

Robert Pietrzykowski, Paweł Kobus
Katedra Ekonometrii i Informatyki,SGGW
e-mail: rpietrzykowski@mors.sggw.waw.pl

ZASTOSOWANIE MODYFIKACJI METODY K-ŚREDNICH W ANALIZIE PORTFELOWEJ.

Streszczenie: W pracy przedstawiono modyfikację metody k - średnich oraz jej zastosowanie do analizy spółek giełdowych notowanych na Warszawskiej Giełdzie Papierów Wartościowych w roku 2004. W podziale na grupy uwzględniono 206 spółek, które grupowano ze względu na najczęściej wykorzystywane wskaźniki finansowo - ekonomiczne.

Słowa kluczowe: analiza portfelowa, analiza skupień, metoda k – średnich,

WSTĘP

Prognozowanie w analizie finansowej spółek odbywa się zwykle na podstawie danych historycznych. Zwykle wykorzystuje się ceny poszczególnych akcji aby wnioskować o ich zakupie bądź sprzedaży. Jednak oprócz tych danych wykorzystuje się również inne wskaźniki, które określają kondycję finansową spółki. W podejmowaniu trudnych dla inwestora decyzji wykorzystuje się różne metody statystyczne. Ze względu na złożoność problemu i branie pod uwagę wielu wskaźników, rozważa się wielowymiarowe metody statystyczne i tak do lat osiemdziesiątych najbardziej popularną metodą była analiza dyskryminacyjna. W latach dziewięćdziesiątych zaczęto stosować metody regresji logistycznej oraz sieci neuronowych. Natomiast w ostatnich latach wykorzystuje się metody taksonomiczne takie jak: analiza skupień, analizy k-średnich i inne.

W pracy zaprezentowano modyfikację metody k-średnich, którą zastosowano do podziału spółek giełdowych ze względu na najczęściej wykorzystywane wskaźniki w analizie fundamentalnej.

METODA K-ŚREDNICH.

Metoda k – średnich należy do metod podziałowych analizy skupień. Metody podziałowe polegają na dzieleniu całego zbioru obiektów zgodnie z ogólną zasadą maksymalizacji wariancji pomiędzy poszczególnymi grupami, przy jednoczesnej minimalizacji wariancji wewnątrz badanych grup. Idea metody k-średnich została opracowana w latach pięćdziesiątych przez T. Daleniusa, który przedstawił iteracyjną procedurę podziału populacji na k grup, tak by zminimalizować wielkość wewnątrzgrupowej wariancji. D.R.Cox [Cox, 1957]

w swojej pracy podał funkcję mierzącą wielkość strat związanych z podziałem obiektów na k grup według jednowymiarowej zmiennej o rozkładzie normalnym. Uogólnienie dla przypadku wielowymiarowego przedstawił G. S. Sebestyen [Sebestyen, 1962]. Autorstwo metody k -średnich, przypisuje się jednak J. McQueen'owi [McQueen, 1967], który rozpatrywał efektywność tejże metody z punktu widzenia losowego doboru obiektów do wyróżnionych grup. [Grabiński, 1992].

Metoda k -średnich należy do metod optymalizacyjno-iteracyjnych. Istota tej grupy metod polega na tym, iż optymalizowana jest pewna funkcja jakości podziału obiektów. Funkcję kryterium można zapisać w postaci formuły minimalizującej ślad macierzy wariancji wewnątrzgrupowej lub maksymalizującej ślad macierzy wariancji międzygrupowej [Gatnar, 1998]. Metodę k -średnich wykorzystuje się do analizy dużych ilości danych, a jej istota polega na zredukowaniu dużej ilości nagromadzonych informacji do kilku podstawowych kategorii, co pozwala na łatwe zorientowanie się w danym zjawisku, wyciągnięcie wniosków uogólniających. Zastosowanie metody k -średnich daje możliwość ustalenia typologii w zakresie badanych obiektów oraz określenie jednorodnych przedmiotów analizy, w której łatwiej jest wyodrębnić czynniki systematyczne oraz ewentualne związki przyczynowo-skutkowe. Jej zastosowanie może prowadzić do zmniejszenia nakładów czasu i kosztów badań przez ograniczenie rozważań do najbardziej typowych faktów, zjawisk czy obiektów przy stosunkowo niewielkich stratach informacji. Działanie metody k -średnich można zawrzeć w następujących punktach. Punktem wyjścia jest wstępny podział zbioru na k skupień, arbitralnie wrzucając obiekty do tych grup. Poszukuje się takiego przypisania obiektów do grup, by w ich obrębie osiągnąć maksymalne podobieństwo przy zachowaniu maksymalnych różnic międzygrupowych. Algorytm stosuje się do momentu otrzymania takiego podziału jednostek, aby uzyskać jak najbardziej istotne wyniki analizy wariancji. Problem z jakim spotyka się badacz w analizie k -średnich to ustalenie wstępnego podziału na liczbę skupień. Podziału tego można dokonać w sposób losowy lub opierając się na ocenie ekspertów, która wynika z intuicji lub znajomości przedmiotu badań. Można również wykorzystać inne metody taksonomiczne.

