t-5} \\ i_{1,t-1} \\ i_{1,t-2} \\ i_{1,t-3} \\ i_{1,t-4} \\ i_{1,t-5} \\ i_{2,t-1} \\ i_{3,t-1} \\ i_{4,t-1} \\ i_{5,t-1} \end{bmatrix} + \begin{bmatrix} v_{h,t} \\ 0 \\ 0 \\ 0 \\ 0 \\ v_{i,1,t} \\ 0 \\ 0 \\ 0 \\ 0 \\ v_{i,2,t} \\ v_{i,3,t} \\ v_{i,4,t} \\ v_{i,5,t} \end{bmatrix} \tag{33}$$

$$\begin{bmatrix} h_t \\ h_{t-1} \\ h_{t-2} \\ h_{t-3} \\ h_{t-4} \\ i_{1,t} \\ i_{1,t-1} \\ i_{1,t-2} \\ i_{1,t-3} \\ i_{1,t-4} \\ i_{2,t} \\ i_{3,t} \\ i_{4,t} \\ i_{5,t} \end{bmatrix} = \begin{bmatrix} \phi & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_5 \end{bmatrix} \begin{bmatrix} h_{t-1} \\ h_{t-2} \\ h_{t-3} \\ h_{t-4} \\ h_{t-5} \\ i_{1,t-1} \\ i_{1,t-2} \\ i_{1,t-3} \\ i_{1,t-4} \\ i_{1,t-5} \\ i_{2,t-1} \\ i_{3,t-1} \\ i_{4,t-1} \\ i_{5,t-1} \end{bmatrix} \tag{34}$$

For presented model input data set consists of one quarterly indicator ($y_{1,t}$) and four monthly indicators ($y_{j,t}$, $j = 2, 3, ..., 5$).

Second very useful feature of Mariano and Murasawa model is its embedded mechanism of handling missing observations. Solution of this problem is the same without distinction to gaps within sample and at the beginning or at the end of time series' set. Observations which are not available are replaced with draws from Gaussian distribution with parameters equal to original variables estimators:

$$\overline{y}_{j,t} = \begin{cases} \tilde{y}_{j,t} \; if \; \tilde{y}_{j,t} \; is \; available \\ w_{j,t} \; if \; \tilde{y}_{j,t} \; is \; not \; available \end{cases} \tag{35}$$

where $w_{j,t} \backsim N(\mu, \sigma)$, μ, σ are estimators of the first two moments of original series.

Moreover missing data handling requires change of model specification (example for handling missing values for first variable):

$$\begin{gathered} \begin{bmatrix} y_{1,t} \\ y_{j,t} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_j \end{bmatrix} + \begin{bmatrix} \gamma_1(\frac{1}{3}h_t + \frac{2}{3}h_{t-1} + h_{t-2} + \frac{2}{3}h_{t-3} + \frac{1}{3}h_{t-4}) \\ \gamma_j h_t \end{bmatrix} + \\ + \begin{bmatrix} \frac{1}{3}i_{1,t} + \frac{2}{3}i_{1,t-1} + i_{1,t-2} + \frac{2}{3}i_{1,t-3} + \frac{1}{3}i_{1,t-4} \\ i_{j,t} \end{bmatrix} \\ if \; \tilde{y}_{j,t} is \; available \\ \begin{bmatrix} y_{1,t} \\ y_{j,t} \end{bmatrix} = \begin{bmatrix} 0 \\ \delta_j \end{bmatrix} + \begin{bmatrix} 0 \\ \gamma_j h_t \end{bmatrix} + \begin{bmatrix} 0 \\ i_{j,t} \end{bmatrix} + \begin{bmatrix} w_{1,t} \\ 0 \end{bmatrix} if \; \tilde{y}_{j,t} is \; not \; available \\ w_{1,t} = N(\delta_1, \sigma_1) \\ \ldots \end{gathered} \tag{36}$$

Estimation of dynamic factor models on input variables' growth is connected with identification issue. According to the last model first moment of $y_{j,t}$ ($E[y_{j,t}]$) consists of two parameters, given particular value of γ_j:

$$E[y_{j,t}] = \delta_j + k * \gamma_j E[h_t] \tag{37}$$

where k = 3 for j =1 and k=1 otherwise. Due to this feature $E[y_{j,t}]$ can be generated by unlimited number of δ_j and $E[h_t]$. To solve this problem each particular input variable is demeaned. Hence deviations of the unobserved factors from their means are modeled ($h_t^{wm} = h_t - E[h_t]$). This solution however brings the following question: how should this mean be restored when the factor with trend is needed? The answer can be found in the output of basic Kalman filter. It gives estimate of h_t as a linear projection of concurrent and lagged data entered into the model. The relationships between tem can be noted with lag polynomial:

$$h_{t|t} = W(L)y_t \tag{38}$$

$$h_{t|t}^{wm} = W(L)y_t^{wm} \tag{39}$$

Deviations from mean can be determined with help of steady state Kalman gain computed in practice as Kalman gain for $t \to T$:

$$W(L) = (I - (I - KH)FL)^{-1}K \tag{40}$$

Finally unobserved component forecast can be restored using expectations of (38):

$$E[h_t] = W(L)E[y_t] \Rightarrow mean(h_t) = W(1)\tilde{y}_t \tag{41}$$

This solution however has some drawbacks which should be considered while applying the model to real data. Using one estimation of latent component's mean doesn't allow to model trends changing across time (stochastic), which are very common macroeconomic time series.

4. STATISTICAL DATA AND PROCEDURES USED IN MODELS' ESTIMATION

Estimation of new coincident indicator of Polish GDP and validation of its nowcasting and forecasting power in a pseudo real-time exercise required creation of a database with contents allowing simulation of historical inflow of observations of component time series. Time span for series enclosed in this database was set from 1996:M1 to 2010:M5 (date of Polish GDP publication for 2010:Q1). Selection of this database contents was started from determining reference series. For this role GDP in average constant prices from 2000 was chosen. Moreover set of potential coincident indicator's components were gathered (all with monthly and quarterly frequency): hard data such as Building Permits, Export, Employment, Import, Industrial Production Index, Production of Cement/Coal, Registered Cars, Retail Trade Turnover Index and soft indicators from the group of industrial confidence indicators and consumer sentiment indicators. PMI and financial indicators such as 3M Interbank Interest Rate Spread, Broad Money (M3) and Index of Warsaw Stock Exchange (WIG) were included into potential set of indicator components as well. Main source of mentioned time series was OECD MEI database, some data came also from other sources as ECB and Eurostat statistical data warehouses.

All procedures used in this survey were written in Matlab. Stock and Watson and Mariano and Murasawa models' codes were implemented by the author. Chistiano and Fitzgerald filtering procedure was upgraded version of the code prepared by Kowal [14] (2005), Bry-Boschan business cycle dating routines were taken from the webpage of Inklaar [12]. The author created also consistent framework which allowed him to semi-automati ze substantial parts of pseudo real-time nowcasting, forecasting and dating exercise.

5. BUILDING COINCIDENT INDICATORS AND EVALUATING THEIR NOWCASTING, FORECASTING AND BUSINESS CYCLE DATING ACCURACY

In this section the procedures of coincident indicators building and using them for nowcasting/forecasting GDP are presented. Procedures applied for indicators' creation are based on two frameworks described in previous section. The first indicator is computed with help of Stock and Watson (S&W) model and the second one with schema proposed by Mariano and Murasawa (M&M). Then both of them are used in pseudo real-time exercise to check their usefulness for nowcasting and forecasting. Finally possibility of business cycles dating is checked as well.

Procedure of coincident indicator building was initiated with time series selection. Dependencies among transformed time series (first differences of logs) were checked simultaneously in time (with cross-correlation) and frequency domain (with coherence). Beside statistical testing, economic importance of each potential indicator's component was taken into consideration. During this stage four indicators were chosen: Industrial Production Index (IPI), Retail Trade Turnover Index (RTTI), Export (EXP) and Import (IMP). This selection of time series allows to monitor process of GDP creation (IPI) and distribution (EXP, IMP) and take a look at demand side of Polish economy (RTTI). Moreover inclusion of two external trade time series gave possibility of quantifying influence of global economic situation on Polish small open economy.

In the second phase S&W and M&M frameworks were applied for estimation of coincident indexes. To simulate as realistic as possible situation faced by analysts trying to use these indexes in real time for assessment of current and future economic situation, computations were performed on growing datasets. Introduction of particular observations at the ends of each time series were done in consistence with Polish official statistical Data Release Calendar available in ECB Statistical Data Warehouse website (http://sdw.ecb.int). This observation by observation growing dataset was used during final computati ons of coincident indicators. Earlier, in the phase of model parameters estimation, datasets reflecting situation at the end of each year of pseudo real-time exercise time span were used. Constructing this pseudo real-time data panel was one of the most time and effort consuming part of the survey.

To be consistent with stationary assumption used by both analytical frameworks, all selected time series were checked for presence of unit roots with standard ADF and KPSS test. Results of the estimation were gathered in the next two tables.

Table 1.

ADF Unit Root Test						
		GDP	IPI	RTTI	IMP	EXP
Loglevels	Spec.	T, C, 2L	T, C, 1L	T, C, 2L	T, C, 0L	T, C, 1L
	Statistics	-1.427	-2.028	-0.858	-1.235	-2.007
	p-value	0.842	0.551	0.947	0.882	0.572
First diff.	Spec.	C, 1L	C, 0L	C, 1L	C, 0L	C, 0L
	Statistics	-4.013	-15.381	-13.264	-14.244	-16.237
	p-value	0.0002	0.0000	0.0000	0.0000	0.0000

Legend: T – trend, C-constant, XL – number of lags (X)

Source: own computations

Table 2.

KPSS Test						
		GDP	IPI	RTTI	IMP	EXP
Loglevels	Spec.	T,C,BW4	T,C,BW10	T,C,BW10	T,C,BW10	T,C,BW10
	LM stat.	0.174	0.232	0.201	0.245	0.319
	5% level	0.146	0.146	0.146	0.146	0.146
First diff.	Spec.	T,C,BW4	T,C,BW6	T,C,BW6	T,C,BW4	T,C,BW4
	Statistics	0.232	0.057	0.172	0.153	0.178
	p-value	0.463	0.463	0.463	0.463	0.463

Legend: T - trend, C-constant, BW -bandwidth

Source: own computations

Achieved results proved that proposed transformation assured stationarity of selected time series.

Estimation of coincident indicators on growing datasets leads to the situation when edge of time series panel is ragged. Moreover in the case of M&M model which embraces mixed time series frequencies some observations within time sample are missing as well.

The table below shows part of time series sample available for estimation on 20.04.2010

As it has been mentioned in the previous section Mariano and Murasawa prepared their model for this type of dataset. Gaps within and at both ends of dataset are replaced with numbers drawn from Gaussian distribution. If any observation is missing in the particular row of input data matrix matrices H and R used in model estimation are tailored as well. In the case of S&W model, which does not include such mechanism,

Table 3.

Date	IPI	RTTI	IMP	EXP	GDP
Oct 09	-0.08	-0.15	5.12	1.64	NA
Nov 09	4.16	0	-3.37	-0.97	NA
Dec 09	-3.21	0.45	-0.66	0.16	1.2
Jan 10	2.75	-1.50	7.65	-0.67	NA
Feb 10	1.49	-1.37	-2.96	0.08	NA
March 10	2.252	TBO	TBO	TBO	TBO
TBO – to be observed, NA – not available					

Source: OECD MEI

this kind of gaps at the end of quarterly dataset were filled with values computed with Expectation Maximization (EM) algorithm.

After dealing with datasets' adjustment problem, both frameworks were ready for parameters estimation. On this stage likehood function maximization procedure with embedded Kalman filtering was employed. To identify proper structure of the models (determining number of lags in the equations modeling behavior of particular input variables and unobserved components) the set of models was estimated and information criteria (Akaike and Bayesian) were computed for each of them. Then models with lowest information criteria values were chosen for further investigation.

Two sets of parameters obtained for models selected in this way (for S&W and M&M models accordingly) were gathered in the following tables.

Estimated parameters determine intuitive dependencies between unobserved components (interpreted as coincident indicator for the S&W model or direct input to equation for coincident indicator computation for M&M framework) and input variables (positive $\gamma's$) and reflect reliable AR processes modeling common and idiosyncratic component s (ϕ, $\theta's$ which lie inside unit circle due to applied transformations). Moreover achieved values of variances connected with each modelled variable can be perceived as intuitive (for S&W model these variances can be bigger than one as input variables were not standardized).

Computed parameters were used in the next phase for estimation of two coincident indicators, assessment of their nowcasting and forecasting power and checking their business cycles dating capability. First of these two indicators was build with the use of S&W framework (therefor e it was named CISW), second one with M&M schema (named CIMM accordingly). Described phase was divided into loops. Each loop was performed for dataset available at the end of reference period and consisted of:

- Estimation of coincident indicators
- Computation of nowcasts and forecast
- Computation of nowcast and forecast errors

Table 4.

Estimated parameters of selected S&W model

Name	Value	Standard Error
γ_1	1.58	0.2
γ_2	0.75	0.24
γ_3	2.64	0.46
γ_4	3.33	0.53
ϕ_{11}	-0.30	0.01
ϕ_{12}	0.27	0.16
θ_{11}	0.22	0.02
θ_{12}	-0.01	0.001
θ_{21}	-0.17	0.01
θ_{22}	0.31	0.13
θ_{31}	0.18	0.01
θ_{32}	-0.01	0.001
θ_{41}	0.22	0.01
θ_{42}	0.01	0.001
σ_1	1.06	0.06
σ_2	2.64	0.50
σ_3	7.64	1.66
σ_4	8.62	3.41
Log likehood: 485.31 AIC: -2.58 BIC: -2.54		

Source: own computations

- Smoothing of coincident indicators
- Application of business cycle dating procedure

Estimation of coincident indicators were done with the use of filtering procedures of S&W and M&M described in the previous section. As a cut off date for each estimation last day of reference month (in the case of M&M model) or quarter was used. For M&M model one additional transformation were required to be applied because originally this framework produces monthly common component (indicator $CIMM^m$). To gain coincident indicator comparable with reference series and S&W

Table 5.

Estimated parameters of selected M&M model

Name	Value	Standard Error
γ_1	0.59	0.13
γ_2	0.21	0.09
γ_3	0.37	0.08
γ_4	0.63	0.08
γ_5	0.76	0.09
ϕ	-0.35	0.09
θ_1	-0.85	0.07
θ_2	-0.25	0.08
θ_3	-0.33	0.08
θ_4	-0.45	0.08
θ_5	-0.36	0.11
σ_1	0.67	0.17
σ_2	0.91	0.1
σ_3	0.79	0.09
σ_4	0.40	0.1
σ_5	0.31	0.13
Log likehood: 312.69		
AIC: -3.22		
BIC: -3.19		

Source: own computations

version for each period t linear combinations of the model's output were computed:

$$CIMM_t^q = \gamma_1(\frac{1}{3}CIMM_t^m + \frac{2}{3}CIMM_{t-1}^m + CIMM_{t-2}^m + \frac{2}{3}CIMM_{t-3}^m + \frac{1}{3}CIMM_{t-4}^m) \quad (42)$$

In the two charts below S&W and M&M transformed indicators are shown:

As depicted above it seems that both coincident indicators share common oscillations with reference series, although they don't register huge decline in GDP growth at the beginning of the sample.

To check their usability as a forecasting variables both indicators were transformed to annual growths /quarter on previous year quarter (QoPYQ)/ and set as an input for the ARX model (autoregressive model with exogenous variables modeled with

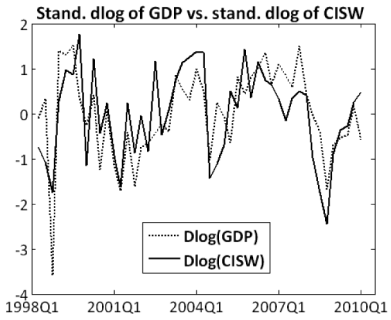


Figure 4.

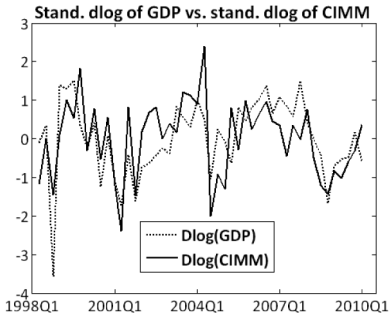


Figure 5.

autoregressive process as well). This framework assumes that forecasted variable and coincident indicator follow an AR process:

$$y_{t+h} = \alpha_0 + \sum_{i=1}^{k} \alpha_i y_{t-i} + \sum_{j=0}^{l} \beta_j . x_{t-j} \tag{43}$$

where y is forecasted variable (GDP in this survey), x is exogenous variable included in the model (CISW and CIMM respectively), h is forecast horizon (h = 0 for nowcast), k is AR process order for forecasted variable and l is AR process order for coincident indicator. Achieved nowcasts and forecast were compared with naive autoregressive model estimated on the same datasets. The standard measures: RMSE, MAE and ME were used as indicators of nowcasting and forecasting power. Results of this exercise are discussed in detail in the next section. In the last phase of the survey computed coincident indicators and reference time series were smoothed with Christiano-Fitzgerald asymmetric filter to highlight their cyclical properties. Finally smoothed indicators and GDP were subject to standard dating algorithm proposed by Bry and Boschan [4] (1971). Similarly to nowcast and forecast results dates of expansions and recessions are reported in the next section.

To recapitulate schema of general coincident indicators building and evaluating procedure is presented:

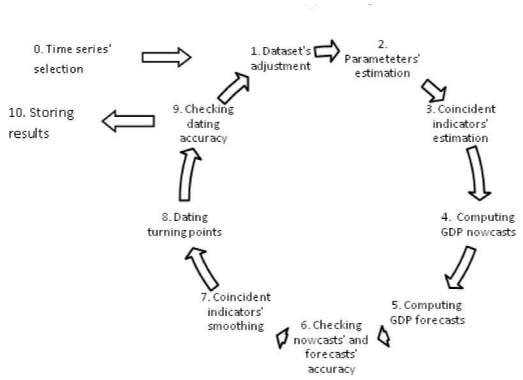


Figure 6.

6. EVALUATION OF NOWCASTING AND FORECASTING POWER

In this section results of the pseudo real-time nowcasting and forecasting exercise are presented. This exercise was performed for last 5 years of the sample, and the forecast horizon was set up to 4 quarters. Therefore the model was estimated for step-by-step growing sample starting on 1st quarter 2004 (March 2003 for MM model) and ending on the 1st quarter 2009 (March 2009). This span of time guaranteed that characteristics of the coincident indicator's nowcasting and forecasting power were checked during expansion and boom phase and in slowdown and recession period as well. Table 6 presents error statistics computed for nowcasting.

Table 6.

GDP nowcast errors

Model	Rank (due to RMSE)	RMSE	MAE	ME
AR	3	1.123	0.961	0.365
CIMM	1	**0.393**	**0.291**	-0.074
CISW	2	0.545	0.424	**-0.029**

Source: own computations

In the nowcast competition CIMM indicator performed better than naive model and its simpler counterpart. In terms of RMSE statistics it was nearly 65% better than naive model and 27% better when compared with S&W indicator. It proves intuitive guess that incorporating fresh monthly data at the end of the sample allowed to react quicker to any changes in the current situation of Polish economy.

During next stage of the survey forecasting power of both indicator was evaluated. Forecasts for each horizon (h=1, 2, 3, 4) were computed separately and then appropriate average statistics were computed.

Table 7 reports gained results:

Table 7.

