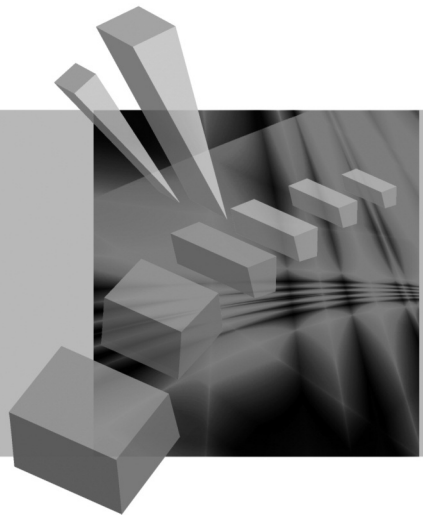
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Alina Bojan
Uniwersytet Ekonomiczny we Wrocławiu

WYKORZYSTANIE METOD WIELOWYMIAROWEJ ANALIZY DANYCH DO IDENTYFIKACJI ZMIENNYCH WPŁYWAJĄCYCH NA ATRAKCYJNOŚĆ WYBRANYCH INWESTYCJI

Streszczenie: 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 artykule dokonano próby identyfikacji zmiennych, które wpływają na to, iż dana grupa inwestycji traci na swej atrakcyjności w porównaniu z pozostałymi. 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.

Słowa kluczowe: wielowymiarowa analiza, metoda *k*-średnich, inwestycje.

1. Wstęp

Inwestor ma do wyboru wiele form na ulokowanie swojego kapitału. Mogą być to m.in. inwestycje w akcje, waluty, surowce, obligacje lub fundusze inwestycyjne. Wraz ze zmieniającą się sytuacją gospodarczą zmienia się często ich atrakcyjność, dlatego umiejętność reagowania na docierające sygnały i realokacja kapitału w odpowiednim momencie jest istotnym, aczkolwiek niezwykle trudnym zadaniem.

Niektóre czynniki mogą oddziaływać w podobny, negatywny bądź pozytywny sposób na wszystkie rozważane opcje inwestycyjne, tym samym ich wystąpienie niekoniecznie powinno skłaniać do podjęcia decyzji np. o wycofaniu kapitału. Analizując więc wpływ wybranych zmiennych na zachowanie potencjalnych inwestycji, warto to zrobić w odniesieniu do pozostałych dostępnych.

Celem artykułu jest zbadanie, jak w czasie zmienia się atrakcyjność wybranych grup inwestycji, jak również próba identyfikacji czynników, które w sposób istotny wpływają na te zmiany. Przy czym „atrakcyjność" będzie tu traktowana w kontekście ryzyka i zyskowności.

Ponieważ analizowane będzie zachowanie w czasie wielu obiektów opisanych pewnymi cechami, rozważane dane są szeregiem przekrojowo-czasowym. Niestety, w literaturze trudno odnaleźć przykłady podobnych danych oraz propozycje podejścia do ich badania. Aby ocenić zachowanie wybranej inwestycji na tle pozostałych potencjalnych, autorka proponuje zastosować metodę klasyfikacji danych, dzięki której dla każdego punktu w czasie otrzymuje się przyporządkowanie obiektów do odpowiedniego segmentu obrazującego poziom atrakcyjności inwestycyjnej. Następnie użycie metody regresyjnej powinno pozwolić na wyróżnienie najważniejszych cech, które wpływają na przemieszczanie się obiektów pomiędzy skupieniami. Do badania wykorzystano dane z portalu stooq.pl o dziennych cenach zamknięcia 32 wybranych form inwestycji oraz podstawowe dane makroekonomiczne pochodzące ze strony Głównego Urzędu Statystycznego w okresie od 1 stycznia 2001 r. do 30 czerwca 2011 r.