Ogólna idea tych procedur polega na poprawianiu danego podziału obiektów z punktu widzenia odpowiednio zdefiniowanego kryterium optymalności podziału. Zakładamy, iż $k \in (2, n-1)$, gdzie n jest liczbą obiektów.

Wariant metody k -średnich można opisać następująco. Niech $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \dots, \mathbf{X}_n$ będą obiektami p cechowymi. (to znaczy $\mathbf{X}_1 = [x_{11}, \dots, x_{1m}]$). Na początku ustala się wyjściową macierz środków ciężkości grup

$$B = [\bar{x}_{lj}] \quad (l = 1, \dots, p; j = 1, \dots, m) \quad (1)$$

gdzie m – liczba zmiennych

Dla każdej z grup obliczamy średnią (położenie centroidu). Wyznacza się odległości pierwszej nieprzydzielonej jednostki od środków ciężkości

poszczególnych grup i kwalifikuje ją do grupy najbliższej położonej. Następnie wyznacza się wartość wyjściowego błędu podziału obiektów między k grup

$$e = \sum_{i=1}^n d_{il}^2 \quad (2)$$

gdzie: d_{il}^2 – odległość Euklidesa między i-tym obiektem a najbliższym l-tym środkiem ciężkości:

$$d_{il}^2 = \sum_{j=1}^m (x_{ij} - \bar{x}_{lj})^2 \quad (i=1, \dots, n) \quad (3)$$

Zestaw odległości euklidesowych obliczany jest pomiędzy poszczególnymi elementami zbioru a kolejnymi centroidami. Dla pierwszego obiektu określa się zmiany błędu podziału wynikające z przyporządkowania go kolejno do wszystkich aktualnie występujących grup:

$$\Delta e_l^{(1)} = \frac{n_k d_{1k}^2}{n_k + 1} - \frac{n_{k_1} d_{1k_1}^2}{n_{k_1} - 1} \quad (4)$$

gdzie: n_k – liczebność k - tej grupy, d_{1k} – odległość pierwszego obiektu od środka ciężkości k - tej grupy, n_{k_1} - liczebność grupy zawierającej pierwszy obiekt, d_{1k_1} – odległość pierwszego obiektu od najbliższego środka ciężkości.

Jeżeli minimalna wartość wyrażenia $\Delta e_l^{(1)}$ dla wszystkich $l \neq l_1$ jest ujemna, to pierwszy obiekt przypisuje się do grupy, dla której $\Delta e_l^{(1)} = \min$. Następnie powtarza się obliczenia to znaczy od nowa oblicza się środki ciężkości grup **B** uwzględniając dokonaną transformację obiektu oraz wyznacza aktualną wartość błędu podziału. Jeżeli minimalna wartość wyżej przedstawionego wyrażenia jest dodatnia lub równa zero, to nie dokonujemy już żadnych zmian. Operacje opisane powyżej powtarza się dla każdego następnego obiektu. Gdy nie obserwujemy już żadnych przesunięć obiektów z grupy do grupy, czyli gdy każdy element jest w grupie, w której centroid jest mu najbliższy, wówczas postępowanie się kończy w pierwszej wersji podziału. W przeciwnym wypadku rozpoczyna się następną iterację, aż do momentu, w którym ich liczba nie przekroczy zadanej wartości [Zeliaś i in., 1989, Witkowska, 2002]. Istnieje wiele modyfikacji tej metody niektóre z nich można znaleźć u Grabińskiego, Zeliasia i innych. [Grabiński, 1992, Zeliaś i in., 1989]

MODYFIKACJA METODY K - ŚREDNICH.