GDP h-periods ahead forecast errors

Model	Rank (due to RMSE)	RMSE	MAE	ME
h = 1				
AR	3	1.257	1.089	0.413
CIMM	**1**	**0.861**	**0.694**	**0.092**
CISW	2	1.169	1.067	0.528
h = 2				
AR	3	1.944	1.654	0.658
CIMM	**1**	**0.951**	**0.796**	**0.147**
CISW	2	1.207	1.014	0.364
h = 3				
AR	3	2.696	2.301	1.038
CIMM	**1**	**1.485**	**1.259**	**0.370**
CISW	2	1.608	1.410	0.461
h = 4				
AR	3	2.851	2.560	1.224
CIMM	**1**	**2.033**	**1.805**	**0.896**
CISW	2	2.285	2.153	1.119

Source: own computations

In a manner analogous to nowcasting part of the exercise second generation dynamic factor indicator allows to formulate forecasts which are the best approximation of Polish GDP. For all verified horizons it offers minimal error associated with forecast (with over 50% reduction in comparison with naive model for 2 quarters horizon in terms of RMSE). Despite the fact that its advantage diminishes across the time, even in the perspective of 4 quarters it is still by more than 25% better than the AR model. CISW also allows to reduce forecast error, but it is not so effective as its first generation competitor. Hence, it can be stated that although both coincident indicators are helpful in forecasting process, M&M index brings more useful information about future development of economic environment in Poland.

Three figures below present evolution of nowcast and forecast errors with respect to particular horizon index.

For both indicators the steepest error growth can be noticed for change from nowcast to one quarter ahead forecast and for shift from horizon of 3 to 4 quarters.

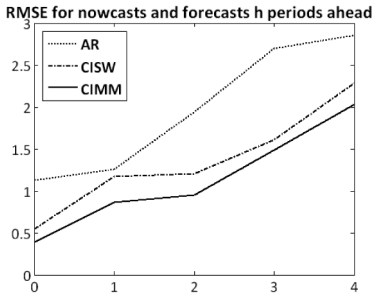


Figure 7.

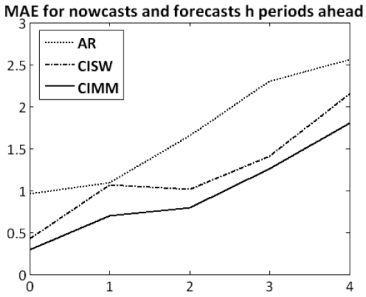


Figure 8.

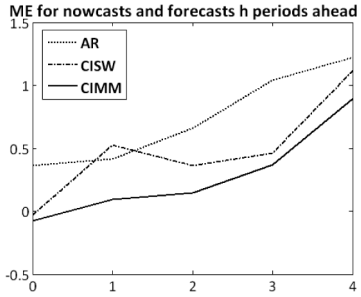
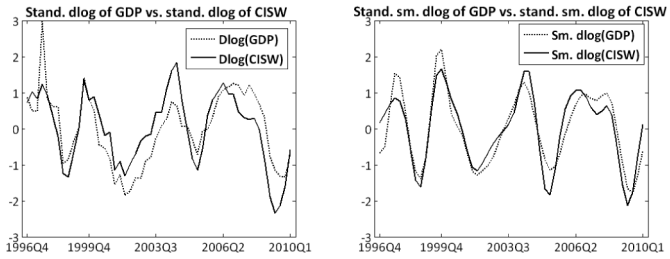


Figure 9.

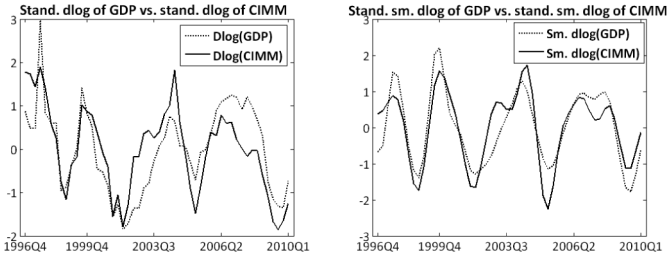
The biggest discrepancy in the forecast based on coincident indicators was registered for 3 quarters ahead perspective.

Last part of this section is devoted to presentation of the results of pseudo real-time business cycle dating exercise. The aim of this part of the survey was to check usefulness of coincident indicators for the process of early critical points detection. The aforementioned feature is very useful for decision makers and agents observing published coincident indicators because it allows them to determine changes in future trends of economic situation development in particular country. Dating procedure has been ran on indicators transformed to quarter on previous year's quarter growths smoothed with Christiano-Fitzgerald filter. Additionally turning points dates were also determined for original series and gained results were almost the same as in case of

transformed data. Set of figures grouped in two panels below shows inputs to the dating procedure for full time span (1996:Q4: 2010:Q1)



Panel 1.



Panel 2.

Basic comparison of above plots confirms that frequency of the business cycles determined by both coincident indicators are consistent with these recorded for the reference indicator. They are also coherent with previous research of the author devoted to synchronization of Polish business cycles (Lupinski [16], 2009), which suggest that average business cycle span in Poland, measured for GDP series, lasts 3 years.

Pseudo real-time dating procedure, like nowcasting and forecasting exercise, was run on step-by-step growing sample. In the table below first occurrences of peaks and troughs are reported.

It generally takes more than one year to detect turning point on reference time series. However use of M&M indicator allows to shorten this period of time (with exception to the second trough) even by more than 50%. Moreover its use guarantees detecting all turning points observed during pseudo real-time exercise horizon (5 years). This crucial feature is not shared by the Stock and Watson indicator, which misses one peak. It is also not so effective as its counterpart in shortening turning points detection lag.

7. CONCLUSIONS

In this paper second generation dynamic factor model with mixed frequencies and missing data handling was applied to create coincident indicator of Polish GDP and formulate with its help reference series' nowcast/short-term forecasts and date business

Table 8.
Business cycles dates determined with Bry-Boschan procedure for GDP, CISW and CIMM

Peak/trough real date	Date of detection			Lag of detection (months)		
	GDP	CISW	CIMM	GDP	CISW	CIMM
Peaks						
June 2006 (2006:Q2)	November 2007	April 2007	February 2007	17	10	8
December 2007 (2007:Q4)	August 2008	–	May 2008	8	–	5
Troughs						
June 2005 (2005:Q2)	August 2006	April 2006	March 2006	14	10	9
June 2007 (2007:Q2)	November 2008	January 2009	November 2008	14	16	14

Source: own computations

cycles' phases. Despite general drawbacks of the dynamic factor models inherited by their second generation version (e.g. identification problems influencing ability of stochastic trend modeling) and their native problems connected with use of Gaussian distribution to replace missing observations, they turned out to be robust econometric tools for helping decision makers and other agents to summarize current and future state of Polish economy. Results achieved during simulated real-time exercise based on small dataset have shown that coincident indicators built with help of M&M model give basis for formulation of forecast that significantly outperform naive benchmark models and dynamic models of the first generation. When compared with S&W coincident indicators they offer smaller forecast errors especially in the perspective of one and two quarters. They also allow to detect business cycle turning points earlier than their counterparts. Hence, intuitive guess that including more recent monthly information in the econometric model can improve its explanatory and forecasting features has found its measurable proof.

During his planned work the author is going to augment presented econometric framework with module introducing endogenous stochastic trend /this plan is inspired by Fernandez-Macho's idea [9] (1997)/. Second considered extension is connected with introduction into base model switching regime schema. The last augmentation is crucial in the context of changes in the structure of the Polish and global economy observed recently. Good example of such approach was presented in the paper of Camacho, Perez-Quirios and Poncela [7] (2010) in which Markov switching framework was exploited. Nevertheless both two mentioned extensions need to be substantially customized to Polish economic data characteristics and carefully tested.

REFERENCES

[1] Arouba S.B., Diebold F.X., Scotti C., 2009, Real-Time Measurement of Business Conditions, Journal of Business and Economic Statistics, vol 27 (4), pp. 417-27.
[2] Barhoumi K., Benk S., Cristadoro R., Den Reijer A., Jakaitiene A., Jelonek P., Rua A., Rünstler G., Ruth K. and Van Nieuwenhuyze Ch., 2004, Short-Term Forecasting of GDP Using Large Monthly Datasets a Pseudo Real-Time Forecast Evaluation Exercise, ECB Occasional Working Paper, no. 84.
[3] Boivin J., Ng S., 2006, "Are More Data Always Better for Factor Analysis?", Journal of Econometrics, 132, pp. 169-194.
[4] Bry G., Boschan C., 1971, Cyclical Analysis of Time Series: Selected Procedures and Computer Programs, NBER.
[5] Burns, A.F., Mitchell W.C., 1946, Measuring Business Cycles, NBER.
[6] Camacho M., Perez-Quiros G., 2009, "Ñ-STING: España Short Term INdicator of Growth," Banco de Espana Working Papers 0912, Banco de España.
[7] Camacho M., Pérez-Quiros G., Poncela P., 2010, "Green Shoots? Where, When and How?", Working Papers 2010-04, FEDEA.
[8] Forni M., et al, 2001, "Coincident and Leading Indicators for the Euro Area," Economic Journal, Royal Economic Society, vol. 111(471), pp. 62-85.
[9] Fernandez-Macho F.J., 1997, A Dynamic Factor Model for Economic Time Series, Kybernetika, vol. 33 (6), pp. 583-606.
[10] Mariano R.S., Murasawa Y., 2003, "A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series," Journal of Applied Econometrics, vol. 18 (4), pp 427-443, John Wiley&Sons.
[11] Inklar R., Jacobs J., Romp W. E., Business Cycle Indexes: Does a Heap of Data Help?, CCSO Working Paper 2003/12.
[12] http://www.rug.nl/staff/r.c.inklaar/research (accessed: 16.04.2010).
[13] Kim Ch. J., Nelson Ch. R., 1999, State Space Models with Regime Switching, MIT.
[14] Kowal P., 2005, "Matlab Implementation of Commonly used Filters" Computer Programs 0507001, EconWPA.
[15] Lupinski M., 2007, Konstrukcja wskaźnika wyprzedzającego aktywnoście ekonomicznej w Polsce, (Construction of Polish Economic Activity Leading Indicator), PhD Dissertation, Warsaw University.
[16] Lupinski M., 2009, Four Years After Expansion. Are Czech Republic, Hungary and Poland Closer to Core or Periphery of EMU?, Economics, vol. 22, pp. 75-103.
[17] Stock J. H., Watson M. W., 1989, "New Indexes of Coincident and Leading Economic Indicators" NBER Chapters, in: NBER Macroeconomics Annual 1989, vol. 4, pp. 351-409, NBER.
[18] Stock J. H., Watson M. W., 1998, "Business Cycle Fluctuations in U.S. Macroeconomic Time Series," NBER Working Papers 6528, NBER.
[19] Shumway R.H., Stoffer D.S., 1982, An Approach to Time Series Smoothing and Forecasting using the EM algorithm, Journal of Time Series Analysis, vol. 3, pp. 253-264.

KRÓTKOTERMINOWE PROGNOZOWANIE I BUDOWA RÓWNOCZESNEGO WSKAŹNIKA AKTYWNOŚCI EKONOMICZNEJ ZA POMOCĄ DYNAMICZNYCH MODELI CZYNNIKOWYCH WYKORZYSTUJĄCYCH DANE O MIESZANYCH CZĘSTOTLIWOŚCIACH I NIEZBILANSOWANYM KOŃCU PRÓBY

S t r e s z c z e n i e

Celem niniejszego artykułu jest prezentacja praktycznych aspektów estymacji nieobserwowalnych zmiennych za pomocą modeli czynnikowych drugiej generacji opartych o zmodyfikowany filtr Kalmana

i metodę największej wiarygodności (MNW). Opisane w pracy modele wykorzystywane są do krótkookresowego prognozowania polskiego PKB i konstrukcji równoczesnego wskaźnika aktywności ekonomicznej w Polsce na podstawie szeregów czasowych o mieszanych częstotliwościach, posiadających niezbilansowany koniec próby. Przedstawione podejście stanowi adaptację koncepcji modeli używanych do krótkookresowego prognozowania przez Camacho oraz Pereza-Quirioza w Narodowym Banku Hiszpanii oraz struktur analitycznych stosowanych przez Aruobę, Diebolda i Scotti'ego przy kompilacji indeksu aktywności ekonomicznej na potrzeby Banku Rezerwy Federalnej w Filadelfii. Zgodnie z wiedzą posiadaną przez autora jest to pierwsza tego rodzaju adaptacja w Europie Środkowo-Wschodniej. W ramach wykonanych badań za pomocą grupy testów statystycznych porównania została jakość prognoz uzyskanych za pośrednictwem zaproponowanych modeli z wynikami uzyskanymi metodami standardowymi. Ponadto ocenione zostały własności statystyczne obliczonego równoczesnego wskaźnika aktywności ekonomicznej w Polsce z wersją tego wskaźnika estymowaną regularnie z wykorzystaniem modeli czynnikowych pierwszej generacji (podejście Stocka-Watsona) począwszy od 2006 r.

Słowa kluczowe: prognozowanie krótkookresowe, wskaźnik równoczesny, modele czynnikowe, mieszane częstotliwości, niezbilansowany koniec próby

SHORT-TERM FORECASTING AND COMPOSITE INDICATORS CONSTRUCTION WITH HELP OF DYNAMIC FACTOR MODELS HANDLING MIXED FREQUENCIES DATA WITH RAGGED EDGES

A b s t r a c t

This paper aims at presenting practical applications of latent variable extraction method based on second generation dynamic factor models, which use modified Kalman Filter combined with Maximum Likehood Method and can be applied for time series with mixed frequencies (mainly monthly and quarterly) and unbalanced beginning and the end of the data sample (ragged edges). These applications embrace short-term forecasting of Polish GDP and construction of composite coincident indicator of economic activity in Poland. Presented approach adopts the idea of short-term forecasting used by Camacho and Perez-Quirioz in Banco de Espana and concept of Arouba, Diebold and Scotti index compiled in the FRB of Philadelphia. According to the author's knowledge, it is the first such adaptation for Central and Eastern Europe country. Quality of the forecast obtained with these models is compared with standard methods used for short-term forecasting with series of statistical tests in the pseudo real-time forecasting exercise. Moreover described method is applied for construction of composite coincident indicator of economic activity in Polish economy. This newly-created coincident indicator is compared with first generation coincident indicator, based on standard dynamic factor model (Stock and Watson) approach, which has been computed by the author for Polish economy since 2006.

Key words: short-term forecasting, coincident indicators, factor models, mixed frequencies, ragged edges

PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012

MAŁGORZATA MARKOWSKA, BARTŁOMIEJ JEFMAŃSKI

FUZZY CLASSIFICATION OF EUROPEAN REGIONS IN THE EVALUATION OF SMART GROWTH[1]

1. INTRODUCTION

It is possible to accomplish smart growth only if an overall innovation potential of EU regions becomes mobilized, since innovation is of great importance for all regions – for the well developed ones, to maintain the role of leaders, and also for these lagging behind to be able to catch up with the best ones [17]. Regions play the crucial role in smart growth accomplishment and in this case smart specialization profile is of decisive importance since it is the regions which take up the role of main institutional partners for universities, other research and educational institutions, as well as small and medium enterprises constituting the key component of innovation processes responsible for overall development [23].

Complex phenomena, such as: development, innovation or smart specialization in regions represent processes difficult for quantification and, in consequence, for measurement. Therefore for the purposes of their identification and characteristics it seems founded to apply tools used in contemporary econometrics, including multidimensional classification methods, the results of which allow for e.g. performing comparative and benchmarking analyses between spatial units, i.e. regions. Such approach has also been applied in the hereby study, which focuses on the evaluation of smart growth level in European regional space. The study applies European space division into fuzzy classes using the output of fuzzy sets theory and therefore allows the option of regions partial membership in particular classes. Substantiality of such approach was characterized, e.g. in the study by Y. Leung [14] who emphasized that numerous spatial analyses were carried out following the assumption that a region represents a strictly defined theoretical construct, characterized by clearly outlined spatial boundaries. This approach, on the other hand, implies the division of space into mutually exclusive regions, in line with Boole's logics. The reality, however, shows that in case of the majority of special phenomena it is not possible to define explicit boundaries between regions, which questions the reasons behind applying, let's call it, dichotomous regionalization.

[1] The study was prepared within the framework of research grant provided by The National Centre of Science no.2011/01/B/HS4/04743 entitled: *Classification of the European regional space in the perspective of smart development concept - dynamic approach.*

Therefore, also in the hereby study, the authors decided to take advantage of fuzzy sets theory output and, in particular, of fuzzy algorithms classification methods. The obtained results indicated both correctness and usefulness of the used approach and confirmed the assumption adopted by the authors that the boundaries between distinguished classes are not clear and in case of some regions it is difficult to indicate their unambiguous membership to the defined classes.

2. SMART SPECIALIZATION AS THE PILLAR OF SMART GROWTH

2.1 SMART GROWTH IN EU STRATEGIC DOCUMENTS

In the perspective of minor effects resulting from Lisbon Strategy implementation and in view of global economic crisis the united Europe needs a strategy which can open opportunities for coming out of the crisis not only stronger, but which may also result in EU economy becoming both smart and sustainable, supporting social inclusion, obtaining high employment rate and capacity indicators, as well as more extensive social cohesion. Therefore it seems crucial to release European innovation potential by improving the effects of education process as well as the quality and results of educational institutions functioning and also by using economic and social opportunities created by a digital society. Impulses should refer to regional, national and EU level.

In times of global crisis and in view of growing competition at an international arena, e.g. on the part of developed and emerging countries and also the need to reform global financial system, as well as challenges determined by climate changes and natural resources, EU needs new strategy based on coordinated economic policy, the strategy which ensures both development and employment rate growth [6].

In view of such challenges the European Commission, on 3^{rd} March 2011, presented the communication *Europe 2020 – Strategy for smart growth and sustainable development enhancing social inclusion* [8]. The Commission proposal regarding new strategy initiation was accepted on 26 March 2010 at The European Council meeting. *Europe 2020* strategy, as the successor of Lisbon Strategy, represents the vision of social market economy for Europe of the 21^{st} century and refers to three related priorities [8]:

– smart growth: knowledge-based economy and innovation development;
– sustainable growth: support for effective economy which uses resources, more environmentally friendly and more competitive one;
– inclusive growth: support for economy representing high employment rate, ensuring social and territorial cohesion.

EU strategic documents define smart growth in terms of obtaining better results in:

– education, by encouraging towards learning, studying and upgrading qualifications,
– research and innovation, by creating new products and services influencing economic growth and increasing employment rate which facilitate solving social problems,

– digital society, i.e. information and communication technologies implementation.

Guideline initiatives stimulating smart growth in EU are as follows:

– The European digital agenda [1, 5], i.e. establishing uniform digital market based on very fast Internet connections and interoperation applications;
– Innovation Union [11], implementing R&D and innovation projects in solving crucial economic problems; strengthening both the process and the role of innovation;
– Mobile youth, i.e. "Youth on the move" project, aims at improving educational results and increasing the attractiveness of European college education at an international arena.

Other initiatives (flagships in Europe 2020 strategy) are as follows [8]:

– Europe using its resources in an effective way,
– Industrial policy in globalization era,
– Programme for new skills and employment,
– European programme for fighting poverty.

Each of the listed above initiatives was also accompanied by measurable and possible to accomplish targets.

2.2 SMART SPECIALIZATION AS THE COMPONENT STIMULATING SMART GROWTH

Europe 2020 strategy represents an integrated approach where, apart from sustainable growth related priorities, emphasis is also placed on inclusive growth and smart growth. Smart growth, in the perspective of strategic documents [8], knowledge-based economy and innovation, means an extended role of knowledge and innovation as the forces stimulating future growth, which requires improvements in the quality of education, higher efficiency of research projects, supporting innovation and knowledge transfer in EU, better implementation of information and communication technologies, as well as ensuring that innovative ideas find their reflection in new products and services which stimulate better growth rate dynamics, facilitate opening new jobs and enhance solving crucial social problems in Europe and worldwide.

The success of Europe 2020 Strategy also depends on such elements as entrepreneurship, financial resources and paying attention to both, users' needs, and opportunities offered by the market.

Smart specialization constitutes an important component of smart growth, which covers enterprises, research centres and higher education institutions cooperating in order to define the most promising areas of specialization in a given region, but also weaknesses making innovation implementation difficult. This approach takes into consideration differences in business opportunities specific for particular regions with reference to innovation, since while the leading regions can invest in upgrading general technologies or innovations in services, in case of the other regions better results may be obtained by investing in innovations in a specific sector, or a few related sectors.

