

## STATYSTYKA OCENY PARAMETRYCZNEJ JAKOŚCI JEDNOSTEK SFERY BADAŃ I WDROŻEŃ W KRAJU

Czesław CEMPEL

Komisja Nauki Rady Nauki; Instytut Mechaniki Stosowanej, Politechnika Poznańska

[czeslaw.cempel@put.poznan.pl](mailto:czeslaw.cempel@put.poznan.pl)

### Streszczenie

Finansowanie sfery badań i innowacji uległo znacznej zmianie po roku 1990, a uzyskało instytucjonalne wsparcie z chwilą powstania KBN w 1991. Wypracowano tam przedmiotowe dofinansowanie Instytutów Badawczych i Wydziałów Uczelnianych, które miało szereg zalet, a i sporo wad. Od kilku lat zatem zaczęły się prace nad nowym systemem oceny **Jednostek**, inaczej mówiąc parametryzacji ich osiągnięć, która wg starych zasad odbyło się ostatni raz w 2000r. W sierpniu 2005 ukazało się zarządzenie Ministra Nauki precyzujące zasady i kryteria nowej parametryzacji, Jednostki złożyły stosowne opracowania, a Ośrodek Przetwarzania Informacji (**OPI**) zebrał to i przetworzył w postaci zbioru danych na CD. Jest to ponad 800 Jednostek badawczych, od teologii do obronności, potraktowanych w ten sam sposób. Jest to zatem doskonała okazja do zbadania możliwości opisu statystycznego tej populacji i wyciągnięcia na wierzch jej istotnych charakterystyk. Praca niniejsza poświęcona jest temu zagadnieniu, a mniej wchodzi w merytoryczne kwestie jakościowe, które były przedmiotem wielu dyskusji w KBN, niestety w większości po opublikowaniu Zarządzenia.

Słowa kluczowe: opis populacji, sfera badań i wdrożeń, jednostki badawcze, opis parametryczny, opis statystyczny, optymalizacja.

## STATISTICS OF PARAMETRIC QUALITY ASSESSMENT OF RESEARCH AND INNOVATION UNITS IN POLAND

### Summary

The change of political system in Poland in 1989 has given also the change in financing of educational and research entities such as University Faculties and research institutions of Polish Academy of Sciences and innovation institution in industry. In the year 1991 the **State Committee for Scientific Research** as an equivalence of Ministry of Science and Technology was created and a new method of financing of research and innovation domain in the country was introduced. In short, two streams of money for every entity from that area were possible. The first the subjective stream, given almost to every entity, have the essential goal to keep the entity in life and work. The second stream based on individual grants given to individuals, or research teams after successful end of the process of assessment of a given application for research or innovation.

The subjective stream of financing of research and innovation entity is connected with the assessment of the quality of every individual entity (*unit*). Such assessment is based on yearly records supplied by the unit in question, and on the basis of that record special quality assessment form has been prepared by OPI, with some number of points for every important activity (*publications, patents, license implementation, etc*) in the area of research and innovation. In this way some parametric assessment of the research quality has been possible. On the basis of such parametric assessment some quality categories has been foreseen, as the main help in financing the basic activity of research and innovation entities in Poland.

The goal of this paper is to characterize statistically the population of research entities in Poland obtained by this assessment process, and in this way to give some methodological guidelines, how to design the quality categories and how to describe objectively the quality of every entity. Using some statistical methods applied to the data describing the population it was found that we have right hand side Pareto like distribution of research quality, and it is possible on this basis to divide population into four quality categories, for example. As the measure of quality the Euclidean distance of every entity from the defined '**best entity**' in a given subpopulation has been proposed. Some open question remains, how to define subpopulation, and how to make normalization of results to a given entity, if any - what was illustrated by some calculations. And this seems to be most important problem awaiting for solution.

Keywords: populations, research units, quality assessment, statistics, quality categories, singular value decomposition, optimization of unit description.

## 1. WSTĘP

Wypracowany w czasach Komitetu Badań Naukowych (KBN, od 1991 -2005r, [1]) system finansowania sfery badań i wdrożeń w kraju polegał na przedmiotowym dofinansowywaniu Jednostek Badawczo Rozwojowych, stosownie do typu jednostki i jej osiągnięć. Pod pojęciem Jednostki mieszczą się **Wydziały** Szkół Wyższych oraz niezależne organizacje i instytucje typu instytutów PAN, instytutów przemysłowych i Ośrodków badawczo Rozwojowych (**OBR**) różnego typu. Dla pokrycia całego spektrum aktywności badawczej tych jednostek, od teologii do badań obronnościowych, utworzono w KBN 12 zespołów merytorycznych<sup>1</sup>, do zadań których należała bieżąca obsługa sfery projektów i grantów różnego typu, oraz ocena jakości badań i wdrożeń w jednostkach. Ta ocena przeprowadzana była średnio raz na cztery lata (ostatnio w 2000r) i dzieliła ogół jednostek na **kategorie jakości od I do IV**, przy czym w najlepszej pierwszej kategorii mogło być maksymalnie **20%** ogółu jednostek w danym Zespole (*prawo Pareto?*).