Obserwowane obiekty $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ są obiektami p cechowymi, to znaczy $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$, gdzie $i = 1, \dots, N$. Ponieważ liczba k skupień jest z góry ustalona, szukamy najlepszego podziału $\mathbf{J}(k) = \{G_1, G_2, \dots, G_k\}$, którym będzie podział zbioru $\{1, \dots, N\}$ na k rozłącznych podzbiorów. Wybieramy najlepszy spośród wszystkich uzyskanych podziałów, to znaczy taki, dla którego zróżnicowanie

wewnątrzgrupowe było najmniejsze oraz zmienność pomiędzy grupami była jak największa czyli oznaczając taki podział przez $\mathbf{J}^*(k)$ (to taki podział na k grup, że zróżnicowanie międzygrupowe w stosunku do zróżnicowania wewnątrzgrupowego jest największe.)

Jako miernik zróżnicowania międzygrupowego przyjęto:

$$S_{A(J(k))}^2 = \frac{1}{k-1} \sum_{i=1}^k \left\| \bar{x}_i - \bar{x}_{J(k)} \right\|^2 = \frac{1}{k-1} \sum_{i=1}^k d_i^2 \quad (5)$$

gdzie:

$$\bar{x}_{J(k)} = \frac{1}{k} \sum_{i=1}^k \bar{x}_{G_i} \quad - \text{środek ciężkości proponowanego podziału } J(k)$$

$$\bar{x}_{G_i} \quad - \text{środek ciężkości } i\text{-tej grupy.}$$

$$\|X\| \quad - \text{oznacza normę euklidesową wektora } \mathbf{X}_1 = (x_{i1}, \dots, x_{ip}) \text{ to}$$

$$\|X\|^2 = \sum_{i=1}^p X_i^2$$

Jako miernik zróżnicowania wewnątrzgrupowego zaproponowano:

$$S_{E(J(k))}^2 = \frac{1}{N-k-1} \sum_{i=1}^k \left(\sum_{j=1}^{k_i} \|x_{ij} - \bar{x}_{G_i}\|^2 \right) = \frac{1}{N-k-1} \sum_{i=1}^k \left(\sum_{j=1}^{k_i} d_{ij}^2 \right) \quad (6)$$

Wtedy $\mathbf{J}^*(k)$ będzie takim podziałem na k grup, że:

$$\frac{S_{A(J^*(k))}^2}{S_{E(J^*(k))}^2} = \max_{J(k)} \frac{S_{A(J(k))}^2}{S_{E(J(k))}^2} \quad (7)$$

oraz niech

$$f(k) = \frac{S_{A(J^*(k))}^2}{S_{E(J^*(k))}^2} \quad (8)$$

Funkcja kryterium to ogólna suma odległości wewnątrzgrupowych liczonych od środka grup, których współrzędne wyznaczono jako średnie arytmetyczne wartości cech obiektów należących do danej podgrupy. Jako optymalny podział $\mathbf{J}^*(k)$ obiektów na skupienia wybiera się ten, dla którego funkcja określona wzorami 7, 8 osiąga maksimum [Pietrzykowski i in., 2005].

DANE DOŚWIADCZALNE.

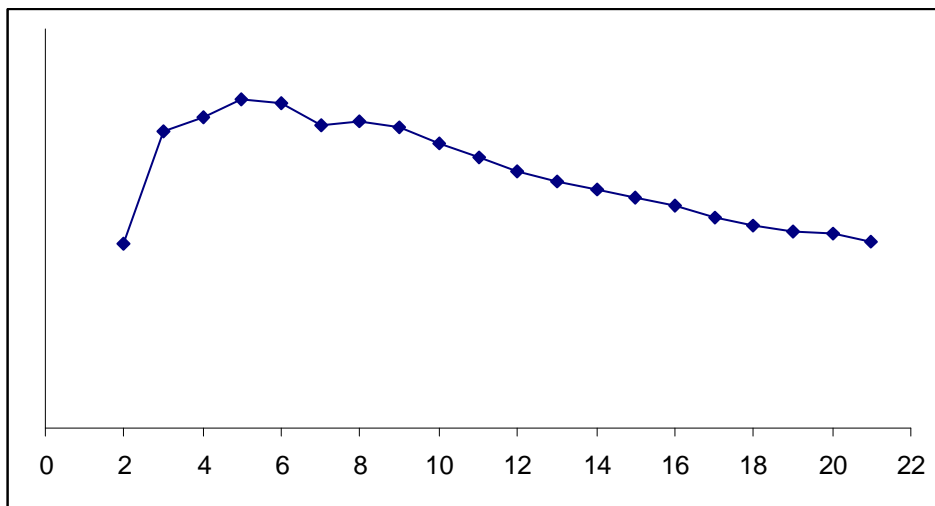
W analizie wykorzystano spółki notowane na Giełdzie Papierów Wartościowych od stycznia do grudnia 2004. Pominięto spółki, które w badanym okresie zostały wycofane z notowań, oraz takie, dla których