Smart specialization accomplishment turns out possible, if regional diversity is taken advantage of, by stimulating cooperation and extending it beyond (regional and

national) boundaries, by regional opening towards new opportunities, by avoiding fragmentation and as the result offering undisturbed knowledge transfer in EU.

Strategic intelligence which allows for defining activities of high added value and therefore offers maximum opportunities for upgrading regional competitiveness, turns out necessary for smart growth accomplishment. Significant influence of investments in R&D and innovation requires obtaining critical mass which should be accompanied by funds aimed at upgrading skills, as well as an overall level of educational institutions and infrastructure extension. State governments (national level) and local governments (regional level) should develop strategies for "smart specialization" in order to obtain optimum effects of the carried out regional policy and other EU policies.

3. FUZZY CLASSIFICATION

Fuzzy set $\tilde{A}$ in $X = \{x\}$ space marked as $\tilde{A} \subseteq X$ is defined by the set of pairs $\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in X\} \forall x \in X$, where $\mu_{\tilde{A}} : X \to [0, 1]$ means the membership to $\tilde{A}$ fuzzy set function, which assigns to each $x \in X$ element its membership level to $\tilde{A}$ fuzzy set, [26 pp. 11-12]. Therefore, a fuzzy set represents the generalization of a classical set, where$x$element belongs ($\mu_{\tilde{A}}(x) = 1$) or does not belong ($\mu_{\tilde{A}}(x) = 0$) to A set [24, p. 100].

In classical classification methods an object's membership in a class is expressed by a binary variable. In other words, either an object belongs to a given class or it does not. The application of fuzzy sets concept in the problem of classification allows for the option of an object's membership in more than just one class. It is possible due to binary variable substitution by a continuous variable presenting values in [0; 1] interval. This procedure allows for describing the situation more precisely, i.e. where the boundaries between classes are "unclear" and assigning an object to a unique class becomes more difficult [12, p. 160]. It also reflects the actual reality to a higher extent and may serve as the protection for a research worker preventing him/her from losing certain information, as opposed to an approach where objects are assigned to one class only [13, p. 146-156].

The profile of fuzzy classification methods has already been discussed in studies, e.g. by: F. Höppner [10], K. Jajuga [12] and F. Wysocki [24]. One of the more frequently applied methods is the fuzzy c-means method suggested by A. Dunn [7] and later generalized by J. C. Bezdek [2] and F. Höppner [10, p. 37]. An interesting and comprehensive review of k-means method development through its diversified modifications and generalizations (including fuzzy classification) was presented by Bock [3]. An extensive description of this method covering its advantages and disadvantages, was also included in the study by Steinley [20]. Its application does not require making assumptions regarding the nature of empirical material under analysis. It is an iteration method the idea of which is very close to classical k-means method. The method focuses on finding such gravity centres of classes which minimize the function below

[15, p. 303]:

$$J_m = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^m d_{ij}^2, \tag{1}$$

where: μ_{ij} – level of j-th object membership in i-th fuzzy class,

d_{ij} – Euclidean distance between i-th fuzzy class gravity centre and j-th object,

m – fuzzification parameter, where $m > 1$.

The algorithm of fuzzy c-means method consists of the following steps [4, pp. 230-238]; [15, pp. 302-303]:

Step 1. Random membership matrix initiation $\mathbf{U} = \left[\mu_{ij}\right]$ where: $\sum_{i=1}^{c} \mu_{ij}$, $\forall j = 1, 2, ..., n$.

Step 2. Calculating gravity centres of classes following the below formula:

$$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m x_j}{\sum_{j=1}^{n} \mu_{ij}^m}, \tag{2}$$

where:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}}. \tag{3}$$

Step 3. Calculating new $\mathbf{U}_{new}$ membership matrix. If $\|\mathbf{U}_{new} - \mathbf{U}\| > \varepsilon$, where $\|\mathbf{U}_{new} - \mathbf{U}\|$ represents Euclidean distance while ε refers to the accepted convergence threshold, then $\mathbf{U} = \mathbf{U}_{new}$ should be accepted and we should proceed to step 2. Procedure is finalized in the situation when $\|\mathbf{U}_{new} - \mathbf{U}\| < \varepsilon$ or the set number of k iterations is obtained.

Before initiating calculations the applied distance measure, number of classes, initial membership of objects in classes and the value of m fuzzy parameter, have to be defined.

M parameter specifies the degree of classification results fuzzification. Parameter value should be $m>1$ where values close to unity will bring about the results similar to these received by means of classical methods. An increase in m parameter value results in the fact that the degree of objects membership in particular classes will take values close to inverse number of classes, i.e. $\frac{1}{c}$ [3, pp. 228-229]; [13, p. 146]. Professional literature does not offer theoretical background for the choice of optimum m parameter value, therefore its choice is frequently made based on experience from previously conducted empirical studies. F. Wysocki's research results suggest that parameter value should be included in [1,3;1,5] interval [24, p. 101].

4. CHARACTERISTICS OF EU REGIONS IN TERMS OF SMART SPECIALIZATION IDENTIFIERS

The implementation of visions described in the strategy requires specifying targets, therefore the targets set for ensuring smart growth are as follows:

1. higher level of outlays on investments (public and private) up to 3% of EU GDP and providing better conditions for R&D and innovation,
2. higher share of the employed (women and men) aged 20-64 up to 75% in 2020 as the result of introducing more people to the job market (mainly women, young and senior citizens, unqualified individuals and legal immigrants),
3. higher level of education, especially by:
   – reducing the share of young people finishing their education prematurely to less than 10%,
   – increasing the share of university graduates aged 30-34 up to at least 40%.

Measurable targets were also defined for other strategy elements, i.e. sustainable growth and inclusive growth. Even though so much has been written about smart specialization of regions, still the smart specialization identifying parameters have not been defined.

Specialization results in different effects depending on technological level it refers to, while the range of its occurrence may be indicated, as below [9]:

– specialization in the domain of scientific knowledge,
– specialization in technology and innovation,
– specialization in production processes,
– specialization related to clusters,
– horizontal vs. vertical specialization.

In professional literature specialization measurement originates from trade theory. Diversified specialization indicators were prepared in order to allow for capturing country specialization, but also different indicators were elaborated and used as technology specialization indicators and obviously, after certain corrections were applied.

The growing service orientation is one of the first specialization symptoms, therefore it seems that defining the importance of service sector at the background of other sectors in economy will be a helpful indicator in defining smart specialization. The service sector may be defined by:

X_1 – share of workforce employed in farming in an overall number of workforce in the region,

X_2 – share of workforce employed in industry in an overall number of workforce in the region,

X_3 – share of workforce employed in services in an overall number of workforce in the region.

Anther manifestation of smart specialization is the growing importance of innovative sectors which may be assessed by means of the following characteristics:

X_4 – workforce employed in knowledge-intensive services as the share of overall workforce in services,

X_5 – workforce employed in high and mid-tech industry (as % of all workforce employed in industry).

The list of due characteristics could be much longer, however, in many cases regional level data are not available for properties potentially helpful in smart specialization identification.

The identification of smart specialization was prepared based on the set of EU NUTS 2 regions, the total number of which amounts to 271 [18], however, due to data gaps referring to selected characteristics about French overseas regions (Guadeloupe, Matinique, Guyane, Réunion) and two Spanish ones (Ciudad Autónoma de Ceuta, Ciudad Autónima de Melilla) further analysis covers 265 out of 271 EU regions. The data regarding regions were collected form Eurostat data base and referred to 2008.

Measures selected for the evaluation of smart specialization, i.e. the five mentioned above characteristics are most diversified (in assessing the variability and also maximum and minimum ratio) regarding the share of workforce employed in farming in the total number of workforce in the region (variability coefficient equals 115% and the max/min ratio amounts to as much as 303), while the least diversified in case of two properties:

– share of workers employed in services in the total number of workers in the region (variability coefficient equals 15,9% and max/min ratio is 3),
– workforce employed in knowledge-intensive services as the share of total workforce in services (variability coefficient – 15,9%, max/min ratio – 2,9).

Extreme values for particular variables were as follows:

– share of workforce employed in farming in the total number of workforce in the region (from 0,16% up to 48,5%) and among regions where the property value was below 1%, 12 regions were listed (Utrecht (NL), Hovedstaden (DK), Wien (AT), Outer London (UK), Berlin (DE), Région de Bruxelless (BE), Bremen (DE), Hamburg (DE), Stockholm (SE), Inner London (UK), Île de France (FR), Praha (CZ), while among regions characterized by the share of workforce employed in farming above 25% 11 regions were included (Nord-Est (RO), Sud-Vest Oltenia (RO), Lubelskie (PL), Sud – Muntenia (RO), Świętokrzyskie (PL), Podlaskie (PL), Peloponnisos (GR), Sud-Est (RO), Nord-Vest (RO), Podkarpackie (PL), Anatoliki Makedonia, Thraki (GR)),
– share of workforce employed in industry sector in the total number of workforce in the region (9,4% – 46,8%), out of which in thirteen regions it was less than 16% (four Dutch regions (Flevoland, Zuid-Holland, Noord-Holland and Utrecht), two British ones (Outer London and Inner London) and Greek (Voreio Aigaio, Ionia Nisia), and also Région de Bruxelless (BE), Hovedstaden (DK), Île de France (FR), Stockholm (SE) and Luxembourg), while in fifteen regions above 40% (6 Czech regions: Severovýchod, Střední Morava, Jihozápad, Jihovýchod, Střední Cechy, Moravskoslezsko, two from Romania (Centru, Vest), two from Slovenia (Západné Slovensko, Stredné Slovensko) and two from Hungary (Közép-Dunántúl, Nyugat-

Dunántúl) and also one Italian (Marche), German (Stuttgart) and Spanish region (La Rioja)),

– workforce employed in services in the total number of workforce in the region (30,4% – 90,2%), out of which above 80% in 19 regions (4 British (Inner London, Outer London, Berkshire, Bucks and Oxfordshire, Surrey, East and West Sussex) and four Dutch (Flevoland, Utrecht, Noord-Holland, Zuid-Holland), two Belgian (Région de Bruxelless, Prov. Vlaams Brabant), German (Berlin, Hamburg) and French (Île de France, Languedoc-Roussillon) and also Wien, Praha, Danish Hovedstaden, as well as Stockholm and Luxembourg), while below 50% in 18 regions (eight Polish: Małopolskie, Lubelskie, Podkarpackie, Świętokrzyskie, Podlaskie, Wielkopolskie, Opolskie and Kujawsko-Pomorskie, seven Romanian: Nord-Vest, Centru, Nord-Est, Sud-Est, Sud – Muntenia, Sud-Vest Oltenia, Vest and also Severovýchod (CZ), Centro (PT) and Vzhodna Slovenija (SI)),
– workforce employed in knowledge-intensive services as the share of total workforce employed in services (25,2% – 73,5%), where above 60% was registered in 14 regions (8 Swedish: Stockholm, Östra Mellansverige, Småland med öarna, Sydsverige, Västsverige, Norra Mellansverige, Mellersta Norrland, Övre Norrland, three British: Inner London, Berkshire, Bucks and Oxfordshire, Surrey, East and West Sussex, two Finnish (Pohjois-Suomi, Åland) and Danish Hovedstaden), while below 35% in 13 regions (7 Greek: Anatoliki Makedonia, Thraki, Ionia Nisia, Sterea Ellada, Peloponnisos, Voreio Aigaio, Notio Aigaio, Kriti, two Portuguese (Algarve, Região Autónoma dos Açores) and Romanian (Centru, Sud-Est) and also Bulgarian (Yuzhen tsentralen) and Spanish (Canarias)),
– workforce employed in high and mid-tech industry sectors (as % of workforce in industry) – from 3,9% to 59%, including 10 regions characterized by the property value below 6% (7 Greek regions: Anatoliki Makedonia, Thraki, Dytiki Makedonia, Ipeiros, Dytiki Ellada, Voreio Aigaio, Notio Aigaio, Kriti, two Spanish (Extremadura, Canarias) and Cyprus), similarly to the situation of regions characterized by the property value above 40% where eight German regions were listed: Stuttgart, Karlsruhe, Freiburg, Tübingen, Oberbayern, Bremen, Braunschweig, Rheinhessen-Pfalz and two French ones (Alsace and Franche-Comté).

The highest (negative) correlation was registered between X_1 and X_3 properties (-0,75) and X_2 and X_3 (-0,72) and also X_1 and X_4 (-0,47), while the next (positive) correlation between the following properties: X_3 and X_4 (0,53) as well as X_4 and X_5 (0,44) which means that an increase (decrease) of workforce share in farming is accompanied by a drop (growth) of workforce share employed in services and knowledge-intensive services, while an increase (decrease) of workforce share in industry goes along with the drop (growth) of workforce share in services. The same orientation is, however, observed in case of:

- changes in the share of workforce employed in services and knowledge-intensive services, as well as

- changes in the share of workforce employed in knowledge-intensive services in high and mid-tech industry sectors.

5. CLASSES OF REGIONS OBTAINED BY APPLYING FUZZY *C*-MEANS METHOD

The application of fuzzy *c*-means method in order to distinguish classes of regions, just like in case of classical *k*-means method variant, requires the number of such classes to be specified. If the non-statistical information about the number of classes is missing two solutions are suggested [24, p. 136]:

– to accept the number of classes specified using disjoint classification methods for the same data matrix,
– to perform fuzzy classification for a different number of classes and select the one for which fuzzy classification quality index reaches an extreme level.

In the hereby study the first procedure was accepted. As Sokołowski [19] emphasizes no classification method exists the advantage of which over other methods would be commonly acceptable. Therefore it is suggested to apply different methods in solving the classification problem and compare the obtained results. In view of the above, in the hereby analysis three, well recognized in professional literature, classification methods were applied in order to divide regions into classes: Ward method, $k-$medoid method and $k-$means method in relation to *Gap* classification quality assessment index. In order to standardize values of all variables, the regarding their variability, the standardization formula was applied. In this way variability, as the basis for objects diversification, was eliminated. Detailed characteristics and properties of variable normalization formulas (including standardization) may be found, among others, in studies by Pawełek [16] as well as Walesiak and Gatnar [22]. The similarity of objects was assessed by means of Euclidean distance (in case of Ward method square of Euclidean distance was applied). *Gap* index values and *diffu* statistics for a different number of classes, specified by means of the three classification methods, are illustrated in Table 1.

Graphical presentation of *Gap* index for a different number of classes is shown in picture 1, while the results of classification obtained by means of Ward method are presented in a dendrogram form in picture 2.

The results of classification obtained using Ward method and $k-$medoid method suggest the division of regions into four classes. In case of $k-$means method *Gap* index values suggest the division into five classes. The assessment of classification stability level in case of four and five classes variant was conducted by means of replication analysis [22, p. 419-420]. The analysis results are presented in Table 2.

The highest stability is observed in the variant of regions classification into four classes using $k-$means method. Classification stability in case of the suggested by *Gap* index division into five classes, and applying $k-$means method, proved lower. High stability is also characteristic for the division into four classes by means of the other

Table 1.

Index values of *Gap* and *diffu* statistics classification quality assessment depending on the number of classes

Class no.	**Classification methods**					
	Ward		***k*-medoid**		***c*-means**	
	Gap	***diffu***	***Gap***	***diffu***	***Gap***	***diffu***
2	1,01530	-0,04552	1,04846	-0,07183	1,0557	-0,05586
3	1,09024	-0,07141	1,15610	-0,01635	1,13291	-0,03074
4	1,18634	0,01484	1,21552	0,06360	1,21083	-0,00120
5	1,20840	0,04596	1,18817	0,05279	1,24119	0,06969
6	1,21571	0,05553	1,19104	0,04110	1,21806	0,03111
7	1,17350	-0,01324	1,16899	0,05565	1,21732	0,04697
8	1,20252	0,03294	1,12882	-0,06117	1,18542	-0,02165
9	1,20446	0,00085	1,22515	0,00730	1,20594	-0,01057
10	1,20486	0,01770	1,21747	0,01469	1,24870	0,06138

Source: Author's compilation using clusterSim package in **R** programme.

Table 2.

Adjusted Rand measure values

Class no.	**Adjusted Rand measure**		
	Ward	***k*-medoid**	***c*-means**
4	0,648875	0,736938	0,7411473
5	0,6402254	0,6096901	0,4975753

Source: Author's compilation using clusterSim package in **R** programme.

two methods. Owing to the above results, related to quality and stability of the obtained classification, for the purposes of fuzzy classification, the division of 265 EU NUTS 2 level regions into four classes was performed.

Fuzzy classification applying fuzzy $c-$means method requires, apart form *a priori* specification of classes number, also the initial classification of objects. Possible approaches in this matter are presented, inter alia, in F. Wysocki's study [24, p. 102]. The hereby analysis applies one of them which consist in random assignment of objects into four classes.

The similarity of objects was assessed by means of Euclidean distance. Fuzzy parameter value was set at the level of $m = 1{,}5$. Gravity centres for the distinguished classes are presented in Table 3.

The first class is characterized by the highest average share of workforce employed in farming sector in the total number of workforce in the region (24,3%) and the second

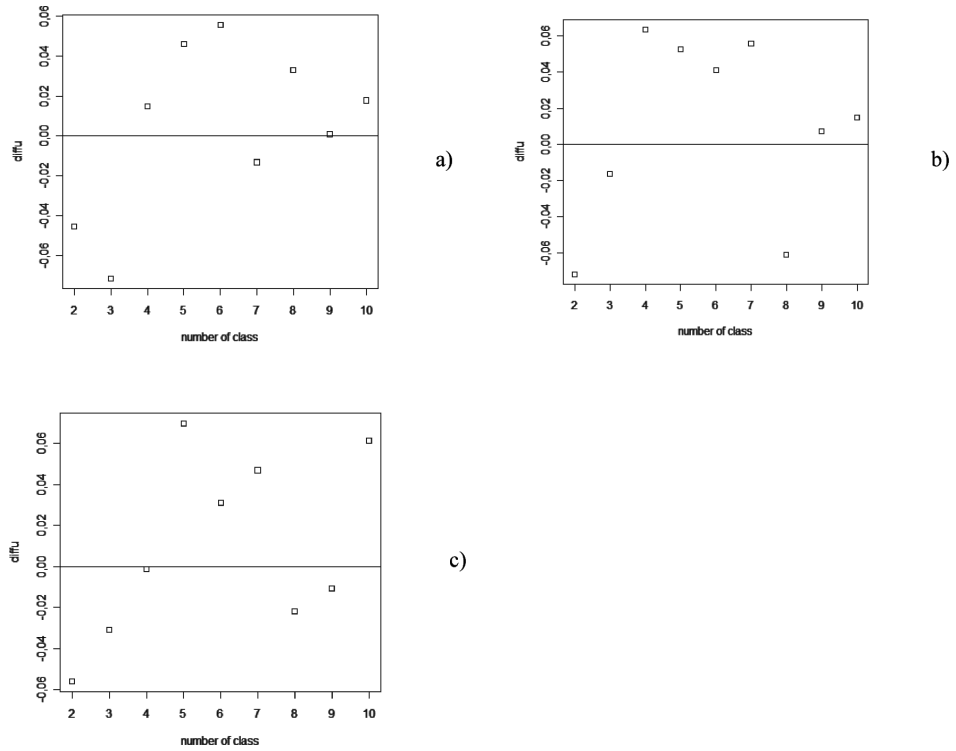


Figure 1. *Gap* index values graphs respectively for the following methods: Ward, *k*-medoid and *c*-means

Source: Author's compilation using clusterSim package in **R** programme.