2. Klasyfikacja inwestycji w zależności od poziomu ich atrakcyjności

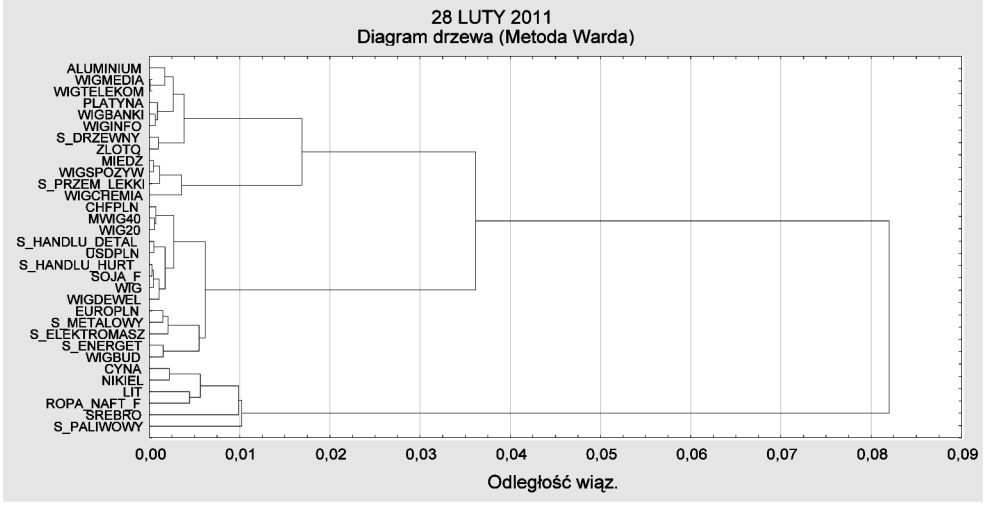
Aby zbadać, jak zmienia się „atrakcyjność" wybranych grup inwestycji w czasie, przeanalizowane zostały ceny 32 obiektów reprezentujących możliwe formy ulokowania kapitału w ciągu 10 lat (od 1 stycznia 2001 r. do 30 czerwca 2011 r.). Posłużyły do tego:

- indeksy GPW w Warszawie: WIG20 (jako przykład inwestycji 20 największych spółek akcyjnych notowanych na giełdzie), MWIG40 (reprezentuje lokowanie kapitału w średnie spółki), WIG-BANKI (banki), WIG-BUDOW (spółki budowlane), WIG-CHEMIA (spółki sektora chemicznego), WIG-DEWEL (spółki deweloperskie), WIG-INFO (spółki sektora informatycznego), WIG-MEDIA (spółki sektora medialnego), WIG-SPOZYW (spółki branży spożywczej), WIG-TELEKOM (spółki sektora telekomunikacyjnego) oraz WIG (jako ogólny przykład inwestycji na GPW),
- indeksy branżowe spółek z sektorów: drzewnego, elektromaszynowego, metalowego, energetycznego, paliwowego, przemysłu lekkiego, handlu detalicznego i hurtowego,
- kursy walut: EUROPLN, USDPLN, CHFPLN,
- metale: złoto, srebro, platyna, aluminium, cyna, lit, miedź, nikiel,
- surowce i towary: ropa naftowa, soja.

Wymienione obiekty opisane zostały dwiema charakterystykami, które są uznawane za podstawowe kryteria inwestycyjne, a mianowicie ryzykiem i oczekiwaną stopą dochodu. Podobnie jak w szeroko stosowanej teorii portfelowej Markowitza, za miarę ryzyka uznaje się tu odchylenie standardowe dziennych stóp zwrotu w miesiącu, natomiast za miarę zysku przyjmuje się średnią z dziennych stóp zwrotu w badanym miesiącu. W ten sposób powstał szereg 32 obiektów opisanych przez 2 zmienne w 126 kolejnych okresach, który stanowił zbiór do analizy. Sprowadzenie

analizy do okresu miesięcznego uzasadnione było tym, iż cechy, których wpływ będzie następnie badany, są publikowane właśnie z taką częstotliwością.