dane ekonomiczno – finansowe były niekompletne. W rezultacie analizę przeprowadzono na 206 spółkach giełdowych.

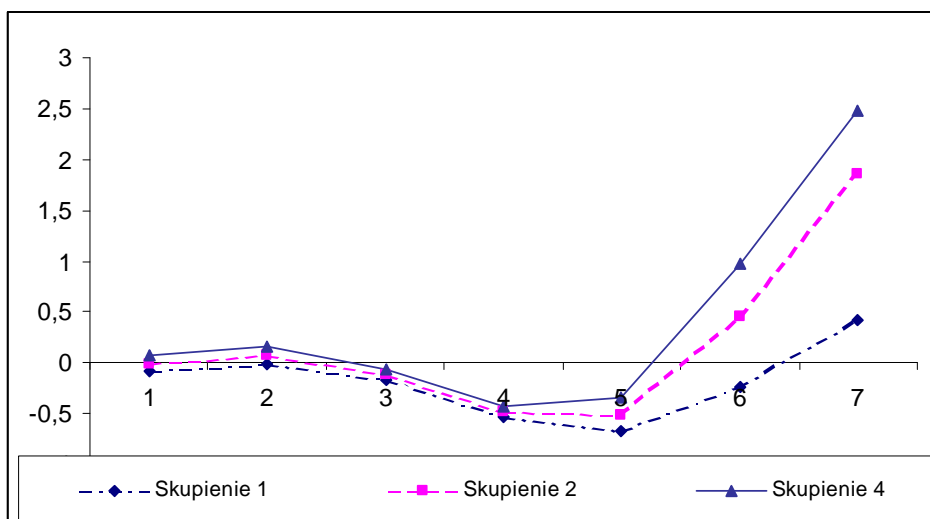
Jako zestaw zmiennych charakteryzujących decyzje inwestycyjne oraz ryzyko ich podejmowania wykorzystuje się pięć grup wskaźników opisujących kondycję ekonomiczno – finansową firmy. Jako zmienne grupujące do analizy wybrano siedem wskaźników ekonomiczno-finansowych. Powodem wybrania takiej kombinacji jest ich duże zróżnicowanie jak i stosunkowo dokładne odzwierciedlenie kondycji finansowej firmy. Wybrane wskaźniki finansowe to miary klasyczne, powszechnie stosowane do określenia kondycji finansowej firmy. Dzięki nim można zidentyfikować mocne i słabe strony działalności gospodarczej firmy. Są one również źródłem informacji o zagrożeniach i szansach rozwoju strategicznego firmy. Największą wadą jest fakt, iż mają znaczenie historyczne. Z reguły wyznacza się je na koniec roku obrachunkowego lub kwartalnie, a w przyszłości ich znaczenie z miesiąca na miesiąc spada. W dalszej analizie wykorzystano następujące wskaźniki ekonomiczno – finansowe (w nawiasie podano przyjęte oznaczenia wskaźników): ROI (W1); ROE (W2); ROA (W3); wskaźnik rentowności sprzedaży (W4); wskaźnik zyskowności netto (W5); wskaźnik kapitałowy (W6); obrotowość (produktywność) aktywów ogółem (W7). Wskaźniki W1 i W4 należą do grupy wskaźników zwrotu z inwestycji. Wskaźniki W2, W3 i W5 zalicza się do grupy wskaźników zyskowności. Wskaźnik W6 to wskaźnik zaliczany do grupy wskaźników kapitałowych, a wskaźnik oznaczony jako W7 należy do grupy wskaźników aktywności gospodarczej.