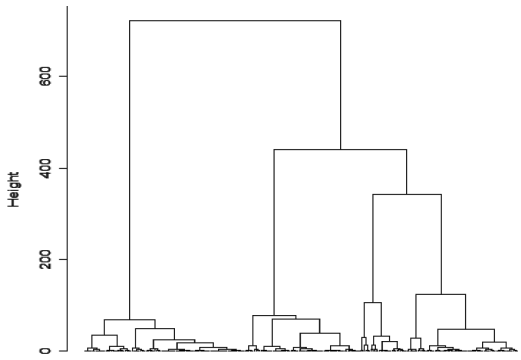


Figure 2. Dendogram prepared using Ward method

Source: Author's compilation using clusterSim package in **R** programme.

largest average share of workforce employed in industry sector (28,8%), and the lowest average share of:

– workforce employed in services in the total number of workforce in the region (46,98%),

Table 3.

Arithmetic means for four classes obtained using fuzzy $c-$means method and for classes obtained at threshold membership value of 0,8

Class no.	variable					variable				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
1	24,26	28,75	46,98	40	13,5	26,02	26,83	45,78	40,37	12,64
2	6,83	26,25	66,92	41,7	15,3	6,71	25,72	67,83	41,57	14,18
3	4,62	35,72	59,66	47,4	31,9	4,05	36,48	59,14	46,74	31,27
4	2,82	21,79	75,38	54,4	25,3	2,12	21,55	75,97	54,73	25,24

Source: Author's compilation using **R** programme.

– workforce employed in knowledge-intensive services in the total number of workforce employed in services in the region (40%),
– workforce employed in high and mid-tech industry sector in the total number of workforce in industry (13,5%).

The first class covers the smallest number of regions, i.e. only 10,2%, i.e. 27 European space regions (9, i.e. 4,3% of EU 15 regions and 18, i.e. 32% of EU 12 regions) and agglomerates NUTS 2 level units representing 5 countries: 10 out of 16 Polish ones, 7 out of 8 Romanian, 7 out of 13 Greek, 2 out of 7 Portuguese and 1 out of 6 Bulgarian regions.

The second class, covering 70 regions, i.e. every fourth EU region, is characterized by the second largest average share of workforce employed in farming (6,8%) and services (67%) and the third highest average level of X_2, X_4 and X_5 properties. The second class group of regions includes:

– 57 EU 15 regions (27,3%), including: 13 from 17 Spanish (76,5%) regions, 12 from 21 Italian (57,1%), 6 from 9 Austrian (66,7%), 6 from 22 French (27, 3%), 6 from 13 Greek (46,2%), 5 from 7 Portuguese (71,4%), 4 from 39 German (10,3%), 2 from 31 British (5,4%) and one from Belgium, Ireland and The Netherlands,
– 13 EU 12 regions, including: three from Poland (out of 16) and Bulgaria (out of 6), one from Hungary and Romania, as well as Cyprus, Estonia, Lithuania, Latvia and also Malta.

The third class, covering 60 regions, (38 from EU15, i.e. 18,2% and 22 from EU 12, i.e. 39,3%) may be called a smart specialization class in industry, since it is characterized by the highest average share of workforce employed in industry (35,7%) and in high and mid-tech industry in the total number of workforce employed in industry (31,9%). The following regions are included in the third class: 21 from 39 German regions (which makes 1/3 of all regions in this class), 7 from 8 Czech, 6 from 21 Italian, 5 from 7 Hungarian, 4 from 22 French, 3 from 4 Slovak, 3 from 16 Polish, 3 from 17 Spanish; 2 Slovenian regions, 2 from 6 Bulgarian and 2 from 9 Austrian regions, as well as one region from Finland and one from Belgium.

The most numerous fourth class, covering 40,8% of all analyzed EU regions, i.e. 108, are characterized by the highest average share of workforce employed in services (75,4%) and also in knowledge-intensive services in the total number of workforce employed in services in the region (54,4%). This class may be called a set of regions following the path towards smart specialization in services. It covers 50,2% of regions from EU 15 and 5,4% of regions from EU 12 ones. Regions from the latest accessions represent three capital regions, or the ones including capital city: Praha, Közép-Magyarország and Bratislavský kraj. Individual regions from UE 15 (Luxembourg, Southern and Eastern, Comunidad de Madrid, Wien and 35 from 37 British regions (94,6%), 14 from 39 German (36%), 12 from 22 French (54,5%), 11 from The Netherlands (91,7%), all Swedish (8) and all Danish regions (5), 9 from 11 Belgian ones (82%), 4 from 5 Finnish (80%) and 3 from 21 Italian (14,3%) regions.

The obtained classes, however, included regions featuring 40% membership regarding the dominating, in a given class, property value and causing that further analyses focused only on these regions which met class membership threshold at the set value of 0,8.

The first class, which covers 16 regions after setting higher membership threshold, is characterized, just like in case of the initial division, by the highest mean value of X_1 property (26,02%) and the second mean value of X_2 property (28,8%) and also the lowest mean value of other properties, such as:

– workforce employed in services in the total number of workforce in the region (45,78%),
– workforce employed in knowledge-intensive services in the total number of workforce in services in the region (40,37%),
– workforce employed in high and mid-tech industry sectors in the total workforce number employed in industry (12,64%).

The first class covers 6% of the European space regions (5, i.e. 2,4% of EU 15 regions and 11, i.e. 19,6% of EU 12 regions) and includes regions from 4 countries: 7 Polish, 4 Romanian and 4 Greek regions and one from Portugal.

In the second class covering 42 EU regions (i.e. 15,8%) the second largest average levels of X_1 (6,7%) and X_3 (67,8%) properties and the third largest average levels of X_2 (25,7%), X_4 (41,6%) and X_5 (14,2%) properties should be registered. This class includes the following regions:

– 37 UE 15 regions (17,7%), including: 11 from 17 Spanish (64,7%), 8 from 21 Italian (38,1%), 5 from 9 Austrian (55,6%), 3 from 22 French (13,6%), 3 from 7 Portuguese (43%), 3 from 39 German (7,7%) and also one form Greece and one from Ireland,

– 5 UE 12 regions (8,9%): one from Bulgaria and Romania, as well as Cyprus, Lithuania and Latvia.

The third class, including 46 regions (28 from UE 15, i.e. 13,4% and 18 from UE 12, i.e. 32%), should be defined as the smart specialization class in industry (the highest average share of workforce in industry (36,84%) and in high and mid-tech

industry in the total number of workforce employed in industry (31,3%). This class includes the following regions: 16 out of 39 German (every third region in this class), 7 out of 8 Czech, 6 out of 21 Italian, 5 out of 7 Hungarian, 2 out of 22 French, 3 Slovakian and 3 Spanish, 2 Polish, one Slovenian and one Austrian region.

The fourth class is the biggest and covers 32,8% of the analyzed EU regions, i.e. 87 of them. It is distinguished by the highest average level of X_3 (75,9%) and X_4 (54,7%) properties, i.e. knowledge-intensive services. This class covers 40,2% regions from EU 15 and 5,4% of regions from EU 12 (capital regions or these including capital city: Praha, Közép-Magyarország and Bratislavský kraj). One region of UE 15 (Southern and Eastern, Comunidad de Madrid, Wien) and 32 from 37 British (86,5%), 9 from 39 German (23,1%), 11 from The Netherland (91,7%), 6 from 22 French (27,3%), 7 from 8 Swedish (87,5%), all Danish regions (5), 7 from 11 Belgian (63,6%), 2 from 5 Finnish (40%) and 2 from 21 Italian (9,5%) regions.

Regions which "left" the classification at the set membership threshold of 0,8 were mostly, following their initial assignment, included in the second class (difficult in terms of its clear and unique specification) – made up of 28 regions, including 5 from Greece (Ionia Nisia, Attiki, Voreio Aigaio, Notio Aigaio and Kriti), 4 from Italy (Provincia Autonoma Trento, Umbria, Abruzzo, Molise), 3 from Poland (Mazowieckie, Zachodniopomorskie i Lubuskie) and 3 from France (Champagne-Ardenne, Picardie, Languedoc-Roussillon), 2 from Bulgaria (Severozapaden, Severoiztochen), Spain (La Rioja, Aragón) and Portugal (Lisboa, Região Autónoma dos Açores), as well as one from Austria (Steiermark), Belgium (Prov. West-Vlaanderen), Germany (Weser-Ems), The Netherlands (Zeeland), Hungary (Dél-Alföld) and also Malta and Estonia. The fourth class (definitely service oriented one and characterized by high innovation level of services) covers 21 regions from such countries as: France (Centre, Pays de la Loire, Aquitaine, Rhône-Alpes, Auvergne, Provence-Alpes-Côte d'Azur), Germany (Oberbayern, Bremen, Mecklenburg-Vorpommern, Hannover, Saarland), Great Britain (Inner London, North Eastern Scotland, Northern Ireland), Finland (Pohjois-Suomi, Åland) and Belgium (Prov. Liège, Prov. Luxembourg), Italy (Valle d'Aosta), Sweden (Småland med öarna) and Luxembourg. The third class (mid and high-tech industry oriented) includes 14 regions from such countries as: Germany (Karlsruhe, Mittelfranken, Gießen, Braunschweig, Rheinhessen-Pfalz), Bulgaria (Severen tsentralen, Yugoiztochen), France (Lorraine, Alsace), Austria (Vorarlberg), Belgium (Prov. Limburg), Finland (Länsi-Suomi), Poland (Śląskie) and Slovenia (Vzhodna Slovenija). The first class (farming profile) – 11 regions from Poland (Wielkopolskie, Opolskie, Warmińsko-Mazurskie), Greece (Thessalia, Ipeiros, Dytiki Ellada) and Romania (Centru, Sud-Vest Oltenia, Vest), Portugal (Norte) and Bulgaria (Yuzhen tsentralen).

The membership threshold increasing strategy, in most cases, also resulted in decreasing extreme values of the analyzed properties (maximum and minimum) and the mean value in each class and, in consequence, in reducing standard deviation and variability coefficient values which resulted in higher cohesion of the obtained classes.

Table 4.
Number of regions from EU countries in classes obtained using fuzzy $c-$means method and for classes obtained at threshold membership value of 0,8

Country (number of regions)	Number of regions in a class and membership thresholds				Number of regions in a class and membership thresholds				Undefined
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	
	0,409 -0,986	0,391 -0,999	0,392 -0,999	0,326 -0,999	0,800 -0,986	0,800 -0,997	0,800 -0,999	0,800 -0,999	<0,800
Austria (9)		6	2	1		5	1	1	2
Belgium (11)		1	1	9				7	4
Germany (39)		4	21	14		3	16	9	11
Denmark (5)				5				5	
Spain (17)		13	3	1		11	3	1	2
Finland (5)			1	4				2	3
France (22)		6	4	12		3	2	6	11
Greece (13)	7	6			4	1			8
Ireland (2)		1		1		1		1	
Italy (21)		12	6	3		8	6	2	5
Luxemburg (1)				1					1
The Netherlands (1)		1		11				11	1
Portugal (7)	2	5			1	3			3
Sweden (8)				8				7	1
Great Britain (37)		2		35		2		32	3
Bulgaria (6)	1	3	2			1			5
Cyprus (1)		1				1			
The Czech Republic (8)			7	1			7	1	
Estonia (1)		1							1
Hungary (7)		1	5	1			5	1	1
Lithuania (1)		1				1			
Latvia (1)		1				1			
Malta (1)		1							1
Poland (16)	10	3	3		7		2		7

cd. Table 4.

Romania (8)	7	1			4	1			3
Slovenia (2)			2				1		1
Slovakia (4)			3	1			3	1	
UE 27 (265)	**27**	**70**	**60**	**108**	**16**	**42**	**46**	**87**	**74**
UE 15 (209)	**9**	**57**	**38**	**105**	**5**	**37**	**28**	**84**	**55**
UE 12 (56)	**18**	**13**	**22**	**3**	**11**	**5**	**18**	**3**	**19**

Source: Author's compilation using **R** programme.

Abbreviations used in the study: UE 27 (regions of all UE countries), UE 12 (regions from countries of the latest two accessions), UE 15 (regions from "old EU 15"), AT (Austria), BE (Belgium), NL (The Nederlands), DE (Germany), DK (Denmark), ES (Spain), FI (Finland), FR (France), GR (Greece), IE (Ireland), IT (Italy), LU (Luxemburg), PT (Portugal), SE (Sweden), UK (Great Britain), BG (Bulgaria), CY (Cyprus), CZ (The Czech Republic), EE (Estonia), HU (Hungary), LT (Lithuania), LV (Latvia), MT (Malta), PL (Poland), RO (Romania), SI (Slovenia), SK (Slovakia).

6. FINAL CONCLUSIONS

The adequately focused regional policy is capable of releasing the EU growth potential by means of promoting innovation in all regions and, at the same time, ensuring complementarity between EU, national and regional support for innovation, research and development, entrepreneurship and also technology and communication. Regional policy represents the basic tool for adjusting EU innovation priorities for the purposes of practical, field activities.

Europe 2020 strategy covers three priority areas, five main goals, ten integrated guidelines and seven leading initiatives and also represents the development vision continuation, outlined by the Lisbon Strategy, as well as an attempt to prevent the slowdown of European economy growth, to counteract crisis effects which brought about the highest, in almost 80 years, economic downturn and unveiled extensive structural weaknesses of European economies.

The problem of R&D specialization is of particular significance for these regions and countries which do not play the role of leaders in each of the major domains in science and technology. Many regional specialists and politicians claim that these regions and countries need higher intensity of knowledge oriented investments in the form of high education and professional trainings quality, outlays on R&D (both public and private) and other activities related to innovation.

However, a question arises whether this presents a better alternative for the so far conducted policy, according to which even modest outlays on investments are necessary in many fields of technology, or in pioneering research (e.g. in biotechnology, information technology, nanotechnology), which unfortunately may not result in sig-

nificant effects in one particular area because of financial means dispersion. In view of the above observations, support and encouragement for investing in programmes, which may function as the supplement to assets held in a region or a country, seems a promising strategy aimed at creating future nationwide opportunities and establishing advantage in regional space.

As the result of smart specialization one should expect diversity among regions as the replacement of the system which creates more or less the same output in every region in a reproductive way, which also favours over extensive correlation and replication of R&D and educational investment programmes, which again reduces possibilities for supplementing each other within the framework of European knowledge base. Smart specialization represents both a concept and a tool supporting regions and countries in their response to the key question about their unique position in knowledge-based economy.

Research results have proved that we face a large number of regions which membership in a distinguished regional classes in Europe, regarding smart growth level, turns out difficult to define. Therefore it confirms that the approach to regions' classification, accepted in the hereby study, which consists in applying fuzzy classification methods is founded, since these methods provide much more additional information about classified regions than it case of classical methods.

Among regions of undefined membership, at the set threshold, 74 regions were listed (more than every fourth EU region), out of which 55 are from EU 15 (26,3%) and 19 from EU 12 (33,9%). The largest group covers German and French regions (11 each), and also Greek (8), Polish (7), Italian and Bulgarian (5 each), Belgian (4), Finnish, Portuguese, British and Romanian (3 each), 2 from Austria, 2 from Spain, and one from each of such countries as: The Netherlands, Sweden, Hungary, Slovenia, Luxemburg, Estonia and Malta. These regions may be defined as the ones "searching" for an optimum intelligent specialization path. They should become objects of particular attention and care for the individuals involved in managing development and responsible for allocating funds, as well as accountable for both interregional and intraregional policy. On the other hand, the regions classified in the third (mid and high-tech industry) and the fourth (knowledge-intensive services) class should be supported in their further development and activities have to be taken up in order to maintain their present development pace.

According to the authors the analysis of transformations dynamics and directions regarding regional membership to particular classes, as well as analyzing changes of such membership in particular EU countries, constitute an interesting focus for further studies.

Wrocław University of Economics

LITERATURE

[1] *A Digital Agenda for Europe*, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, EUROPEAN COMMISSION, COM(2010) 245 final/2, Brussels, 2010.

[2] Bezdek J.C., [1981], *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

[3] Bock H.-H., [2008], *Origins and extensions of the k-means algorithm in cluster analysis*, "Electronic Journal for History of Probability and Statistics", vol. 4, no. 2.

[4] Cox E., [2005], *Fuzzy modeling and genetic algorithms for data mining and exploration*, Morgan Kaufmann Publishers, San Francisco.

[5] *Digital Agenda Scoreboard*, Commission Staff Working Paper, SEC(2011) 708, European Commission, Brussels, 2011.

[6] Domańska W., [2010], *Strategia rozwoju Europy do 2020* r., „Wiadomości Statystyczne" no. 8 (591).

[7] Dunn A., [1973], *Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters*, "Journal of Cybernetics", vol. 3.

[8] *EUROPA 2020. Strategia na rzecz inteligentnego i zrównoważonego rozwoju sprzyjającego włączeniu społecznemu*. Komisja Europejska, Komunikat Komisji, KOM(2010) 2020 wersja ostateczna, Bruksela 2010.

[9] Giannitsis A., Kager M., [2009], *Technology and specialisation: Dilemmas, options and risks*?, Expert group "Knowledge for Growth".

[10] Höppner F., [1999], *Fuzzy cluster analysis: methods for classification, data analysis, and image recognition*, John Wiley&Sons, Chichester.

[11] *Innovation Union Competitiveness report*, Directorate-General for Research and Innovation, Directorate-General for Research and Innovation, Research and Innovation, European Commission, Luxembourg: Publications Office of the European Union, 2011.

[12] Jajuga K., [1990], *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa.

[13] Lasek M., [2002], *Data Mining. Zastosowania w analizach i ocenach klientów bankowych*. Oficyna Wydawnicza „Zarządzanie i finanse", Biblioteka Menadżera i Bankowca, Warszawa.

[14] Leung Y., [1983], *Fuzzy Sets Approach to Spatial Analysis and Planning, a Nontechnical Evaluation*, "Geografiska Annaler. Series B", vol. 65, no. 2.

[15] Nascimento S., Mirkin B., Moura-Pires F., [2000], *A fuzzy clustering model of data and fuzzy c-means*. IEEE International Conference on Fuzzy Systems: Soft Computing in the Information Age, vol. 1.

[16] Pawełek B., [2008], *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*, Wydawnictwo UE w Krakowie, Kraków.

[17] *Polityka regionalna jako czynnik przyczyniający się do inteligentnego rozwoju w ramach strategii Europa 2020*, Komunikat Komisji do Parlamentu Europejskiego, Rady, Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów, KOM(2010) 553, Bruksela, 2010.

[18] *Regions in the European Union. Nomenclature of territorial units for statistics NUTS 2006/EU-27*, Series: Methodologies and Working Papers, European Commission, Luxemburg 2007.

[19] Sokołowski A., [1992], *Empiryczne testy istotności w taksonomii*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.

[20] Steinley D., [2006], K-means clustering: a half century synthesis, „British Journal on Mathematical and Statistical Psychology", vol. 59.

[21] *Strategia na rzecz inteligentnego i zrównoważonego rozwoju sprzyjającego włączeniu społecznemu*, Krajowy program reform Europa 2020, Ministerstwo Gospodarki, Warszawa 2010.

[22] Walesiak M., Gatnar E. (red.), [2009], *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa.

[23] Wintjes R., Hollanders H., [2010], *The regional impact of technological change in 2020 – Synthesis report,* European Commission, DG Regional Policy, Brussels.
[24] Wysocki F., [2010], *Metody taksonomiczne w rozpoznawaniu typów ekonomicznych rolnictwa i obszarów wiejskich*, Wydawnictwo Uniwersytetu Przyrodniczego w Poznaniu, Poznań.
[25] *Youth on the Move*, Publications Office of the European Union, European Union, Luxembourg, 2010.
[26] Zimmermann J.H., [2001], *Fuzzy set theory and its applications. Fourth edition*, Kluwer Academic Publishers, Boston.