Dla każdego punktu w czasie dokonano procedury grupowania. Aby zdecydować, ile skupień należy wyróżnić w zbiorze, w pierwszej kolejności posłużono się metodą aglomeracyjną, której wynikiem jest dendogram (rys. 1) powstały przez rekurencyjne łączenie istniejących grup [Larose 2006]. Do jego wyznaczenia wykorzystano pakiet Statistica10. Stosując metodę Warda z odległością euklidesową, otrzymano w sumie 126 diagramów drzew. Ich analiza pozwoliła w zdecydowanej większości przypadków wysnuć wniosek, iż najbardziej odpowiednia liczba skupień dla analizowanych danych to 3. Drugą rozważaną opcją było wyróżnienie 4 grup.



Rys. 1. Przykładowe wyniki otrzymane po zastosowaniu metody Warda dla danych na 28 lutego 2011 r.

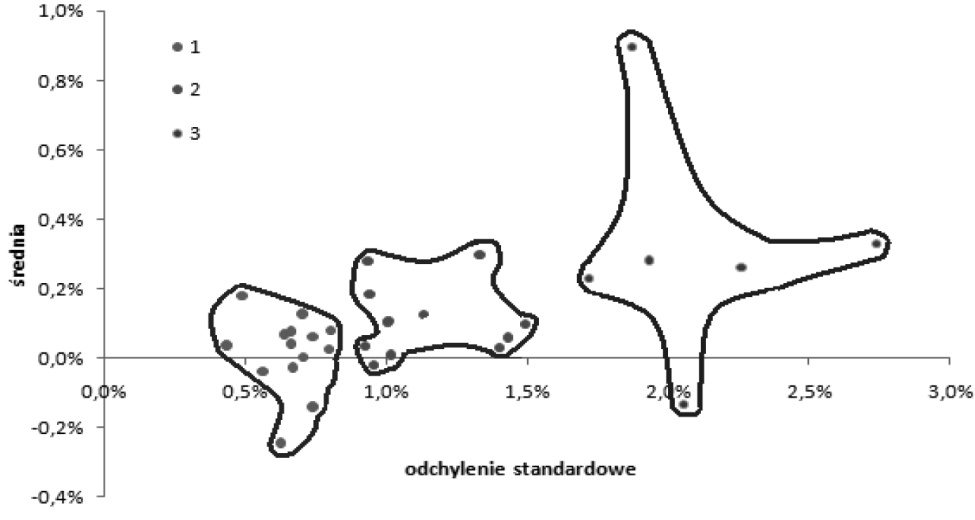
Źródło: opracowanie własne.

Następnie użyto algorytmu *k*-średnich będącego reprezentantem metod optymalizujących wstępny podział zbioru [Walesiak, Gatnar 2009], który dzieli obiekty między klasy, tak aby zmienność wewnątrz grupy była możliwie najmniejsza, a zmienność międzyklasowa wręcz przeciwnie. Procedurę tę przeprowadzono dla 3 i 4 klas. Do podjęcia decyzji o liczbie skupień posłużono się indeksem Calińskiego i Harabasza [Walesiak, Dudek 2006]:

$$G1(u) = \frac{tr(\mathbf{B}) / (u-1)}{tr(\mathbf{W}) / (n-u)},$$

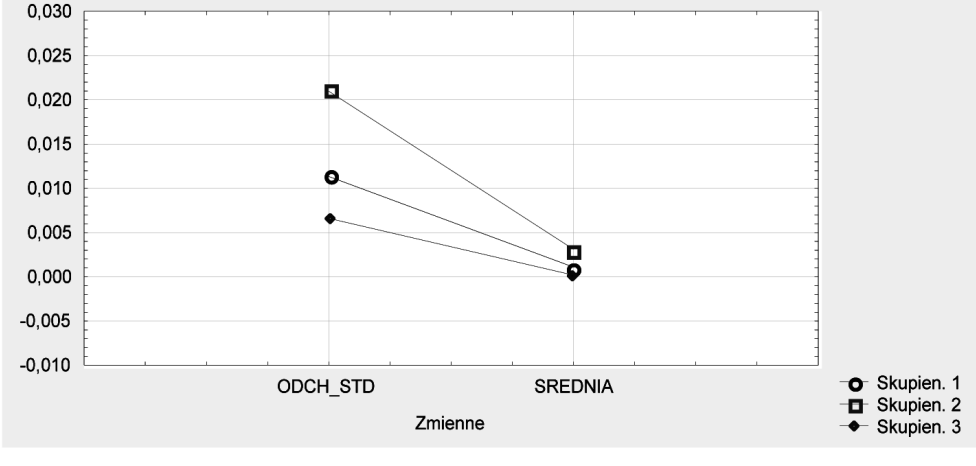
gdzie: *tr* – ślad macierzy, **W** – macierz kowariancji wewnątrzklasowej, ***B*** – macierz kowariancji międzyklasowej, *u* – liczba klas, *n* – liczba obiektów.