WYNIKI

Na wykresie 1 przedstawiono wartości funkcji $f(k)$, gdzie na osi poziomej znajdują się liczby skupień. Jak widać na wykresie funkcja osiągnęła maksimum dla pięciu skupień. Z badań przeprowadzonych dla polskiego rynku kapitałowego wynika, że dobrze zdywersyfikowany portfel otrzymuje się dla akcji z przedziału od pięciu do piętnastu spółek dlatego maksimum funkcji szukano w takim zakresie.



Wykres 1. Wartości funkcji f(k)



Wykres 2. Średnie dla trzech skupień.

Powstaje problem jak wybrać spółki z poszczególnych skupień do naszego portfela akcji. Jak wspomniano wcześniej pomijamy te skupienia, w których jest jedna bądź dwie spółki. Pozostaje więc wybór spółek z trzech skupień. Problemem jest jednak, które spółki wybrać. Na wykresie 2 przedstawiono, wpływ poszczególnych wskaźników na przydzielenie spółek do odpowiednich skupień. Jak można zauważyć wskaźnik który najbardziej różnicował otrzymane podziały to obrotowość (produktywność) aktywów ogółem (W7), chociaż można powiedzieć, że podobnie działał wskaźnik kapitałowy (W6).

W tabeli 1 przedstawiono spółki, które zostały przydzielone do poszczególnych skupień. Najmniej liczne skupienia to skupienie trzecie i piąte. W skupieniu trzecim znalazły się dwie spółki, a w skupieniu piątym tylko jedna. Spółki, które znalazły się w tych skupieniach są spółkami albo bardzo dobrymi lub bardzo słabymi ze względu na badane wskaźniki ekonomiczne. W skupieniu pierwszym, drugim i czwartym znalazły się pozostałe spółki. W poszczególnych skupieniach znalazły się spółki, które były podobne ze względu na wskaźniki finansowo-ekonomiczne, a nie koniecznie podobne ze względu na prowadzona działalność gospodarczą lub oddziaływanie na rynku gospodarczym. Z tych skupień najbardziej liczne jest skupienie czwarte.

Tabela 1. Spółki przydzielone do poszczególnych skupień przy podziale $k = 5$.

Numer skupienia	Nazwa Spółki
Skupienie pierwsze	PRÓCHNIK; AMREST; BICK; BYTOM; ELEKTROEX; IBSYSTEM; MOSTALZAB; PEMUG; SWARZĘDZ; SWISSMED; WFM OBORNIKI;
Skupienie drugie	ATLANTIS; BUDOPOL; TELL; TIM; PGF; AMPLI; KAREN NOTEBOOK; NEONET; JAGO; INDYKPOL; ZREW; ORFE; ODLEWNIE; PROVIMI-ROLIMPEX; FARMACOL; ALMA MARKET; STALPRODUKT; CSS; INTERIA; SOKOŁÓW; ŚRUBEX; MCI; POLMOS BIA.; MEDIATEL; TORFARM; ELDORADO; EUROCASH; ROPCZYCE; WAWEL; PROSPER; IGROUP; OPTIMUS; MACROSOFT; DECORA; LENA; STALPROFIL; CAPITAL P.; PROJPRZEM;
Skupienie trzecie	ŁDA INVEST; FON
Skupienie czwarte	EKODROB; TUP; AL PRAS; PPWK; RESBUD; IDMSAPL; FAMEG; EUROFAKTOR; TVN; PEP; POLNORD; PROCHEM; NKT CABLES; STRZELEC; STALEXPORT; ELMONTWAR; ECHO; PKN ORLEN; POLNA; GANT; ZETKAMA; TRAVELPLANET; RMF FM; HOOP; MOSTALWAR; PROKOM; TRAS-INTUR; BUDIMEX; KOPEX; BEEF SAN; EFEKT; BARLINEK; NAFTOBUDOWA; REDAN; VARIANT; MIESZKO; SOFTBANK; TP SA; BAUMA; 7BULLS.COM; POLIGRAFIA; NOVITA; KOGENERACJA; TECHMEX; PBG; GRAAL; FERRUM; BETACOM (); INTERCARS; RELPOL; KRUK; CERSANIT; EMAX; ELBUDOWA; COMARCH; PRATERM; MUZA; AMICA; KOELNER; WÓLCZANKA; FASING; CENSTALGD; KROSNO; BĘDZIN; RAFAKO; KOZIENICE; ŻYWIEC; MEWA; LPP; KGHM; WANDALEX; KRUSZWICA; HYGIENIKA; ZTS_ERG; CIECH; WISTIL; COMPUTERLAND; AQUA; IRENA; PEPEES; KOMPAP; MNI; IMPEXMETAL; PAGED; WSiP; BIOTON; POLLENA; ARTMAN; MASTERS; SIMPLE; GROCLIN; ELSTAR OILS; PLASTBOX; ELZAB; POLMOS LUB.; APATOR; FORTE; HOGA; BEST; BORYSZEW; HUTMEN; CCC; ORBIS; ODRATRANS; BOLESŁAWIEC; IMPEL; GRUPA ONET; LUBAWA; DWORY; JUTRZENKA; REMAK; AGORA; SANOK; GRAJEWÓ; KĘTY; ŚNIEŻKA; SPIN; DUDA; ZEG; ATM GRUPA; FAM; WILBO; VISTULA; MILMET; DĘBICA; ATM; SANWIL; COMP; OZC; SKOTAN; PONARFEH; JELFA; LZPS PROTEKTOR; MPEC; ZELMER; SUWARY; POLICE; LOTOS; JCAUTO; MONDI; LENTEX; TALEX; NETIA; PUŁAWY; DGA; HYDROTOR; POLCOLORIT; PERMEDIA; ABG STER-PROJEKT; OPOCZNO; MENNICA; PEKAES; NOWAGALA; UNIMIL; .
Skupienie piąte	ELEKTRIM