OCENA INTELIGENTNEGO ROZWOJU REGIONÓW EUROPEJSKICH Z ZASTOSOWANIEM KLASYFIKACJI ROZMYTEJ

Streszczenie

„Strategia Europa 2020" stanowi wizję rozwoju gospodarki europejskiej, dla której jednym z priorytetów jest rozwój inteligentny czyli oparty na wiedzy i innowacjach. Istotnym elementem inteligentnego rozwoju jest inteligentna specjalizacja obejmująca przedsiębiorstwa, ośrodki badawcze oraz szkoły wyższe, które współpracują na rzecz określenia najbardziej obiecujących obszarów specjalizacji w danym regionie. Stanowi ona zarówno koncepcję jak i narzędzie pozwalające regionom i krajom ocenić ich unikalną pozycję w gospodarce opartej na wiedzy. Trudno przecenić tą wiedzę na etapie formułowania założeń polityk regionalnych i interregionalnych oraz ustalania kierunków dystrybucji środków finansowych przeznaczonych na dalszy rozwój regionów budujących swoją przewagę w przestrzeni regionalnej oraz pozycję w gospodarce opartej na wiedzy. Dlatego zasadniczym celem niniejszego opracowania było wyodrębnienie klas regionów w przestrzeni europejskiej ze względu na zjawisko złożone jakim jest inteligentna specjalizacja. W tym celu zastosowano klasyczne i rozmyte metody klasyfikacji. Podejście takie umożliwiło m.in. wskazanie tych regionów, dla których nie można jednoznacznie określić przynależności do wyodrębnionych klas. Są to regiony „poszukujące" optymalnej ścieżki inteligentnego rozwoju, które winne zostać otoczone szczególną uwagą przez podmioty zarządzające rozwojem zarówno na szczeblu regionalnym, krajowym jak i całej wspólnoty europejskiej.

Słowa kluczowe: rozmyta klasyfikacja regionów, rozmyta metoda *c*-średnich, Europa 2020, rozwój inteligentny regionów

FUZZY CLASSIFICATION OF EUROPEAN REGIONS IN THE EVALUATION OF SMART GROWTH

Abstract

"Europe 2020 Strategy" presents the vision of European economy development, in which smart development, i.e. development based on knowledge and innovation, constitutes one of major priorities. Smart specialization which refers to enterprises, research centres and high schools cooperating in defining the most promising areas of specialization in a given region, represents one of crucial smart development

components. Smart specialization refers to both, the concept and the tool, allowing regions and countries to assess their unique position in knowledge-based economy. This knowledge should not be underestimated at the stage of preparing regional and interregional policy assumptions and specifying directions for the distribution of financial means allocated to further development of regions, constructing their advantage in regional space and position in knowledge based economy. Therefore, the essential objective of the hereby study is to distinguish classes of regions in European space with regard to one complex phenomenon, i.e. smart specialization. For this reason both classical and fuzzy classification methods were applied. Such approach facilitated e.g. specifying these regions for which it is difficult to provide clear division regarding their membership in distinguished classes. They are the regions which "keep searching" for their optimum path of smart development and which should be offered particular attention by entities managing development at regional, national and overall EU level.

Key words: fuzzy region classification, fuzzy *c*-means, Europe 2020, smart growth of regions

PRZEGLĄD STATYSTYCZNY
R. LIX – ZESZYT 1 – 2012

SPRAWOZDANIA

KRZYSZTOF JAJUGA, MAREK WALESIAK, FELIKS WYSOCKI

SPRAWOZDANIE Z KONFERENCJI NAUKOWEJ NT. „KLASYFIKACJA I ANALIZA DANYCH – TEORIA I ZASTOSOWANIA"

W dniach 21-23 września 2011 roku w Hotelu Pietrak w Wągrowcu odbyła się XX Jubileuszowa Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXV Konferencja Taksonomiczna) nt. Klasyfikacja i analiza danych – teoria i zastosowania, organizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, Katedrę Finansów i Rachunkowości Wydziału Ekonomiczno-Społecznego Uniwersytetu Przyrodniczego w Poznaniu oraz Katedrę Statystyki Wydziału Informatyki i Gospodarki Elektronicznej Uniwersytetu Ekonomicznego w Poznaniu. Przewodniczącymi Komitetu Organizacyjnego Konferencji byli prof. dr hab. Feliks Wysocki oraz dr hab. Elżbieta Gołata, prof. nadzw. UEP w Poznaniu, natomiast sekretarzami dr inż. Aleksandra Łuczak i dr Izabela Kurzawa.

Zakres tematyczny konferencji obejmował zagadnienia:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczym celem konferencji SKAD była prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 99 osób. Byli to pracownicy oraz doktoranci następujących uczelni: Akademii Górniczo-Hutniczej w Krakowie, Państwowej Wyższej Szkoły Zawodowej im. Prezydenta Stanisława Wojciechowskiego w Kaliszu, Politechniki Białostockiej, Politechniki Łódzkiej, Politechniki Opolskiej, Politechniki Rzeszowskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krako-

wie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Łódzkiego, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwersytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, Wyższej Szkoły Bankowej we Wrocławiu, Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie, a także przedstawiciele NBP, Urzędu Statystycznego w Poznaniu oraz Wydawnictwa C.H. Beck.

W trakcie 3 sesji plenarnych oraz 11 sesji równoległych wygłoszono 64 referaty poświęcone aspektom teoretycznym i aplikacyjnym zagadnień klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 19 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli: prof. dr hab. Józef Pociecha; prof. dr hab. Krzysztof Jajuga; dr hab. Andrzej Sokołowski, prof. nadzw. UEK; prof. dr hab. Iwona Roeske-Słomka; prof. dr hab. Mirosław Krzyśko; dr hab. Janusz Korol, prof. nadzw. US; prof. dr hab. Tadeusz Kufel; dr hab. Barbara Pawełek, prof. nadzw. UEK; dr hab. Andrzej Młodak; dr hab. Jan Paradysz, prof. nadzw. UEP; prof. dr hab. Stanisława Bartosiewicz; prof. dr hab. Eugeniusz Gatnar; prof. dr hab. Jadwiga Suchecka; dr hab. Elżbieta Gołata, prof. nadzw. UEP.

Teksty referatów przygotowane w formie recenzowanych artykułów naukowych stanowią zawartość przygotowywanej do druku publikacji z serii Taksonomia nr 19 (w ramach Prac Naukowych Uniwersytetu Ekonomicznego we Wrocławiu). Zaprezentowano następujące referaty:

Stanisława Bartosiewicz, Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej

Referat jest uzupełnieniem artykułu pt. *Opowieść o skutkach subiektywizmu w analizie wielowymiarowej*. Informuje o skutkach subiektywizmu w wyborze metod klasyfikacji obiektów ze względu na zjawisko złożone. Skutkami subiektywizmu są oczywiście zróżnicowane wyniki klasyfikacji. Wnioski z przeprowadzonych badań są następujące: 1) różnice w klasyfikacji obiektów zależą od liczby założonych klas: im więcej klas, tym większe różnice; 2) skutki subiektywizmu najwyraźniej występują przy wyborze metody klasyfikacji (taksonomia wrocławska znacznie mniej różnicuje wyniki, niż metoda środków ciężkości), na drugim miejscu częściej notuje się skutki subiektywizmu przy wyborze rodzaju odległości, na trzecim (ostatnim) miejscu niekiedy, (rzadziej niż przy wyborze rodzaju odległości) pojawiają się skutki subiektywizmu przy wyborze metody normalizacji cech zjawiska złożonego.

Andrzej Sokołowski, Q uniwersalna miara odległości

W zadaniach taksonomicznych coraz częściej mamy do czynienia z danymi zawierającymi zmienne mierzone w różnych skalach. Jedną z możliwych strategii jest wykonanie analizy osobno dla jednorodnej grupy cech, a potem jakieś scalenie wyników. Druga strategia to zastosowanie miary odległości, która ze swej natury umożliwia jednoczesne wykorzystanie cech mierzonych w różnych skalach. Najbardziej dojrzałą

propozycją jest tu miara Marka Walesiaka GDM. W proponowanym referacie przede wszystkim wykraczamy poza tradycyjny podział cech ze względu na skale pomiaru – na nominalne, porządkowe, przedziałowe i ilorazowe. Wyróżniono 15 rodzajów cech statystycznych. W dalszej kolejności zaproponowano takie przekształcenia tych cech (lub odległości), że składowe sumy w odległości typu Manhattan przyjmują wartości z przedziału [0,1] i są niemianowane. To umożliwia policzenie odległości ogólnej.

Eugeniusz Gatnar, Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)

W referacie przedstawiono analizę jakości danych w systemach statystycznych banków centralnych, biorąc jako przykład Narodowy Bank Polski. Dobrym wskaźnikiem jakości tych danych jest pozycja bilansu płatniczego Polski, nazywana „saldo błędów i opuszczeń". Dane sprawozdawane przez instytucje finansowe oraz Ministerstwo Finansów i GUS nie były dokładne i wskazywały na występowanie wysokiego deficytu na rachunku bieżącym w stosunku do PKB. Podjęte prace przez Departament Statystyki NBP spowodowały rewizję importu w bilansie płatniczym w roku 2011 w wysokości ok. 10 mln PLN.

Marek Walesiak, Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej

W referacie scharakteryzowano trzy strategie postępowania w pomiarze odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej:

1. Kodowanie kategorii (metody: zastąpienie kategorii rangami, kodowanie liniowe lub nieliniowe), potraktowanie zmiennych porządkowych jako zmiennych mierzonych na skali metrycznej (sztuczne wzmocnienie skali pomiaru zmiennych), a następnie zastosowanie miar odległości właściwych dla danych metrycznych (odległość euklidesowa lub miejska).
2. Kodowanie kategorii (zastąpienie kategorii rangami), a następnie zastosowanie odległości bazujących na rangach (np. odległość Kendalla, odległość Podaniego).
3. Zastosowanie miar odległości wykorzystujących dopuszczalne relacje na skali porządkowej (odległość GDM2).

Przedstawiono odpowiednie formuły odległości dla poszczególnych strategii oraz omówiono ich zalety i wady.

Krzysztof Jajuga, Marek Walesiak, XXV lat konferencji taksonomicznych – fakty i refleksje

W referacie w syntetycznej formie ujęto rys historyczny konferencji taksonomicznych oraz statystykę wygłoszonych i opublikowanych referatów w podziale na poszczególne lata i ośrodki akademickie. Ponadto przedstawiono fakty i refleksje płynące z 25 dotychczasowych konferencji taksonomicznych.

Józef Pociecha, Barbara Pawełek, Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w zmieniającej się koniunkturze gospodarczej – problemy teoretyczne i praktyczne

W dotychczas stosowanych modelach bankructwa firm zakłada się, że zmienne występujące w teorii ekonomicznej są bezpośrednio obserwowane. Za źródła losowych odchyleń przyjmuje się głównie błędy w zachowaniu się obiektów lub błędy w równaniu. W pracy zaproponowano zmianę dotychczasowego ujęcia problemu na podejście alternatywne, w którym za źródło losowej zmienności uznane zostałyby błędy pomiaru. Celem pracy jest przedstawienie propozycji wykorzystania modelu SEM w analizie zagrożenia bankructwem przedsiębiorstw w zmieniającej się koniunkturze gospodarczej.

Tomasz Górecki, Mirosław Krzyśko, Funkcjonalna analiza składowych głównych

Kiedy obserwujemy dane jako funkcje czasu (np. finansowe szeregi czasowe) nazywamy je danymi funkcjonalnymi. W wielu zastosowaniach statystycznych realizacje ciągłych szeregów czasowych są dostępne jako obserwacje procesu w dyskretnych momentach czasowych. Z tego powodu kluczowym punktem jest konwersja danych dyskretnych na ciągłe funkcje. Konwersja ta może dotyczyć wygładzania. Jedną z procedur wygładzania jest reprezentacja każdej funkcji jako liniowej kombinacji K funkcji bazowych φk(Ramsay, J.O., Silverman, B.W., *Functional Data Analysis,* Springer, New York 2006). Analiza składowych głównych (PCA) ma na celu zmniejszenie oryginalnego zestawu zmiennych do mniejszego zbioru ich ortogonalnych liniowych kombinacji. Tak zredukowane dane są następnie wizualizowane (najczęściej na płaszczyźnie) przy zachowaniu maksymalnej zmienności oryginalnych danych. Odpowiednikiem PCA dla danych funkcjonalnych jest funkcjonalna analiza składowych głównych (FPCA).

Paweł Lula, Uczące się systemy pozyskiwania informacji z dokumentów tekstowych

Zasadniczym celem pracy jest prezentacja, klasyfikacja i ocena systemów pozyskiwania informacji z dokumentów tekstowych konstruowanych przy udziale mechanizmów uczących. W początkowej części pracy zdefiniowano pojęcie systemu uczącego się oraz zagadnienie pozyskiwania informacji. Następnie zaprezentowano strukturę oraz sposób funkcjonowania uczącego się systemu pozyskiwania informacji. Kluczowym elementem tego typu rozwiązań jest model zawartości informacyjnej. Jego charakterystyka i rodzaje są zasadniczym tematem kolejnego punktu pracy. Następnie przedstawiono rolę wiedzy zewnętrznej i metod uczenia maszynowego w poszczególnych rozwiązaniach. W kolejnej części referatu zaprezentowano rozważania dotyczące lokalnego lub globalnego charakteru poszczególnych rozwiązań.

Ewa Roszkowska, Zastosowanie metody TOPSIS do wspomagania procesu negocjacji

Celem pracy jest prezentacja matematycznych podstaw systemu wspomagania procesu negocjacji z wykorzystaniem procedury TOPSIS. Procedura TOPSIS umożliwia ocenę wartości ofert, ich uporządkowanie od najlepszej do najgorszej, wyznaczenie

ofert alternatywnych, szacowanie i porównanie wartości ustępstw, rozważenie poprawy kompromisu przez poszukiwanie rozwiązań Pareto optymalnych. Istotną zaletą tej metody jest prostota obliczeniowa, łatwość i przejrzystość interpretacji otrzymanych wyników, możliwość uogólnienia na zmienne lingwistyczne, przedziałowe, czy liczby rozmyte.

Andrzej Młodak, Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne

W pracy zaprezentowano porównanie koncepcji sąsiedztwa fizycznego oraz społeczno-gospodarczego obszarów przestrzennych. To pierwsze dotyczy ich położenia na mapie administracyjnej oraz ewentualnej wspólnoty granic, zaś drugie – podobieństwa w zakresie złożonego zjawiska społeczno-ekonomicznego. Praca opisuje klasyczną teorię sąsiedztwa i jego macierzy oraz ukazuje jak można zaadoptować ją w przypadku wielowymiarowej analizy danych. Na zakończenie zaproponowano efektywną metodę porównywania obu typów sąsiedztwa wykorzystującą znaną z algebry postać normalną Smitha macierzy całkowitoliczbowych.

Andrzej Bąk, Modele kategorii nieuporządkowanych w badaniach preferencji – zmienne i dane

Wśród mikroekonometrycznych modeli kategorii nieuporządkowanych wyróżnia się najczęściej wielomianowy model logitowy, warunkowy model logitowy i mieszany model logitowy. Podstawę rozróżnienia tych typów modeli stanowi głównie charakter zmiennych objaśniających uwzględnionych w modelu. Rozróżnienie to nie jest jednak jednoznacznie interpretowane. Celem referatu jest wskazanie podstawowych różnic między typami modeli logitowych oraz prezentacja przykładów estymacji różnych typów tych modeli dla różnych typów danych z wykorzystaniem programu R.

Jacek Kowalewski, Zintegrowany model optymalizacji badań statystycznych

Referat stanowi próbę sprecyzowania teoretycznego modelu, który pozwalałby na przeprowadzenie optymalizacji procesu badań statystycznych. Złożoność zagadnienia powoduje, że prezentowany model ma charakter wieloetapowego zagadnienia wielokryterialnego z uwzględnieniem aspektów jakościowych, ekonomicznych oraz aktualności informacji. Część rozważań poświęcono przejściu z kryteriów jednostkowych dla poszczególnych informacji na globalne funkcje celu dla całego systemu badań statystycznych. Przedstawiono także warunki niezbędne do praktycznego zastosowania zaprezentowanego modelu.

Jan Paradysz, Karolina Paradysz, Obszary bezrobocia w Polsce – problem benchmarkowy

W statystyce małych obszarów istnieje wiele estymatorów pośrednich, co do których powstaje problem kryteriów wyboru oraz oceny wyników estymacji. Jedno z bardziej perspektywicznych rozwiązań stwarza statystyka lokalnego rynku pracy. Badanie

aktywności ekonomicznej ludności jest systematyczne, powtarzalne, obejmuje wszystkie gospodarstwa domowe. Można go zintegrować ze statystyką ludności oraz rejestrami administracyjnymi, dzięki czemu dysponujemy względnie stałą bazą zmiennych wspomagających. Szacując wszystkie elementy rynku pracy i sumując realizacje odpowiednich estymatorów mogliśmy porównać je z liczbą ludności i ocenić ich dokładność. W wyniku tego porównania trochę lepszym okazał się estymator EB. Estymatory pośrednie charakteryzowały się też mniejszymi standardowymi błędami szacunku.

Tomasz Szubert, W co grać aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich

W podjętym opracowaniu próbowano udowodnić, że jakakolwiek strategia gry w zakładach bukmacherskich jest nieopłacalna dla graczy, a zyski z tej rozrywki czerpią jedynie firmy organizujące te zakłady. Dla realizacji postawionego zadania wykorzystano informacje o kursach meczów piłkarskich z 6 ostatnich lat w wybranych krajach europejskich. W badaniu posłużono się m.in. analizą korelacji i analizą skupień oraz regresją logistyczną. W toku przeprowadzonych analiz okazało się, że jedynie tzw. progresja z nieograniczoną, maksymalną stawką zakładu jest w stanie przynieść dodatnią stopę zwrotu z gry, ale wymaga ona zaangażowania bardzo dużego kapitału (niekiedy rzędu kilku milionów złotych), co dla wielu graczy jest barierą nie do pokonania, a ponadto w praktyce firmy bukmacherskie stosują limity takich maksymalnych stawek, blokując tym samym szanse osiągnięcia zysku przez graczy.

Izabela Szamrej-Baran, Identyfikacja ubóstwa energetycznego z wykorzystaniem metod taksonomicznych

Głównym celem referatu jest próba identyfikacji i analizy zjawiska ubóstwa energetycznego. Podjęto w nim próbę specyfikacji listy zmiennych opisujących ubóstwo energetyczne. W oparciu o nie skonstruowano ranking krajów UE. Następnie, za pomocą hierarchicznych i niehierarchicznych metod grupowania dokonano ich podziału na grupy krajów podobnych pod względem cech charakteryzujących ubóstwo energetyczne. Przeprowadzono także klasyfikację z wykorzystaniem rozmytej metody *c-średnich*.

Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski, Próba wykorzystania modelu grawitacji w analizie dojazdów do pracy

W referacie podjęto próbę wykorzystania modelu grawitacji do opisu ciążenia określonych jednostek podziału przestrzennego kraju do ośrodków miejskich, pełniących istotną rolę w analizie lokalnych i regionalnych rynków pracy. Dotychczas możliwość aplikacji modeli grawitacji w warunkach polskich była ograniczona dostępnością danych, a także ograniczonymi możliwościami technicznymi związanymi z przetwarzaniem informacji przestrzennej. Dopiero prace podjęte w 2008 roku przez Ośrodek Statystyki Miast Urzędu Statystycznego w Poznaniu wypełniły lukę informacyjną w zakresie badań dojazdów do pracy, która istniała w Polsce od 1988 roku.

Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz, Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych

Celem podjętych badań była klasyfikacja gospodarstw domowych uprawnionych ze względu na niskie dochody do korzystania z pomocy społecznej oraz próba identyfikacji czynników warunkujących skłonność gospodarstw do korzystania z tej pomocy. Przyjęto założenie, że ocena sytuacji bytowej rodziny nie zależy wyłącznie od dochodu na jednego członka rodziny. Analizowano zatem cechy dotyczące stanu posiadania konkretnych dóbr trwałego użytku oraz subiektywnej oceny sytuacji życiowej gospodarstwa domowego (cechy psychologiczne). Zastosowanie drzew klasyfikacyjnych pozwoliło na identyfikację tych zmiennych, które ze względu na wysoką zdolność dyskryminacyjną różnicowały gospodarstwa korzystające i niekorzystające z pomocy społecznej.