Dla większości miesięcy indeks ten przyjmował najwyższe wartości przy $u = 3$. Potwierdziło to wysnute wcześniej wnioski o słuszności wyróżnienia 3 skupień (rys. 2).



Rys. 2. Wynik analizy skupień metodą *k*-średnich dla 3 klas na 28 lutego 2011 r.

Źródło: opracowanie własne.



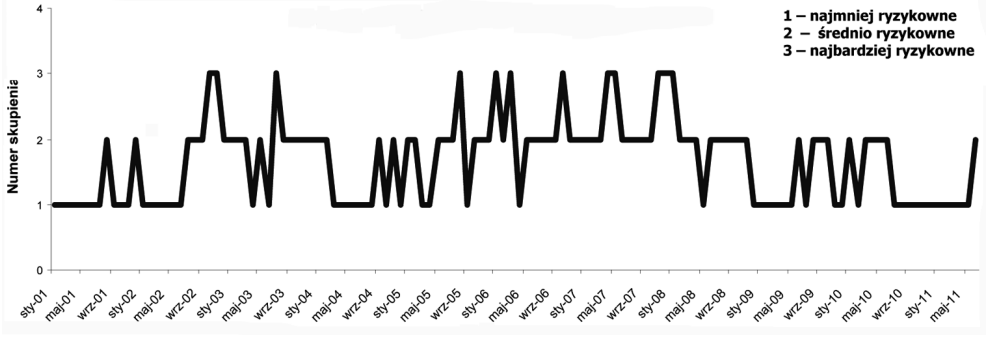
Rys. 3. Wykres średnich dla każdego skupienia na 28 lutego 2011 r.

Źródło: opracowanie własne.

Następnie dla otrzymanych segmentów przeprowadzono analizę wariancji, aby przekonać się, w jakim stopniu poszczególne cechy różnicują skupienia w czasie (rys. 3) oraz określić profil każdego z nich. Najbardziej istotnym czynnikiem charakteryzującym segmenty okazała się miara ryzyka, dlatego uzyskane grupy określono jako: inwestycje najmniej ryzykowne (1), inwestycje o średnim poziomie ryzyka (2) oraz te najbardziej ryzykowne (3).

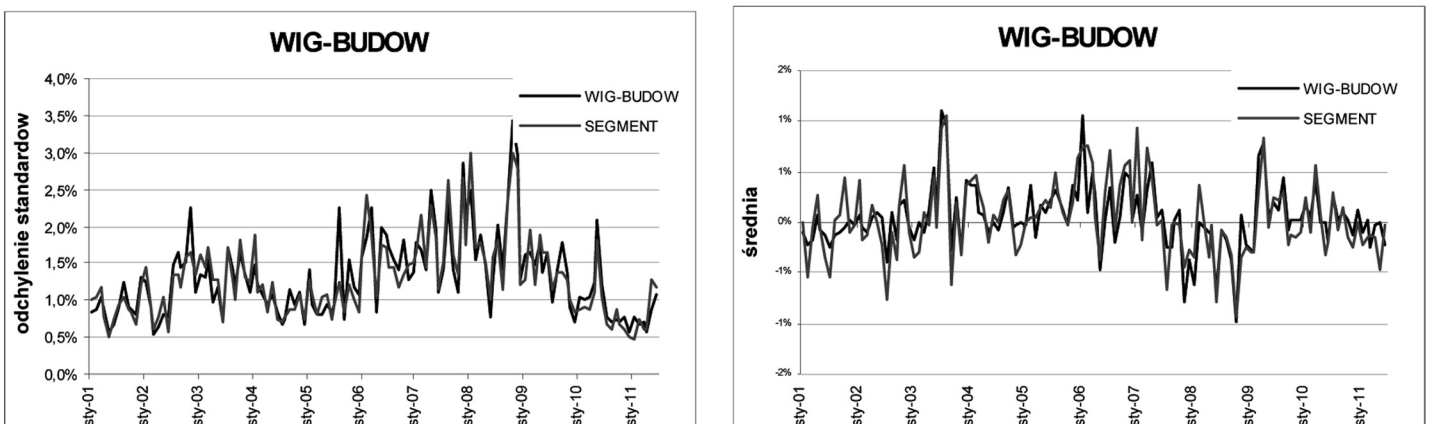
3. Analiza zmian skupień w czasie i identyfikacja wpływających na to zmiennych