Do portfela przyjęto te spółki, które najmniej różniły się od średnich odległości określających poszczególne wskaźniki ekonomiczno - finansowe ze szczególnym uwzględnieniem wskaźnika obrotowość (produktywność) aktywów ogółem. W efekcie uzyskano portfel składający się z pięciu spółek: Próchnik, Amrest, PGF, Mondy, Budimex.

PODSUMOWANIE.

Metoda k-średnich jest pomocna w tworzeniu portfela akcji. Ze względu na pominięcie spółek ze skupień jedno i dwuelementowych, powstaje jednak problem wyboru akcji tak by zdywersyfikować portfel. Należy więc stwierdzić, że metoda k-średnich nie jest metodą która określa, które spółki powinny znaleźć się w portfelu. Można powiedzieć, że jest ona prostym narzędziem dzięki któremu można przeprowadzać szybkie analizy wielowymiarowe. Zastosowanie modyfikacji tej metody pozwoliło określić na ile skupień podzielić badane obiekty. W standardowej metodzie k-średnich liczbę skupień przyjmujemy arbitralnie, a tym samym nie jesteśmy do końca pewni czy nasz podział można uznać za „najlepszy”. Dzięki tej metodzie udało się dobrać spółki, z różnym gałęzi gospodarki w oparciu o wskaźniki finansowo-ekonomiczne czyli oprócz dywersyfikacji poziomej uwzględniono również dywersyfikację pionową. Wybrane spółki wydają się być podobne ze względu na badane wskaźniki co można wpływać na dywersyfikację ryzyka portfela.

LITERATURA

- Cox D.R. "Note of grouping", Journal of The American Statistical Association, 1957.
 McQueen J. "Some methods for classification and analysis of multivariate observations", 5'th Berkaley Symposium on Mathematics, "Statistics and Probability", 1967.
 Gatnar E. „Symboliczne metody klasyfikacji danych” PWN, Warszawa, 1998
 Grabiński T. „Metody taksonometrii” wyd. Akademii Ekonomicznej w Krakowie, 1992
 Pietrzykowski R., Zieliński W., Koziół D. „Wykorzystanie metody k-średnich w taksonomii portfela akcji” wyd. Wyższej Szkoły Ekonomiczno – Informatycznej, Warszawa, 2005, s. 3, 74-76
 Sebestyen G.S. "Decision making process in pattern recognition", John Wiley and Sons, New York, 1962.
 Witkowska D. Sztuczne sieci neuronowe i metody statystyczne. C. H. Beck, Warszawa 2002
 Zeliaś A, Grabiński T, Wydmus S., Metody taksonomii numerycznej w modelowaniu zjawisk społeczno – gospodarczych PWN Warszawa, 1989

Application of modify k-means method to portofolio analysis.

Summary: In the paper propose some modification method of k-means to portofolio analysis. As an example, consider partitions into k clusters of 206 stocks for Warsaw Stock Exchange in 2004 year.

Key words: portofolio analysis, method k-means, cluster analysis.