Hanna Dudek, Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiąganych dochodów

W referacie przedstawiono wyniki estymacji skal ekwiwalentności na podstawie danych o subiektywnej ocenie własnej sytuacji dochodowej przez gospodarstwa domowe. W tym celu zastosowano metodę częściowo uogólnionych uporządkowanych modeli logitowych. Analizę przeprowadzono na podstawie danych z badań budżetów gospodarstw domowych zrealizowanych przez GUS w 2009 roku. Proponowana metodyka jest rozszerzeniem podejścia zaprezentowanego w publikacji [Schwarze J., *Using panel data on income satisfaction to estimate equivalence scale elasticity*, „Review of Income and Wealth" 2003, 49, s. 359-372].

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka, Segmentacja szkół wyższych metodą analizy skupień versus konkurencja technologiczna ustalona metodą DEA — studium komparatywne

W referacie przedstawiono możliwość wykorzystania analizy skupień i idei konkurencji technologicznej w metodzie DEA do segmentacji szkół wyższych. Na przykładzie danych z Rankingu Szkół Wyższych „Perspektyw" i „Rzeczpospolitej", wykorzystując różne kombinacje zborów zmiennych kryterialnych oraz metod segmentacji, dokonano czterech klasyfikacji uczelni. Następnie określono charakterystyczne cechy analizowanych obiektów należących do wyłonionych jednorodnych grup. Dokonano również oceny zgodności wyników klasyfikacji, wykorzystując miarę Randa oraz sformułowano kilkukryterialne zestawienie porównujące zastosowane w badaniu metody segmentacji.

Ewa Chodakowska, Wybrane metody klasyfikacji w konstrukcji ratingu szkół

W referacie zaprezentowano możliwość wykorzystania analizy skupień (metoda Warda, k-średnich) oraz metod porządkowania liniowego (odległość euklidesową, miarę Hellwiga, miarę GDM Walesiaka) do konstrukcji ratingu szkół. Oceny podobieństw wyników klasyfikacji dokonano za pomocą miar zgodności Randa i Nowaka. Przeprowa-

dzono walidację wyników grupowania, wykorzystując indeks silhouette. Zinterpretowano wyniki otrzymanej klasyfikacji. Celem badania było zweryfikowanie hipotezy o użyteczności wybranych metod klasyfikacji do ratingu efektywności nauczania w szkołach.

Tomasz Górecki, Klasyfikacja szeregów czasowych bazująca na pochodnych

W ostatnich latach popularność szeregów czasowych stale wyrasta. Biorąc pod uwagę szerokie wykorzystanie nowoczesnych technologii informacyjnych, duża liczba szeregów czasowych powstaje w biznesie, medycynie, biologii itp. W konsekwencji nastąpił znaczny wzrost zainteresowania w badaniu tego typu danych, co z kolei spowodowało wysyp wielu prac wprowadzających nowe metody indeksowania, klasyfikacji, grupowania oraz aproksymacji szeregów czasowych. W szczególności, wprowadzono wiele nowych miar odległości pomiędzy szeregami. W pracy zaproponowana została nowa miara odległości bazująca na pochodnych funkcji. W przeciwieństwie do dobrze znanych z literatury podejść, miara ta uwzględnia również kształt szeregu czasowego. W celu kompleksowego porównania, przeprowadzone zostały eksperymenty badające wydajność proponowanej miary na 20 szeregach czasowych z różnych dziedzin. Eksperymenty wykazały, że nasza metoda zapewnia wyższą jakość klasyfikacji.

Bartosz Soliński, Sektor energetyki odnawialnej w krajach UE – klasyfikacja w świetle strategii zarządzania zmianą

Celem referatu jest sklasyfikowanie sektora energetyki odnawialnej w poszczególnych krajach UE, na klasy obserwacji o podobnych właściwościach, dające możliwość wyodrębnienia krajów mających główny wpływ na sukces wykorzystania odnawialnych źródeł energii w świetle strategii zarządzania zmianą. W analizach przeprowadzonych w referacie zostały wykorzystane metody analizy skupień (tzw. drzewa klasyfikacyjne), które pozwoliły na uzyskanie jednorodnych grup krajów będących przedmiotem badania, w ramach których nastąpiło wyodrębnienie ich zasadniczych cech.

Elżbieta Gołata, Grażyna Dehnel, Rejestry administracyjne w analizie przedsiębiorczości

Celem referatu jest ocena możliwości zastosowania metodologii statystyki małych obszarów do szacunku podstawowych charakterystyk ekonomicznych dotyczących małych, średnich i dużych przedsiębiorstw, na podstawie informacji zawartych w źródłach statystycznych i administracyjnych. Dokonano oceny użyteczności informacji zawartych w rejestrach administracyjnych pod kątem ich wykorzystania w estymacji pośredniej typu GREG czy EBLUP. Estymacja ta pozwala na dostarczenie szacunków na niskim poziomie agregacji, na które popyt nieustannie rośnie.

Tadeusz Kufel, Marcin Błażejowski, Paweł Kufel, Modelowanie niefinansowych danych transakcyjnych (tickowych)

W referacie zostały przedstawione płaszczyzny modelowania niefinansowych danych transakcyjnych (tickowych). Niefinansowe dane transakcyjne dotyczące danych

sprzedażowych – danych paragonowych wykorzystywane są w analizie koszykowej. Analiza czasów trwania obsługi klienta, czyli obsługi transakcji, a także przerw w obsłudze klienta pozwala na optymalizację pracy kanałów obsługi klientów. Analiza cen transakcyjnych sprzedaży nieruchomości i pojazdów pozwala na ocenę roli poszczególnych cech nieruchomości (pojazdów) i wyznaczanie cen ofertowych. Normowanie danych transakcyjnych poprzez agregowanie do danych o jednakowych jednostkach czasu (dane minutowe, godzinowe, dobowe, tygodniowe) pozwala na ocenę natężenia zjawiska w czasie, co umożliwia ocenę cykliczności zjawisk.

Katarzyna Chudy, Marek Sobolewski, Kinga Stępień, Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce

Badania dotyczące rentowności banków wymagają zgromadzenia odpowiednich danych finansowych. Informacje na temat działalności banków są prezentowane w różnych źródłach i różnych okresach sprawozdawczych. Główny problem, jaki występuje na tym etapie to dostęp do ujednoliconych i aktualnych danych. Rozwiązaniem tego problemu może być zastosowanie metod prognostycznych bazujących na danych historycznych. Celem referatu jest skonstruowanie prognozy wartości wskaźników rentowności ROA i ROE dla banków działających na GPW w Warszawie na lata 2010 i 2011. Analizą zostało objętych 14 banków giełdowych. Prognoza została ustalona dla grup banków (metoda Warda), które w 2007 i 2009 roku charakteryzowały się podobnymi wartościami wskaźników ROA i ROE.

Katarzyna Dębkowska, Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod statystycznej analizy wielowymiarowej

Celem referatu jest ocena sprawności wybranych metod wielowymiarowej analizy statystycznej w prognozowaniu upadłości przedsiębiorstw. Porównano wyniki klasyfikacji trzech metod: drzew klasyfikacyjnych, regresji logitowej oraz analizy dyskryminacyjnej. W ramach badania stworzono bazę polskich przedsiębiorstw reprezentujących różne sektory, wśród których znaleźli się zarówno bankruci jak i niebankruci, a proporcja między jednymi a drugimi wyniosła 1:1. Każde przedsiębiorstwo zostało opisane za pomocą zmiennych diagnostycznych w postaci wskaźników finansowych. Dane do analizy zebrano na podstawie informacji zamieszczonych w bazie Emerging Markets Information Service (EMIS).

Alina Bojan, Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji

Inwestor ma wiele możliwości na ulokowanie swego kapitału. Są to np. inwestycje w akcje wybranych branż, surowce, waluty czy też obligacje. Wraz ze zmieniającą się sytuacją gospodarczą zmieniać się może również ich atrakcyjność. Dlatego ważne jest umiejętne reagowanie na docierające sygnały i realokacja kapitału w odpowiednim momencie. W referacie dokonano próby identyfikacji zmiennych, które wpływają na to, iż dana grupa inwestycji traci na swej atrakcyjności w porównaniu do pozostałych.

Najpierw za pomocą metod klasyfikacyjnych wyróżnione zostały najbardziej i najmniej atrakcyjne grupy inwestycji z punktu widzenia stopy zwrotu i ryzyka, a następnie wyodrębnione zostały zmienne istotnie wpływające na zmianę segmentu wybranej inwestycji w czasie.

Justyna Brzezińska, Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE

Analiza logarytmiczno-liniowa jest metodą badania niezależności zmiennych niemetrycznych, która pozwala na analizę dowolnej liczby zmiennych przy jednoczesnym uwzględnieniu interakcji zachodzących pomiędzy nimi. Modele, które budowane są hierarchicznie, oceniane są za pomocą statystyki chi-kwadrat, ilorazu wiarygodności oraz kryterium informacyjnego *AIC* oraz *BIC*. W niniejszym referacie zaprezentowane zostanie także kryterium Aitkina. Spośród wszystkich modeli wybrany zostaje model o najmniejszej złożoności, który jest dobrze dopasowany do danych. W programie **R** analiza logarytmiczno-liniowa dostępna jest dzięki funkcji loglm w pakiecie MASS. W niniejszym referacie analiza ta zaprezentowana zostanie do analizy przyczyn umieralności w krajach UE.

Aneta Rybicka, Bartłomiej Jefmański, Marcin Pełka, Analiza klas ukrytych w badaniach satysfakcji studentów

Badania satysfakcji klientów stanowią jedno z ważniejszych zagadnień poruszanych w ramach badań marketingowych. Prowadzenie tego typu badań oraz analiza ich wyników wymaga od badacza m.in. dużej wiedzy statystycznej bowiem obszar ten dotyka zarówno zagadnienia zmiennych ukrytych, jak i sposobów analizy zmiennych, których wartości mierzone są na skalach niemetrycznych. Dlatego celem opracowania jest charakterystyka możliwości zastosowania modeli klas ukrytych w badaniach satysfakcji konsumentów, ponieważ spełniają one wszystkie postulaty stawiane analizie danych gromadzonych w ramach badań satysfakcji oraz pozwalają modelować zależności między zmiennymi ukrytymi. Zaprezentowane w referacie badanie empiryczne dotyczące pomiaru i analizy satysfakcji studentów jednego z wydziałów Uniwersytetu Ekonomicznego we Wrocławiu pozwoliło scharakteryzować możliwości aplikacyjne jednego z trzech głównych typów modeli klas ukrytych.

Marta Jarocka, Zastosowanie wybranych metod wielowymiarowej analizy porównawczej w hierarchizacji polskich uczelni

Celem referatu jest przedstawienie możliwości wykorzystania wybranych metod wielowymiarowej analizy porównawczej w konstrukcji rankingu szkół wyższych. Zbiór potencjalnych zmiennych diagnostycznych opracowany przez Kapitułę Rankingu Szkół Wyższych „Perspektyw" i „Rzeczpospolitej" zweryfikowano ze względu na ich wartość informacyjną. Do analizy zdolności dyskryminacyjnej zmiennych zastosowano klasyczny i medianowy współczynnik zmienności, zaś do zbadania stopnia skorelowania zmiennych wykorzystano metodę parametryczną zaproponowaną przez Z. Hellwiga.

Do budowy rankingu uczelni wykorzystano jedną z metod wielowymiarowej analizy porównawczej – metodę wzorcową tzw. miarę Hellwiga.

Alicja Grześkowiak, Agnieszka Stanimir, Badanie efektów kształcenia i warunkujących je czynników

W referacie podjęto próbę oceny wyników egzaminacyjnych osiąganych przez uczniów. Z jednej strony zanalizowano punktacje osiągane przez piętnastolatków w badaniu PISA, z drugiej strony przeprowadzono ocenę krajowych osiągnięć uczniów w trakcie egzaminu gimnazjalnego. Każdy z typów egzaminu ocenia zbliżoną grupę uczniów w podobnych obszarach wiedzy. Zanalizowanie wyników młodych Polaków wraz z wynikami uczniów z innych krajów jest niewątpliwie użyteczne ze względu na ciągłe modyfikacje naszego systemu kształcenia. Jednoczesne wykorzystanie w analizie uzyskanych przez uczniów punktów i innych czynników, które mogą wpływać na poziom wiedzy również jest wskazówką do prowadzenia dalszych zmian systemowych.

Bartłomiej Jefmański, Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska **R**

Opracowanie stanowi propozycję pomiaru opinii respondentów polegającą na połączeniu zastosowania dyskretnych skal szacunkowych oraz liczb rozmytych. Zaprezentowano dwa warianty konstrukcji tzw. pytań pomocniczych w kwestionariuszu ankiety, które umożliwiają przekształcenie wyrażeń lingwistycznych w liczby rozmyte (np. o trójkątnych lub trapezoidalnych funkcjach przynależności). Proponowane podejście zastosowano w badaniu opinii doradców zawodowych na temat jakości kształcenia ustawicznego w podregionie wałbrzyskim. Wyniki analizy potwierdziły różnice w interpretacji przez respondentów wartości lingwistycznych stanowiących punkty na zastosowanej w opracowaniu skali szacunkowej. Operacje arytmetyczne na liczbach rozmytych niezbędne do analizy danych ankietowych z zastosowaniem proponowanego podejścia przeprowadzono w programie **R**.

Julita Stańczuk, Porównanie metod statystyki klasycznej i technik sztucznych sieci neuronowych w wielostanowej klasyfikacji obiektów ekonomicznych

Głównym celem referatu jest analiza porównawcza jakości klasyfikacji przedsiębiorstw notowanych na Giełdzie Papierów Wartościowych w Warszawie z wykorzystaniem analizy dyskryminacyjnej oraz sztucznych sieci neuronowych. Wykazano, że sieci neuronowe, a konkretniej perceptron wielowarstwowy, są lepszym klasyfikatorem dla wielostanowej klasyfikacji obiektów ekonomicznych niż analiza dyskryminacyjna.

Daniel Papla, Przykład zastosowania metod analizy wielowymiarowej w badaniu kryzysów finansowych

Kryzysy finansowe są ważnym zjawiskiem dla całej gospodarki, ponieważ podczas kryzysu zwiększają się koszty pośrednictwa oraz koszty kredytów, trudniejszy jest również dostęp do kredytów. Powoduje to ograniczenia działalności sektora realnego, co może prowadzić do kryzysu również i w tym sektorze. Dosyć duża częstość

występowania kryzysów finansowych może prowadzić do wniosku, że sektor finansowy jest szczególnie wrażliwy na różnego rodzaju zaburzenia. Zwłaszcza kryzys z lat 2008-2010 pokazał, jak gospodarka światowa jest wrażliwa na zaburzenia w epoce globalizacji. Aby zbadać uwarunkowania rozprzestrzeniania się kryzysów finansowych wykorzystano w tym referacie metody analizy powiązań między światowymi rynkami kapitałowymi, takie jak wielowymiarowe rozkłady warunkowe, funkcje powiązań czy też dynamiczne modele warunkowej korelacji. Powinny one dać odpowiedź na pytanie, czy podczas kryzysu występuje istotna zmiana w zależnościach między rynkami, co w części tłumaczyłoby tak szybkie rozprzestrzenianie się kryzysu. W pierwszej części pracy omówiono obszerną literaturę dotyczącą kryzysów finansowych oraz metody wykorzystane w dalszej części referatu. Druga część zawiera wyniki badań empirycznych, a w ostatniej zawarta jest interpretacja otrzymanych wyników.

Jerzy Krawczuk, Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500

Kluczowe w prognozowaniu zachowania rynków finansowych jest określenie kierunku zmiany, czyli określenie czy nastąpi wzrost, czy spadek. Narzędziem, które może zostać wykorzystane do tego celu są używane w eksploracji danych klasyfikatory. Klasyfikator na podstawie zbioru uczącego (danych historycznych), może przydzielić klasę wzrostu bądź spadku. Skuteczność takiej klasyfikacji (prognozy) dla indeksu giełdy amerykańskiej S&P500 została zbadana w niniejszym referacie. Użyto siedmiu popularnych klasyfikatorów, w tym drzew decyzyjnych i klasyfikatora SVM.

Anna Czapkiewicz, Beata Basiura, Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych w oparciu o modele *Copula*-GARCH

W pracy przedstawiono badanie symulacyjne, dotyczące badania poprawności przyjętej metody grupowania w oparciu o parametr wyznaczony z modelu Copula-GARCH. Ponadto zbadano wpływ zaburzeń rozkładów warunkowych procesu GARCH na wynik klasyfikacji. W szczególności zbadano jaki wpływ na wynik klasyfikacji ma nieuwzględnienie istniejącej skośności w modelowaniu szeregów czasowych.

Radosław Pietrzyk, Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych

Celem referatu jest zbadanie umiejętności zarządzających polskimi funduszami inwestycyjnymi. Efektywność inwestycji zostanie w referacie oceniona pod kątem umiejętności wykorzystania trendów rynkowych i dostosowywania strategii inwestycyjnych do zmieniającej się sytuacji na rynku giełdowym. Drugim czynnikiem oceny efektywności inwestycji jest dodatkowa stopa zwrotu z tytułu doboru papierów wartościowych. Pozwala ona na porównanie portfeli inwestycyjnych między sobą oraz z portfelami zarządzanymi pasywnie.

Aleksandra Witkowska, Marek Witkowski, Zastosowanie metody Panzara-Rosse'a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych

W pracy podjęto próbę pomiaru poziomu konkurencji w sektorze banków spółdzielczych funkcjonujących w Polsce, jako że poziom konkurencji ma istotne znaczenie dla stabilności tych banków, wpływa bowiem na ich dochodowość oraz jakość i dostępność oferowanych produktów. Dążono w niej do uzyskania odpowiedzi na następujące pytania:

1. Jaki był poziom i typ konkurencji w badanym segmencie banków spółdzielczych i jak dalece był on zmienny w czasie,

2. Czy zmiany w poziomie konkurencji są powiązane ze zmianami poziomu koncentracji w przedmiotowym segmencie banków spółdzielczych.

Jako narzędzie badawcze zastosowano jedną z metod WAS – metodę Panzara-Rosse'a.

Marcin Pełka, Podejście wielomodelowe z wykorzystaniem metody boosting w analizie danych symbolicznych

Celem referatu jest zaprezentowanie możliwości wykorzystania metody *boosting* w agregacji modeli dla danych symbolicznych z zastosowaniem metody *k*-najbliższych sąsiadów dla danych symbolicznych jako klasyfikatora bazowego. W referacie przedstawiono podstawowe pojęcia z zakresu analizy danych symbolicznych, metody *k*-najbliższych sąsiadów dla danych symbolicznych. Zaprezentowano w nim także możliwość zastosowania podejścia wielomodelowego *boosting* dla danych symbolicznych. W części empirycznej przedstawiono zastosowanie podejścia wielomodelowego dla danych symbolicznych dla przykładowych zbiorów danych.

Justyna Wilk, Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych

Celem referatu jest ocena przydatności dostępnego na rynku oprogramowania statystycznego w klasyfikacji danych symbolicznych. W referacie wyjaśniono podstawowe pojęcia analizy danych symbolicznych. Wskazano metody analizy skupień i procedurę klasyfikacji danych symbolicznych. Scharakteryzowano podstawowe własności pakietów stosowanych w analizie skupień i zakres oprogramowanych metod. Określono ich użyteczność w poszczególnych etapach procedury klasyfikacji danych symbolicznych. Z uwagi na relatywnie duże możliwości aplikacyjne, bliżej scharakteryzowano pakiety R i SODAS.

Tomasz Bartłomowicz, Justyna Wilk, Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych

Treścią referatu jest propozycja wykorzystania metodologii analizy danych symbolicznych w filtrowaniu dziedzinowych baz danych. Proponowane rozwiązanie, obok zmiennych klasycznych uwzględnia zmienne symboliczne, które opisują obiekty bez utraty informacji. Ponadto w rozwiązaniu wykorzystano znormalizowaną miarę Ichino-Yaguchiego dla danych symbolicznych. W opinii autorów, połączenie to umożliwia przeszukiwanie baz danych w oparciu o wszystkie możliwe kryteria bez względu na

rodzaj występujących zmiennych. W praktyce oznacza to rozszerzenie możliwości występujących w dostępnych powszechnie mechanizmach filtrowania dziedzinowych baz danych, co zostało zaprezentowane na przykładzie przeszukiwania ofert nieruchomości.