Po przyporządkowaniu każdemu z obiektów odpowiadającego mu segmentu zbadana została ich zmienność w czasie (rys. 4). Najbardziej stabilne w analizowanym okresie na tle innych były inwestycje w soję oraz waluty. Najczęściej przynależały do segmentu 1, który charakteryzował się istotnie mniejszym poziomem ryzyka niż pozostałe. Tym samym czynniki, które wywoływały wzrost ich zmienności, musiały działać równocześnie na alternatywne opcje. Jednakże konsekwencją mniejszej wariancji stóp zwrotu jest zwykle ograniczona możliwość osiągania ponadprzeciętnych zysków. Również lokowanie w średnie spółki (reprezentowane przez indeks MWIG40) przez bardzo długi okres wiązało się ze stosunkowo niskim ryzykiem. W zupełnej opozycji znajdowała się inwestycja w ropę naftową, która klasyfikowana była najczęściej do skupień 2 oraz 3. Dużą zmiennością w przynależności do poszczególnych skupień cechował się sektor drzewny, metalowy, indeks WIG-TELEKOM reprezentujący inwestycje w spółki branży telekomunikacyjnej oraz platyna i srebro. Złoto okazało się najbardziej stabilną w czasie opcją inwestycyjną wśród badanych metali szlachetnych ze stosunkowo niższym poziomem zmienności. Indeks WIG-BANKI przez długi okres należał do segmentu 1 i 2, jednak w miesiącach przypadających na czas kryzysu znacznie wzrosła częstość przyporządkowania go do skupień 2 i 3.



Rys. 4. Wykres zmiany skupień w czasie dla WIG-BUDOW

Źródło: opracowanie własne.



Rys. 5. Porównanie w czasie faktycznych miar ryzyka i zysku WIG-BUDOW z wartościami właściwymi dla skupienia, do którego w danym czasie obiekt został przyporządkowany

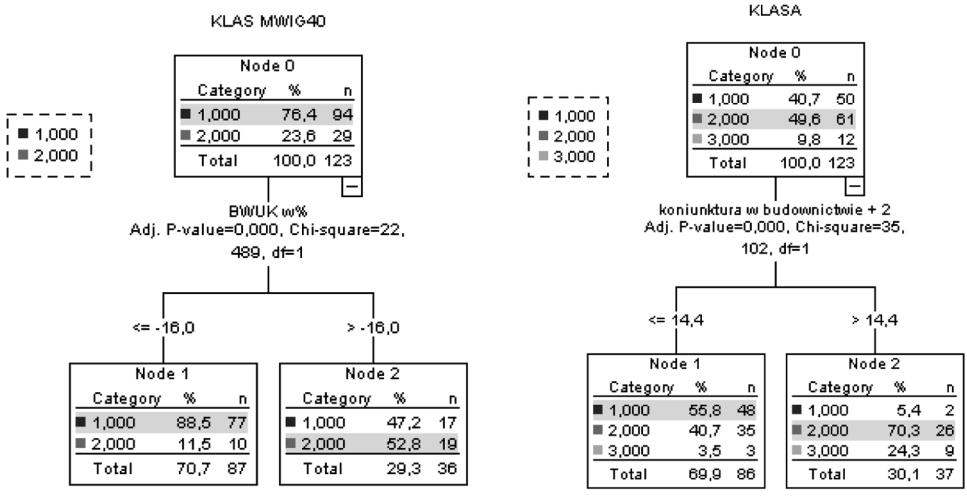
Źródło: opracowanie własne.