Andrzej Magruk, Analiza skupień w metodyce badawczej foresightu

W referacie przedstawiono wyniki autorskiej – opartej na analizie skupień – klasyfikacji metod badawczych projektów foresight. Autor zidentyfikował bogatą listę metod badawczych możliwych do zastosowania w omawianych projektach. Z uwagi na fakt, że dotychczas wiele metod należy do kilku typów autor pokusił się o dokonanie własnej klasyfikacji, ukazując grupy metod charakteryzujących się podobnymi cechami, odkrywając jednocześnie nieznaną do tej pory strukturę analizowanych danych. Nowatorskie grupowanie może być pomocne, z jednej strony dla początkujących praktyków foresightu, z drugiej natomiast dla ekspertów zajmujących się w sposób profesjonalny metodyką badawczą foresight.

Kamila Migdał Najman, Propozycja hybrydowej metody grupowania opartej na sieciach samouczących

W referacie autorka dokonuje prezentacji hybrydowej metody grupowania opartej na sieciach neuronowych samouczących typu SOM i GNG. Autorka weryfikuje potencjał proponowanej hybrydowej metody grupowania typu: sieć SOM + metoda k-średnich. Proponowane podejście weryfikuje na przykładzie badania preferencji i zachowań komunikacyjnych mieszkańców Gdyni w 2010 roku.

Dorota Rozmus, Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody bagging

Stosując metody taksonomiczne w jakimkolwiek zagadnieniu klasyfikacji ważną kwestią jest zapewnienie wysokiej poprawności wyników grupowania. Stąd też w literaturze wciąż proponowane są nowe rozwiązania, które mają przynieść poprawę dokładności grupowania w stosunku do tradycyjnych metod. Przykładem mogą tu być metody polegające na zastosowaniu podejścia zagregowanego oraz algorytmy spektralne. Głównym celem tego referatu jest porównanie dokładności zagregowanych algorytmów taksonomicznych opartych na idei metody *bagging* oraz spektralnego algorytmu taksonomicznego zaproponowanego przez Ng A. Y., Jordan M. I., Weiss Y., *On spectral clustering: Analysis and an algorithm,* „Advances in Neural Information Processing Systems", 2001, 849-856.

Krzysztof Najman, Grupowanie dynamiczne w oparciu o samouczące się sieci GNG

Wraz ze stale rozwijającą się techniką informatyczną, dynamicznie zwiększa się także ilość danych zbieranych w różnych systemach komputerowych. Jedną z cech baz danych tworzonych dynamicznie w systemie on-line jest między innymi dynamicznie zmieniająca się struktura grupowa jednostek. Grupowanie jednostek w takiej bazie danych wymaga zastosowania specjalnych metod. W referacie sformułowano wymagania stawiane takiej metodzie, a także zaproponowano użycie w opisywanym celu

samouczącej się sieci neuronowej typu GNG. W referacie w oparciu o badania teoretyczne i eksperyment badawczy weryfikowano hipotezę o przydatności sieci GNG w dynamicznym grupowaniu danych.

Małgorzata Misztal, Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna

Drzewa klasyfikacyjne należą do tych algorytmów uczących, które mogą być wykorzystane w sytuacji występowania braków danych w zbiorze danych. W pracy porównano kilka wybranych technik postępowania w sytuacji występowania braków danych. Wykorzystano podejście symulacyjne generując różne proporcje i mechanizmy powstawania braków danych w zbiorach danych pochodzących z repozytorium baz danych na Uniwersytecie Kalifornijskim w Irvine oraz z badań własnych. Celem badań była ocena wpływu wybranych metod imputacji danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności.

Mariusz Kubus, Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych

Paul D., Bair E., Hastie T., Tibshirani R. [*"Pre-conditioning" for feature selection and regression in high-dimensional problems,* Annals of Statistics 36(4), 2008, s. 1595-1618] zaproponowali metodę wstępnego uwarunkowania zmiennej objaśnianej, którą realizuje się w dwóch krokach. W pierwszym przewidywane są wartości zmiennej objaśnianej za pomocą metody głównych składowych z nauczycielem. W drugim budowany jest jeszcze raz model regresji, ale na zbiorze uczącym, w którym rolę zmiennej objaśnianej pełnią jej wartości teoretyczne estymowane w kroku pierwszym. Tu wykorzystuje się LASSO, które pozwala wyeliminować zmienne nieistotne. W referacie zaproponowano modyfikację oryginalnej metody oraz za pomocą symulacji oceniono ich przydatność w zadaniu selekcji zmiennych w przypadku modelu liniowego z interakcjami.

Barbara Batóg, Jacek Batóg, Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym

W referacie wykorzystano analizę dyskryminacyjną do wyboru zbioru wskaźników (zmiennych dyskryminacyjnych) determinujących stopę zwrotu z inwestycji na rynku kapitałowym w Polsce. Wśród rozpatrywanych przekrojów badania uwzględniono okres wzrostów oraz okres spadków cen akcji, jak również wybrane sektory gospodarcze. Ocenie poddano dodatkowo trafność uzyskiwanych klasyfikacji w zależności od sposobu podziału badanych obiektów na grupy.

Katarzyna Wójcik, Janusz Tuchowski, Analiza porównawcza miar podobieństwa tekstów bazujących na macierzy częstości z bazującymi na wiedzy dziedzinowej

Zasadniczym celem niniejszej pracy jest próba oceny przydatności znanych z literatury miar podobieństwa tekstów bazujących na macierzy częstości z tymi bazującymi na wiedzy dziedzinowej w postaci ontologii. W kolejnych punktach referatu przedstawione zostały najpierw dokumenty tekstowe, które były porównywane w badaniu, a następnie wybrane miary podobieństwa oparte ma macierzy częstości i ich analiza symulacyjna. W dalszej części zaprezentowana została ontologia wykorzystana w badaniach oraz wyniki przeprowadzonej analizy symulacyjnej miar opartych na wiedzy reprezentowanej przez tą ontologię. Na tej podstawie została podjęta próba oceny przydatności tych miar.

Iwona Staniec, Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach

Celem referatu jest przedstawienie empirycznego wykorzystania analizy czynnikowej w badaniu obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach. Zastosowanie w tym przypadku analizy czynnikowej umożliwia pełniejszą diagnozę współzależności między poszczególnymi elementami systemów zarządzania, a ich istotnością dla doskonalenia oraz zależnością od branży.

Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk, Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami

W praktycznych zadaniach klasyfikacji, na przykład w analizie danych medycznych, występuje stosunkowo często konieczność wnioskowania w oparciu o dane niekompletne. Celem pracy jest porównanie efektywności wybranych podejść do rozwiązywania problemu klasyfikacji z brakującymi obserwacjami dla wybranych klasyfikatorów prostych i złożonych oraz dla różnych metod transformacji cech z brakującymi obserwacjami. W badaniach zastosowano implementacje technik eksploracji danych w środowisku STATISTICA Data Miner oraz systemie uczenia maszynowego WEKA. Jako dane do klasyfikacji wykorzystano bazę danych o pacjentach leczonych operacyjnie z powodu raka płuca we Wrocławskim Ośrodku Torakochirurgii w latach 2000-2011.

Iwona Foryś, Wielowymiarowa analiza wpływu cech mieszkań na ich atrakcyjność cenową w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego

Przedmiotem analizy jest obrót na wtórnym rynku mieszkaniowym. Pozytywnie zweryfikowano hipotezę o stałości cech jakościowych wpływających na cenę, niezależnie od sytuacji na rynku mieszkaniowym. W tym celu wykorzystano analizę logliniową. Wykazano, iż rozszerzenie modelu o kolejne zmienne oraz interakcje rzędu wyższego niż drugi nie poprawiało dopasowania modelu. W badanych latach najlepsze dopasowanie modelu przy zmiennej zależnej cena uzyskano dla lokalizacji oraz powierzchni lokalu. Najczęściej nieistotne okazały się zmienne opisujące lokalizację mieszkania w budynku oraz zbywane prawo do lokalu.

Ewa Genge, Analiza skupień oparta na mieszankach uciętych rozkładów normalnych

W referacie przedstawione zostało najnowsze podejście modelowe, polegające na usuwaniu obserwacji oddalonych, w trakcie estymacji parametrów mieszanki. W podejściu tym na macierze kowariancji nakładane są pewne ograniczenia, wyznaczane są funkcje wiarygodności uciętych zmiennych losowych, a parametry takiej mieszanki szacowane są za pomocą zmodyfikowanej wersji algorytmu EM, tj. algorytmu TCLUST [García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar A., *A review of clustering methods*, „Advances of data analysis and classification", Springer, 2010, 4, s. 89-109]. Przeprowadzone zostały również badania porównawcze z bardziej popularnym, odpornym modelem mieszanek rozkładów normalnych.

Jerzy Korzeniewski, Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień

Metodę Ichino można stosować do zmiennych mierzonych na dowolnych skalach. Pełka i Wilk zbadali tę metodę, ale badanie ograniczało się do kilku modeli struktur skupień zaś kryterium oceny była stabilność grupowania mierzona indeksem sylwetkowym. Metoda uśredniania zmiennych nieistotnych dla struktury skupień (Fraiman R., Justel A., Svarc M. (2008), Selection of Variables for Cluster Analysis and Classification Rules, JASA, 103) jest nową metodą niezbadaną dotychczas w żadnym obszerniejszym eksperymencie. W referacie przedstawione są wyniki badania symulacyjnego, w którym kryteriami są pamięć, precyzja i indeks Randa. Metody konkurencyjne, z którymi porównywana jest metoda Ichino to; HINoV oraz dwie modyfikacje HINoV (VSKM i VAF z indeksem skupialności). Andrzej Dudek, SMS – propozycja nowego algorytmu analizy skupień

Klasyfikacja spektralna (*spectral clustering* – Ng, A., Jordan, M., Weiss, Y. (2002), *On spectral clustering: analysis and an algorithm*, W: T. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, 849-856) i klasyfikacja za pomocą średniej przesunięcia okna w kierunku wektora średniej (*Mean Shift clustering* – Wang, S., Qiu, W., and Zamar, R. H. (2007), CLUES: *A non-parametric clustering method based on local shrinking. Computational Statistics & Data Analysis*, Vol. 52, issue 1, 286-298) to dwa stosunkowo nowe podejścia w analizie skupień, dające, zwłaszcza dla skupień o nietypowych kształtach lepsze rezultaty niż klasyczne metody k-średnich, k-medoidów czy hierarchiczne metody aglomeracyjne. Referat zawiera propozycję algorytmu o roboczej nazwie SMS (*Spectral-Mean Shift*) łączącego cechy obu podejść wyróżniającego się wśród innych algorytmów analizy skupień m.in.: możliwością analizy skupień o nietypowych kształtach; możliwością automatycznego rozpoznawania liczby skupień; lepszą odpornością na zmienne zakłócające.

Artur Mikulec, Metody oceny wyniku grupowania w analizie skupień

W referacie dokonano przeglądu metod oceny wyniku grupowania. Omówiono trzy metody wyboru właściwej liczby skupień dla metod aglomeracyjnych zaproponowane przez Mojenę i Wisharta oraz zaimplementowane w programie *ClustanGraphics 8*. Wspomniane kryteria bazują na relatywnych wartościach rożnych poziomów połączeń obiektów na wykresie drzewa – *best cut significance test* (*upper tail, moving average*) oraz sprawdzaniu losowości podziału obiektów na wykresie drzewa – *tree validation*. Treść referatu zilustrowana została przykładem empirycznym.

Artur Zaborski, Analiza PROFIT i jej wykorzystanie w badaniu preferencji

Celem referatu jest prezentacja analizy PROFIT będącej połączeniem skalowania wielowymiarowego oraz analizy regresji wielorakiej. Dla konfiguracji punktów reprezentujących obiekty otrzymanej za pomocą skalowania wielowymiarowego przeprowadza się analizę regresji wielorakiej, w której zmiennymi objaśniającymi są współrzędne obiektów na mapie percepcyjnej, a zmiennymi zależnymi oceny marek ze względu na poszczególne cechy. Na zakończenie zaprezentowano przykład badania preferencji marek samochodów z wykorzystaniem analizy PROFIT.

Karolina Bartos, Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena

W referacie zaprezentowano wyniki analizy skupień 86 państw ze względu na udział różnego typu wydatków konsumpcyjnych obywateli w ich wydatkach konsumpcyjnych ogółem. W badaniu uwzględniono 12 typów wydatków, takich jak wydatki na: żywność i napoje bezalkoholowe, utrzymanie domu, odzież i obuwie, transport, wypoczynek i rekreację, alkohol i wyroby tytoniowe, telekomunikację, edukację, zdrowie, usługi hotelowe i catering, artykuły gospodarstwa domowego i usługi dla domu, pozostałe dobra i usługi. Do stworzenia grup wykorzystano sieci Kohonena.

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak, Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych

Celem referatu jest klasyfikacja gospodarstw domowych na rynku ubezpieczeń w Polsce ze względu na bodźce skłaniające gospodarstwa domowe do zawierania umów ubezpieczeniowych na życie, dożycie i pokrycie kosztów leczenia. W badaniu wykorzystano drzewa klasyfikacyjne oraz dwumianowe modele logitowe, w których zmienną objaśnianą był popyt zrealizowany i potencjalny na ubezpieczenia danego rodzaju. Natomiast zmiennymi objaśniającymi były bodźce skłaniające do zawierania umów ubezpieczeniowych takie jak doświadczenia najbliższego otocznia, strach przed katastrofami, niewydolność społecznego sytemu ubezpieczeń i służby zdrowia. Analizę przeprowadzono na próbie 500 gospodarstw domowych.

Izabela Kurzawa, Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto-wieś

Celem pracy było zbadanie przydatności liniowej aproksymacji prawie idealnego systemu funkcji popytu LA/AIDS do wyznaczenia elastyczności cenowych popytu dla wybranych grup artykułów żywnościowych w relacji miasto-wieś. Parametry tego modelu oszacowano na podstawie danych z badania indywidualnych budżetów gospodarstw domowych w Polsce w 2006 roku. Na tej podstawie wyznaczono elastyczności cenowe własne i mieszane dla wybranych produktów żywnościowych. Klasa miejscowości zamieszkania gospodarstwa domowego różnicuje elastyczności cenowe.

Aleksandra Łuczak, Feliks Wysocki, Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych

W referacie przedstawiono trzy podejścia do konstrukcji syntetycznego miernika rozwoju obiektów opisanych za pomocą cech mierzonych na skalach metrycznych i porządkowej. Podejście pierwsze polega na sprowadzeniu cech metrycznych i porządkowych do określeń lingwistycznych i liczb rozmytych (osłabienie skali cech metrycznych). Podejście drugie wiąże się z zastosowaniem uogólnionej miary odległości GDM zaproponowanej przez Walesiaka (zachowanie skal pomiarowych). Podejście trzecie polega na zamianie wartości cech porządkowych (lingwistycznych) na liczby rozmyte, co prowadzi do wyrażenia wartości tych cech w skali ilorazowej (wzmocnienie skali cech porządkowych). Proponowane podejścia zostały zastosowane do porządkowania liniowego powiatów województwa wielkopolskiego według stanu i jakości infrastruktury technicznej.

Grzegorz Kowalewski, Konstrukcja wskaźników złożonych w testach koniunktury

Wskaźniki złożone w testach koniunktury są zazwyczaj konstruowane w sposób arbitralny. W referacie przedstawiono sposób konstrukcji wskaźników złożonych z wykorzystaniem informacji zawartych w danych. Rozważania przedstawiono na przykładzie badań ankietowych prowadzonych przez GUS w przemyśle przetwórczym.

Agnieszka Sompolska-Rzechuła, Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim

Celem referatu było porównanie efektywności dwóch podejść w konstrukcji miernika rozwoju wykorzystujących klasyczny miernik Hellwiga oraz pozycyjny wektor medianowy Webera w badaniu obiektywnej jakości życia w województwie zachodniopomorskim. W analizie wykorzystano wskaźniki z następujących dziedzin: środowisko naturalne, demografia, rynek pracy, infrastruktura komunalna i mieszkania, edukacja, kultura i turystyka, ochrona zdrowia, transport drogowy, dochody i wydatki budżetów jednostek samorządu terytorialnego. Badanie dotyczyło 2008 roku. Podejście klasyczne dało lepsze wyniki pod względem homogeniczności, heterogeniczności i poprawności grupowania.

Krzysztof Szwarc, Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną

Procesy demograficzne zachodzące w naszym kraju nie przebiegają jednakowo w poszczególnych województwach, powiatach, a nawet gminach. Przemiany, które mają miejsce w Polsce przejawiają się między innymi w ograniczaniu płodności, przesunięciu wieku zawierania małżeństw a także wieku rodzenia pierwszego dziecka, wydłużaniu się przeciętnego dalszego trwania życia. Konsekwencją tych zmian jest starzenie się społeczeństwa. Celem niniejszego referatu jest dokonanie klasyfikacji powiatów badanego województwa ze względu na poziom podstawowych wskaźników, charakteryzujących sytuację demograficzną. Badaniem objęto lata 2002-2009, a źródłem danych była baza demograficzna zamieszczona na stronie internetowej Głównego Urzędu Statystycznego.

Ponadto w trakcie sesji plakatowej zaprezentowano następujące referaty:

Joanna Banaś, Małgorzata Machowska-Szewczyk, Analiza ocen skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego

Celem referatu jest analiza opinii studiujących użytkowników kont pocztowych na temat posiadanych skrzynek poczty. Źródło danych stanowiły kwestionariusze ankiety wypełniane przez studentów wybranych wydziałów szczecińskich uczelni, które zawierały pytania dotyczące zarówno charakterystyki skrzynki poczty elektronicznej, jak i intensywności jej używania. Do oceny obsługi poczty elektronicznej wykorzystano uporządkowany model probitowy, w którym występuje jakościowa zmienna zależna, opisująca kategorie uporządkowane przy założeniu rozkładu normalnego składnika losowego.

Iwona Bąk, Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę

W referacie przedstawiono wyniki badań dotyczące segmentacji gospodarstw domowych emerytów i rencistów pod względem ich przeciętnych miesięcznych wydatków na rekreację i kulturę. Do klasyfikacji gospodarstw domowych wykorzystano drzewo regresyjne. Podstawę informacyjną badań stanowiły nieidentyfikowalne jednostkowe dane o dochodach i wydatkach indywidualnych gospodarstw domowych, pochodzące z badań budżetów gospodarstw domowych przeprowadzonych w 2009 r. przez GUS. Segmentacja została przeprowadzona dla reprezentatywnej próby 1308 gospodarstw domowych emerytów i 327 gospodarstw domowych rencistów.

Aneta Becker, Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010

W referacie przedstawiono wyniki uporządkowania województw Polski pod względem wykorzystania technologii informacyjno-telekomunikacyjnych w przedsiębiorstwach (ICT, ang. Information and Communication Technology), w latach 2008-2010. W badaniach wykorzystano metodę wielokryterialnego wspomagania decyzji ANP (ang. Analytic Network Process). Zastosowany algorytm pozwolił na analizę położenia

województw z uwagi na różne zaawansowanie obszarów w przestrzeni teleinformatycznej.