Cała analiza ze względu na ilości danych jest dość rozległa i z powodu ograniczeń, jakie nakłada artykuł, szerzej przedstawione zostaną dwie grupy inwestycji: WIG-BUDOW (rys. 5) oraz MWIG40.

W wyniku grupowania każdemu obiektowi nadano numer segmentu na koniec miesiąca i sprawdzono, które ze zmiennych ekonomicznych mają największy wpływ na to, jakim profilem inwestycyjnym charakteryzuje się dana inwestycja.

Jako zmienne objaśniające wykorzystano tu dane pochodzące ze strony Głównego Urzędu Statystycznego, które prezentowane są w odstępach miesięcznych, a mianowicie: wskaźnik inflacji, wartość eksportu towarów, wartość importu towarów, bieżący wskaźnik ufności konsumenckiej (BWUK), wyprzedzający wskaźnik ufności konsumenckiej (WWUK), wskaźnik ogólnego klimatu koniunktury w przetwórstwie przemysłowym, wskaźnik ogólnego klimatu koniunktury w budownictwie, liczba mieszkań oddanych do użytkowania, ceny żywność i napojów bezalkoholowych, ceny napojów alkoholowych i wyrobów tytoniowych, ceny odzieży i obuwia, ceny użytkowania mieszkania i nośników energii, ceny transportu, ceny łączności, ceny rekreacji i kultury oraz stopę bezrobocia. Powyższe zmienne włączono do analizy również dla przesunięcia w czasie do 3 miesięcy, co miało odzwierciedlić opóźnienia w udostępnieniu danych do wiadomości publicznej. W pierwszej kolejności dokonano analizy korelacji. Ponieważ, jak wykazała analiza wariancji, odchylenie standardowe zawsze różnicowało otrzymane klasy najbardziej, można je było uporządkować rosnąco ze względu na miarę ryzyka. A skoro zmienna objaśniana może być traktowana jako zmienna porządkowa, to do wstępnej analizy wpływu czynników można się było posłużyć współczynnikiem korelacji rang Spearmana.

W przypadku indeksu WIG-BUDOW współczynnik ten przyjmował największe wartości dla wskaźnika ogólnego klimatu koniunktury w budownictwie z przesunięciem czasowym o 3 okresy ($\rho_s = 0{,}48$). Nieznacznie mniejszy współczynnik korelacji miały pozostałe zmienne odnoszące się do koniunktury w budownictwie, jak również wskaźnik ogólnego klimatu koniunktury w przetwórstwie przemysłowym (0,40) i BWUK (0,38). Wysoką, lecz ujemną korelacją wynoszącą –0,40 cechowały się natomiast ceny wyposażenia mieszkania i prowadzenia gospodarstwa domowego. Następnie dla rozważanego indeksu dokonano próby budowy drzewa decyzyjnego. Jest to metoda nieparametryczna, polegająca na stopniowym podziale wielowymiarowej przestrzeni cech na rozłączne podzbiory, aż do uzyskania ich pełnej homogeniczności ze względu na wyróżnioną cechę [Gatnar 2000]. Użyto algorytmu CHAID, a jako warunkiem stopu posłużono się regułą, iż minimalna liczba obserwacji w węźle potomnym (*child node*) powinna być równa co najmniej 30. Metoda ta wskazała, iż najbardziej istotny okazał się wskaźnik ogólnego klimatu koniunktury w budownictwie z przesunięciem czasowym o 2 miesiące. Użycie wykrytego warunku pozwoliło poprawnie zakwalifikować 60% przypadków (rys. 6). Przesunięcie o 2 okresy może wynikać z opóźnienia w publikacji danych przez GUS. Pozwala to wysnuć wniosek, iż ta informacja w sposób istotny wpływa na decyzje inwestycyjne. Analizując wpływ wybranych czynników na zmiany skupień MWIG40 reprezen-



Rys. 6. Drzewo regresyjne dla WIG-BUDOW oraz MWIG40 z wykorzystaniem algorytmu CHAID

Źródło: opracowanie własne.

tującej inwestycje w średnie spółki, największą dodatnią korelacją można zauważyć dla zmiennych: bieżący wskaźnik ufności konsumenckiej (BWUK), wskaźnik ogólnego klimatu koniunktury w przetwórstwie przemysłowym oraz wskaźnik ogólnego klimatu koniunktury w budownictwie (rys. 6). Wysoka ujemna korelacja jest natomiast widoczna dla stopy bezrobocia. Po zbudowaniu drzewa regresyjnego z użyciem algorytmu CHAID okazało się, iż największy wpływ na profil atrakcyjności inwestycyjnej ma BWUK, a użycie wskazanej reguły decyzyjnej pozwoliło poprawnie zakwalifikować 78% przypadków.

4. Podsumowanie

W większości prac dotyczących inwestycji analizowane są bądź to szeregi czasowe dla konkretnej spółki lub też szeregi przekrojowe dotyczące wielu obiektów w jednym punkcie czasu. Ponieważ autorka była zainteresowana znalezieniem cech, które w analizowanych 10 latach w sposób istotny oddziaływały na wybrane inwestycje, biorąc pod uwagę zachowanie alternatywnych form, należało podejść do problemu szeregu przekrojowo-czasowego. Zaproponowane podejście pozwoliło ocenić, w jaki sposób zmieniał się profil atrakcyjności wybranych inwestycji mierzony za pomocą ryzyka i zysku. Ponieważ w badanym przypadku otrzymane segmenty można było uznać za zmienną porządkową, do wysnucia pierwszych wniosków można było się posłużyć prostą analizą korelacji rang. Kolejnym krokiem w analizie była budowa drzewa decyzyjnego. W przypadku inwestycji w spółki budowlane okazało się, iż czynnikiem specyficznym wpływającym na zmianę klasy jest wskaźnik

ogólnego klimatu koniunktury w budownictwie. Jak wykazała analiza, inwestorzy przede wszystkim kierują się jego znaną na dany moment w czasie wartością. Dla spółek średnich najbardziej istotny wpływ miał natomiast bieżący wskaźnik ufności konsumenckiej. Niestety 10 lat okazuje się zbyt krótkim szeregiem do wyciągnięcia głębszych wniosków, gdy zmienna objaśniana jest mierzona na skali nominalnej lub porządkowej. Analizę ograniczają bowiem restrykcje odnośnie do minimalnej liczebności dla każdej z reprezentowanych wartości zmiennej endogenicznej, a wykrycie cech o mocy dyskryminacyjnej wymaga większych zbiorów. Mimo to autor uważa, iż samo podejście może zwiększyć wiedzę badacza, a w przypadku posiadania informacji o zachowaniu się obiektów w dłuższym okresie lub też mierzenia wartości potencjalnych czynników z większą częstotliwością niż miesięczna otrzymana analiza przyniosłaby dodatkowe wnioski.

Literatura

Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa 2000.

Larose E., *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, PWN, Warszawa 2006.

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

Walesiak M., Dudek A., *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – charakterystyka problemu*, Zeszyty Naukowe Uniwersytetu Szczecińskiego, Prace Katedry Ekonometrii i Statystyki nr 17, Szczecin 2006.

IDENTIFICATION OF VARIABLES WHICH INFLUENCE ATRACTIVENESS OF GIVEN INVESTMENTS WITH THE USAGE OF MULTIVARIATE ANALYSIS

Summary: An investor has many possibilities to alocate his capital. He can invest in shares of selected branches, commodities, currency or bonds. Their attractiveness, however, differs in time together with the changeable economic situation. Thus, it is very important to know how to react correctly to the reaching signals, in order to realocate the capital in the right moment. The aim of this article is to identify the most significant variables which have influence on making a given group of investment less attractive in comparison to other ones available on the market. Classification methods allowed to distinguish the most and the least atractive groups of ivestments from risk and rate of return point of view. Next, variables which influence changes of segments in time were identified.

Keywords: multivariate analysis, k-means, investments.