Katarzyna Dębkowska, Ocena sytuacji finansowej sektorów przy użyciu metod wielowymiarowej analizy statystycznej

Przy ocenie sytuacji finansowej przedsiębiorstw pomocne są informacje dotyczące sektorowych wskaźników finansowych. Celem referatu jest dokonanie klasyfikacji sektorów pod względem wskaźników finansowych, wykorzystując metody wielowymiarowej analizy statystycznej: analizę skupień oraz drzewa klasyfikacyjne. Wykorzystane metody pozwoliły na pogrupowanie sektorów ze względu na wskaźniki finansowe. Klasyfikacją objęte zostały 53 sektory ze względu na 14 zmiennych diagnostycznych w postaci wskaźników finansowych. Źródłem informacji do przeprowadzenia klasyfikacji były dane dotyczące wskaźników sektorowych za 2009 r. Otrzymana klasyfikacja może stanowić cenne źródło informacji o sektorach dla przedsiębiorstw lub inwestorów.

Małgorzata Machowska-Szewczyk, Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych

Większość opracowanych metod klasyfikacji umożliwia grupowanie obiektów, opisanych za pomocą zmiennych ustalonego typu. W praktycznych zastosowaniach wiele obiektów może być charakteryzowanych przez różne typy cech. Celem pracy jest prezentacja rozmytego algorytmu klasyfikacji obiektów, które mogą być opisane jednocześnie za pomocą zmiennych numerycznych, symbolicznych lub rozmytych. Algorytm ten został zaproponowany przez Yanga, Hwanga i Chena, którzy zdefiniowali miarę niepodobieństwa między mieszanymi obiektami oraz zmodyfikowali rozmytą metodę c-środków. W referacie przedstawiono także numeryczny przykład zastosowania tej metody do obiektów o cechach mieszanych na podstawie danych rzeczywistych.

Anna Domagała, Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru wprzód)

Referat przedstawia propozycję metody doboru zmiennych (nakładów oraz wyników) do modeli DEA (Data Envelopment Analysis). Metoda polega na stopniowym dodawaniu zmiennych do modelu DEA, zgodnie z opracowanym algorytmem. Decyzję o wprowadzeniu danej zmiennej podejmuje się na podstawie dwóch kryteriów – procentowej zmiany średniej efektywności, wywołanej dodaniem takiej zmiennej do układu oraz współczynnika korelacji pomiędzy rezultatami badania bez tej zmiennej, a wynikami badania z jej uwzględnieniem.

Nina Drejerska, Mariola Chrzanowska, Iwona Pomianek, Taksonomiczna analiza porównawcza poziomu rozwoju gmin wiejskich i miejsko-wiejskich województwa mazowieckiego w latach 2002 i 2009

Rozwój społeczno-gospodarczy zachodzący w gminach wiejskich i miejsko-wiejskich w sąsiedztwie dużych aglomeracji powoduje tworzenie się specyficznych obszarów funkcjonalnych. Wyniki porównawczej analizy poziomu rozwoju badanych gmin

metodą k-średnich umożliwiają określenie obszaru funkcjonalnego Warszawy. Rozwój lokalny jednostek administracyjnych jest mierzony głównie wskaźnikami rozwoju społeczno-gospodarczego, stąd do badań wybrano zmienne charakteryzujące aspekty społeczne, gospodarcze i infrastrukturalne oraz uwzględniono zmienną określającą odległość danej gminy od centrum Warszawy. Badaniu poddane zostały wiejskie (229) i miejsko-wiejskie (50) gminy województwa mazowieckiego w latach 2002 i 2009.

Henryk Gierszal, Karina Pawlina, Maria Urbańska, Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej

W referacie zaprezentowano możliwości wykorzystania statystycznych analiz wielu zmiennych w badaniach popytu na usługi telekomunikacyjne na przykładzie planowanej ogólnokrajowej radiowej sieci łączności dyspozytorskiej. Wyniki przedstawiono dla analizy kanonicznej, dyskryminacyjnej, czynnikowej, korespondencji oraz skupień, które pozwoliły zidentyfikować zachowania użytkowników determinujące zapotrzebowanie na tego rodzaju usługi.

Hanna Gruchociak, Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej

Głównym celem referatu jest przedstawienie przydatności metodologii modelowania dwupoziomowego w szacowaniu wartości zmiennych społeczno-gospodarczych. W pierwszej części opracowania omówiona została idea konstrukcji estymatora dla danych o strukturze dwupoziomowej. W części drugiej przeprowadzone zostało badanie empiryczne, mające na celu zastosowanie opisanego estymatora do szacowania wskaźnika zatrudnienia w przekroju powiatów. Po dokonaniu oszacowań porównano wiarogodności estymatora uwzględniającego dwupoziomową strukturę danych oraz zwykłego estymatora regresyjnego. Przeprowadzona analiza wykazała istotną poprawę jakości oszacowań uzyskanych przy zastosowaniu modelowania dwupoziomowego.

Tomasz Klimanek, Marcin Szymkowiak, Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy

Referat przedstawia propozycje wykorzystania metod estymacji pośredniej (w tym także tej metody, która uwzględnia korelację przestrzenną) do oszacowania odsetka osób bezrobotnych w populacji osób w wieku 15 lat i więcej w przekroju podregionów w Polsce w I kwartale 2008 roku. Jest to bardziej szczegółowy poziom agregacji przestrzennej niż ten prezentowany w publikacjach Głównego Urzędu Statystycznego opartych na wynikach Badania Aktywności Ekonomicznej Ludności. Drugim celem jest porównanie miar precyzji estymatora bezpośredniego z precyzją estymatora typu EBLUP (*empirical best linear unbiased predictor*) oraz estymatora typu SEBLUP (uwzględniającego korelację przestrzenną).

Wojciech Kuźmiński, Wykorzystanie metod statystycznej analizy wielowymiarowej do oceny realizacji zadań statutowych Agencji Nieruchomości Rolnych

W referacie wykorzystano metody analiz wielowymiarowych do oceny stopnia realizacji zadań stawianych Agencji Nieruchomości Rolnych. Przeprowadzono analizy klasyfikacyjne, celem określenia w których częściach kraju obowiązki wynikające z ustaw realizowane są lepiej, a w których gorzej. Podobnie wykonano grupowanie pod względem podobieństwa regionów.

Jarosław Lira, Prognozowanie opłacalności żywca wieprzowego w Polsce

W pracy analizowano trafność krótkoterminowych prognoz opłacalności produkcji trzody chlewnej na podstawie relacji cenowych żywca wieprzowego w stosunku do cen targowiskowych żyta i jęczmienia oraz cen prosiąt. Do prognozowania targowiskowych cen zbóż zastosowano model wyrównywania wykładniczego Wintersa, a do sporządzenia prognoz targowiskowych cen żywca wieprzowego oraz cen prosiąt wykorzystano zmodyfikowaną metodę klasyczną opartą na modelu multiplikatywnym. Uzyskane prognozy posłużyły do określenia relacji cenowych, które następnie poddano ocenie trafności.

Christian Lis, Wykorzystanie metod taksonomicznych w ocenie konkurencyjności portów południowego Bałtyku w kontekście pogłębienia toru wodnego Szczecin-Świnoujście do głębokości 12,5 m

W badaniu autor wykorzystał metody taksonomiczne do oceny poprawy pozycji konkurencyjnej portu w Szczecinie wśród portów południowego Bałtyku w wyniku podjęcia inwestycji, polegającej na pogłębieniu toru wodnego ze Szczecina do Świnoujścia do głębokości 12,5 m. W referacie wykorzystana została uogólniona miara odległości, dla potrzeb badania nazwana taksonomicznym miernikiem konkurencyjności portów i analiza skupień. Wyniki badań dostarczają istotnych argumentów za koniecznością podjęcia kluczowej inwestycji w ramach Programu SONORA (*South-North Axis*). Program ten ma na celu poprawę infrastruktury transportowej w korytarzu transportowym Bałtyk-Adriatyk.

Beata Bieszk-Stolorz, Iwona Markowicz, Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych

W 2010 roku z Powiatowego Urzędu Pracy w Szczecinie zostało wyrejestrowanych ponad 20 tys. osób bezrobotnych. Przyczyny były bardzo różne. Celem referatu jest analiza szansy na podjęcie pracy oraz ocena prawdopodobieństwa podjęcia pracy o określonym charakterze w zależności od cech osób bezrobotnych: płci, wykształcenia i wieku. Analizę przeprowadzono dwuetapowo. Najpierw, przy wykorzystaniu dwumianowego modelu logitowego, zbadano wpływ cech osób bezrobotnych na szansę wyjścia z bezrobocia poprzez podjęcie pracy. W drugim etapie skupiono się na ocenie prawdopodobieństwa podjęcia określonej formy zatrudnienia. W tym celu zastosowano wielomianowy model logitowy.

Ewa Piotrowska, Statystyczny portret innowacyjności i rozwoju gospodarczego województw Polski

Celem opracowania jest konstrukcja statystycznego portretu innowacyjności i rozwoju gospodarczego województw Polski w oparciu o metody wielowymiarowej analizy statystycznej. Ocenę poziomu innowacyjności regionów przeprowadzono na podstawie odpowiednio dobranych agregatów wskaźników szczegółowych bazujących na metodologii tworzenia syntetycznego indeksu SII. Analizie poddano dane liczbowe z 2008 roku. Na podstawie otrzymanego portretu innowacyjności i rozwoju regionalnego dokonano oceny pozycji województwa podlaskiego w polskiej przestrzeni regionalnej. Otrzymane wyniki mogą zostać wykorzystane w zakresie strategii rozwojowych obejmujących wskazywanie wzorca rozwoju czy niwelowanie dysproporcji rozwojowych.

Lucyna Przezbórska-Skobiej, Jarosław Lira, Przestrzeń agroturystyczna Polski i jej waloryzacja

Celem niniejszego referatu była waloryzacja przestrzeni agroturystycznej Polski na poziomie powiatów ziemskich, z wykorzystaniem cech opisujących tę przestrzeń, uznanych w literaturze jako istotne dla rozwoju agroturystyki. Do przeprowadzenia waloryzacji przestrzeni agroturystycznej wykorzystano zbudowany agregatowy miernik waloryzacji agroturystycznej. Na podstawie wartości miernika wyodrębniono i opisano 7 klas powiatów, o różnej jakości przestrzeni agroturystycznej

Roma Ryś-Jurek, Olga Stefko, Zastosowanie modeli SARIMA do prognozowania cen produktów ogrodniczych

Celem opracowania było wykazanie przydatności zastosowania modeli SARIMA do prognozowania cen w ogrodnictwie, na przykładzie dwóch produktów, a mianowicie jabłek i kapusty. Posłużono się danymi zbieranymi za pomocą wywiadu bezpośredniego przez serwis informacyjny „Fresh-market.pl", a dotyczącymi minimalnych dziennych cen sprzedaży jabłek i kapusty w zł/kg, które zebrano między 11 maja 2007, a 24 maja 2011 roku z 10 rynków hurtowych w całej Polsce. Oszacowane modele SARIMA przeszły pozytywnie weryfikację, więc wykorzystano je do celów prognostycznych, sporządzając przykładowe prognozy minimalnych cen sprzedaży jabłek i kapusty na kolejne 30 dni.

Paweł Ulman, Model rozkładu wydatków a funkcje popytu

W pracy podjęto problem wpływu zmian poziomu dochodów na kształtowanie się rozkładu wydatków na dany typ dóbr i usług, a nie jedynie na przeciętny ich poziom. Wykorzystując jako model rozkładu wydatków teoretyczny rozkład Burra typu III (Daguma) oraz uzmienniając jego parametry poprzez odpowiednie funkcje dochodów uzyskano możliwość badania wpływu zmian dochodu na kształtowanie się rozkładu wydatków. Jako funkcje dochodów przyjęto funkcję liniową, potęgową oraz funkcję Törnquista. Odpowiednie modele wydatków oszacowano na podstawie danych z budżetów gospodarstw domowych z 2009 roku przy wykorzystaniu MNW.

Maria Urbańska, Tadeusz Mizera, Henryk Gierszal, Zastosowanie metod analizy statystycznej w badaniach przyrody ożywionej

W referacie przedstawiono możliwości wykorzystania różnych metod analizy statystycznej do wnioskowania o przyrodzie ożywionej na przykładzie badań nad dwoma typami królestwa zwierząt obejmujących bezkręgowce (mięczaków) oraz kręgowców (strunowców). W pierwszym przypadku badano gromadę ślimaków (Gastropoda), zaś w drugim gatunek orła bielika *Haliaeetus albicilla*. Do porównania bogactwa malakofauny na różnych terenach wykorzystano wskaźniki podobieństw, które posłużyły do wykreślenia dendrogramów. Z kolei w badaniach nad bielikiem wykorzystano wieloczynnikową analizę wariancji MANOVA oraz regresję wielokrotną w celu określania płci ptaków na podstawie pomiarów biometrycznych.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu, po wyborze, przewodniczył prof. dr hab. Feliks Wysocki. Ustalono plan przebiegu zebrania obejmujący punkty:

1. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
2. Zapowiedzi kolejnych konferencji SKAD PTS.
3. Wybór przedstawiciela Sekcji SKAD PTS do IFCS Council.
4. Sprawy różne.

Prof. dr hab. Marek Walesiak otworzył posiedzenie Sekcji SKAD PTS. Przypomniał, że na stronie internetowej Sekcji znajduje się nowy regulamin, a także deklaracja członkowska. Poinformował, że został opublikowany zeszyt z serii „Taksonomia 18", PN UE we Wrocławiu nr 176, a także podkreślił, aby w przygotowaniu artykułów do kolejnego zeszytu uwzględnić wszystkie wymogi edytorskie zawarte w „Instrukcji dla autorów". W „Przeglądzie Statystycznym" z. 1/2 z 2011 r. ukaże się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Toruniu w dniach 15-17 września.

W następnym punkcie posiedzenia podjęto kwestię kolejnych konferencji SKAD. Mgr inż. Ewa Chodakowska, w imieniu prof. dra hab. Joanicjusza Nazarko, potwierdziła organizację przyszłorocznej konferencji SKAD. Poinformowała, że odbędzie się ona w dniach 12-14 września 2012 r. w miejscowości Borki w hotelu Lipowy Most. Prof. dr hab. Krzysztof Jajuga zasugerował pozostawienie na niezmienionym poziomie opłat konferencyjnych. W związku z tym, że nie padły deklaracje w sprawie organizacji konferencji w 2013 r., prof. dr hab. M. Walesiak zaproponował, że w przypadku braku innych propozycji, podejmie się organizacji konferencji.

W dalszej części zebrania prof. dr hab. Krzysztof Jajuga podsumował, że International Federation of Classification Societies (IFCS) skupia 13 towarzystw klasyfikacyjnych, a SKAD jest trzecim pod względem liczebności. Krótko wyjaśnił organizację i strukturę oraz sposób powoływania prezydenta IFCS. W związku z tym, że w 2011 r. kończy się kadencja prof. UEK dr hab. Andrzeja Sokołowskiego jako członka Rady i jednocześnie przedstawiciela SKAD w IFCS, należało przeprowadzić głosowanie w sprawie wyboru reprezentanta na kolejną kadencję.

Prof. dr hab. Krzysztof Jajuga zaproponował powołanie do Komisji Skrutacyjnej prof. dra hab. Tadeusza Kufla oraz prof. UE dr hab. Elżbietę Gołatę oraz zaproponował

kandydaturę prof. Andrzeja Sokołowskiego. W związku z brakiem innych kandydatur przystąpiono do głosowania. Komisja podliczyła głosy: 45 głosów za, 0 głosów przeciw, 0 osób się wstrzymało. W wyniku głosowania przyjęto kandydaturę prof. Andrzeja Sokołowskiego jako reprezentanta Sekcji SKAD PTS w IFCS Council na kolejną kadencję.

W ostatniej części zebrania prof. Marek Walesiak przypomniał, że w dniach 30 sierpnia – 2 września 2011 r. we Frankfurcie odbyła się konferencja IFCS oraz GfKl. Prof. Andrzej Sokołowski podsumował, że podczas konferencji wygłoszono 175 referatów, z tego znaczna część dotyczyła grupowania szeregów czasowych. Prof. Marek Walesiak podkreślił, że w konferencji udział wzięły osoby z następujących uczelni: UE Wrocław (4 osoby); UWM w Olsztynie (4 osoby); UE Kraków (3 osoby); UE Katowice (2 osoby); Politechnika Opolska (1 osoba); IBS PAN (1 osoba). Poinformował również, że w dniach 21-26 sierpnia w Dublinie odbył się 58th World Statistics Congress of the International Statistical Institute (ISI), w którym udział wziął prof. dr hab. Eugeniusz Gatnar.

Prof. Krzysztof Jajuga zapowiedział, że kolejna konferencja IFCS odbędzie się w Porto w lipcu 2013 r., a w 2015 r., prawdopodobnie w Kolumbii. Zachęcał do udziału w przyszłorocznej konferencji GfKl, która odbędzie się w Hildesheim. Prof. Andrzej Sokołowski podkreślił ważność publikowania artykułów w czasopismach z Listy Filadelfijskiej, w tym w *Journal of Classification*.

REDAKCJA PRZEGLĄDU STATYSTYCZNEGO składa podziękowania RECENZENTOM artykułów nadesłanych do publikacji w roku 2011 w osobach:

Prof. dr hab. Sławomir Dorosiewicz
Prof. dr hab. Eugeniusz Gatnar
Prof. dr hab. Stanisław Maciej Kot
Prof. dr hab. Tadeusz Kufel
Prof. dr hab. Władysław Milo
Prof. dr hab. Magdalena Osińska
Prof. dr hab. Emil Panek
Prof. dr hab. Antoni Smoluk
Prof. dr hab. Bogdan Suchecki
Prof. dr hab. Józef Stawicki
Prof. dr hab. Mirosław Szreder
Prof. dr hab. Łucja Tomaszewicz
Prof. dr hab. Marek Walesiak
Dr hab. Andrzej Bąk
Dr hab. Tadeusz Bołt
Dr hab. Marek Gruszczyński
Dr hab. Jerzy Marzec
Dr hab. Paweł Miłobędzki
Dr hab. Kazimierz Krauze
Dr hab. Mateusz Pipień
Dr hab. Krystyna Pruska
Dr hab. Krystyna Strzała
Dr Jerzy Korzeniewski
Dr Grzegorz Krzykowski
Dr Kamila Migdał Najman
Dr Krzysztof Najman
Dr Ewa Wycinka
Dr Anna Zamojska

SPIS TREŚCI

Od Redakcji ... 3
Andrzej S. B a r c z a k – Wspomnienie profesora Michała Kolupy ... 5
Emil P a n e k – O pewnej interpretacji systemu Walrasa z zapasami ... 9
Marek W a l e s i a k – Klasyfikacja spektralna a skale pomiaru zmiennych ... 13
Henryk G u r g u l, Tomasz W ó j t o w i c z – Hipoteza Brody w świetle wyników badań empirycznych ... 32
Marcin Ł u p i ń s k i – Short-term forecasting and composite indicators construction with help of dynamic factor models handling mixed frequencies data with ragged edges ... 48
Małgorzata M a r k o w s k a, Bartłomiej J e f m a ń s k i – Fuzzy classification of European regions in the evaluation of smart growth ... 74

SPRAWOZDANIA

Krzysztof J a j u g a, Marek W a l e s i a k, Feliks W y s o c k i – Sprawozdanie z Konferencji Naukowej nt. ”Klasyfikacja i analiza danych – teoria i zastosowania” ... 94

CONTENTS

From the Editor 3
Andrzej S. B a r c z a k – In Memory of Professor Michał Kolupa 5
Emil P a n e k – A Certain Interpretation of the Walras's System and Stocks 9
Marek W a l e s i a k – Spectral Clustering and Measurement Scales of Variables 13
Henryk G u r g u l, Tomasz W ó j t o w i c z – Brody's Hypothesis Versus Results of Empirical Research 32
Marcin Ł u p i ń s k i – Short-Term Forecasting and Composite Indicators Construction with Help of Dynamic Factor Models Handling Mixed Frequencies Data with Ragged Edges 48
Małgorzata M a r k o w s k a, Bartłomiej J e f m a ń s k i – Fuzzy Classification of European Regions in the Evaluation of Smart Growth 74

REPORTS

Krzysztof J a j u g a, Marek W a l e s i a k, Feliks W y s o c k i – "Classification and Data Analysis - Theory and Applications" – SKAD2011 Conference Report 94