Krótko po tym zaczęło się w KBN myślenie o utworzeniu w miejsce KBN Ministerstwa Nauki, oczywiście z nowymi zasadami pracy i współpracy ze sferą badań i wdrożeń, w tym też z nowymi (*lepszymi* /) zasadami kategoryzacji jednostek. Stało się to faktem w roku 2005, kiedy w początku roku KBN przekształcono w Ministerstwo Nauki i Informatyzacji, a 4 sierpnia 2005 Minister wydał zarządzenie precyzujące nowe zasady oceny i kategoryzacji **Jednostek** na podstawie ich ankiet z lat 2001- 2004 przesyłanych corocznie do Ośrodka Przetwarzania Informacji (**OPI**). Jak się wydaje w chwili obecnej, już po zebraniu nowych Ankiet i Kart Oceny Jednostek, nie są one ulepszeniem poprzednich kryteriów ocenowych stosowanych przez KBN, lecz zostały napisane od nowa z silnym uwypukleniem strony innowacyjno wdrożeniowej<sup>2</sup> (*najmniejsze wdrożenie jest równoważne nowej profesurze – 30pkt*).

W nowej **Karcie Oceny** Jednostek całość ich możliwej sfery aktywności podzielono na **trzy strumienie**. Pierwszy strumień aktywności ujmuje nowo pozyskane stopnie, tytuły, laboratoria, projekty badawcze i celowe, oraz nagrody. Każdemu składnikowi opisowemu tego strumienia przypisano stosowną liczbę punktów, po to by ich sumę  $P_I$  podzielić przez liczbę zatrudnionych w jednostce  $N$  otrzymując efektywność pierwszego strumienia aktywności  $E_I$ . Podobnie został sparymetryzowany drugi strumień aktywności ujmujący publikacje i monografie wszelkiego typu, z naciskiem na piśmiennictwo anglojęzyczne, co dało punkty  $P_{II}$  i efektywność  $E_{II}$ . Trzeci strumień aktywności

w charakterystyce jednostki, to patenty, licencje i wdrożenia efektów badawczo rozwojowych (**B+R**) zobrazowany punktowo przez  $P_{III}$  i efektywność policzoną jak poprzednio  $E_{III}$ . Te trzy strumienie aktywności mają w rozporządzeniu Ministra z sierpnia 2005 różne **wagi** w ocenie różnych jednostek. Dla jednostek powiązanych z gospodarką waga wdrożeń dochodzi nawet do **5** w skali **10** stopniowej, a z wagą **1** dla wydziałów humanistycznych szkół wyższych. Ponadto nie jest pewne, który z tak stworzonych symptomów lepiej ocenia jednostkę; punkty  $P$  czy efektywności  $E$ . A może ich kombinacja? Warto to obiektywnie zbadać<sup>3</sup>.

W lutym br. członkowie Rady Nauki (**RN**) otrzymali ten całościowy materiał do oceny parametrycznej jednostek przedstawionych na nośniku CD, gdzie zawarto cząstkowe punkty  $P_i$  oraz efektywności  $E_i$  dla **832** Jednostek. Znaczy to, że weryfikację i zastosowanie wag przy ocenie łącznej liczby punktów  $P$  i łącznej efektywności  $E$  dla jednostek pozostawiono weryfikatorom z **RN**. Weryfikacja ta potrwa jeszcze trochę, bo w nadesłanym materiale wykorzystano skutecznie wiele **niejednoznaczności** pytań Ankiety, łącznie z liczbą zatrudnionych  $N$ . Weryfikacja ta zmieni położenie niektórych jednostek na **skali jakości** badań i wdrożeń (*symptomy*  $P$  i  $E$ ), ale jak się wydaje **nie zmieni** w sposób istotny własności tak licznego zbioru ocenianych jednostek. Takie założenie przyświeca poniższymi rozważaniami, a ich celem jest danie przesłanek metodologicznych do zadania **kategoryzacji** jakości jednostek sfery badań i innowacji poprzez badanie własności statystycznych całego zbioru lub wyodrębnionych części.

## 2. DOMINUJĄCY TYP ROZKŁADU OCEN JAKOŚCI JEDNOSTEK

Jak wynika ze wstępu w bazie danych mamy generalnie dwie miary jakości jednostek (*dwa symptomy jakości*), punkty  $P$  oraz efektywności  $E$ . Ponieważ wspomniane już współczynniki wagi będą stosowane do mało licznych grup jednostek, zastosujemy w tej analizie populacji dla wszystkich jednostek wagi jednostkowe, tak jak np. wdrożenia dla wydziałów humanistycznych, czyli z **wagą 1**.

Zatem badając statystykę ocen posłużymy zwykłą się na początku sumą ocen punktowych;  $P = P_I + P_{II} + P_{III}$ . Z tych samych powodów podobnie postąpimy dla efektywności<sup>4</sup>;  $E = E_I + E_{II} + E_{III}$ .

Wiedząc o tym popatrzymy na zbiory jednostek przez pryzmat tych dwu miar jakości, a

<sup>1</sup> Autor przez lata należał do **Zespołu T07 Mechanika, Budownictwo i Architektura**.

<sup>2</sup> Jak twierdzą naukowcy w Stanford University „**dobra nauka to taka która zmienia świat, a można to zrobić jedynie wdrażając wyniki badań**, [2].

<sup>3</sup> W chwili pisania tego materiału istniało już zarządzenie o szczegółach obliczeń kategoryzacji, ale umysł badacza zawsze pyta czy można lepiej.

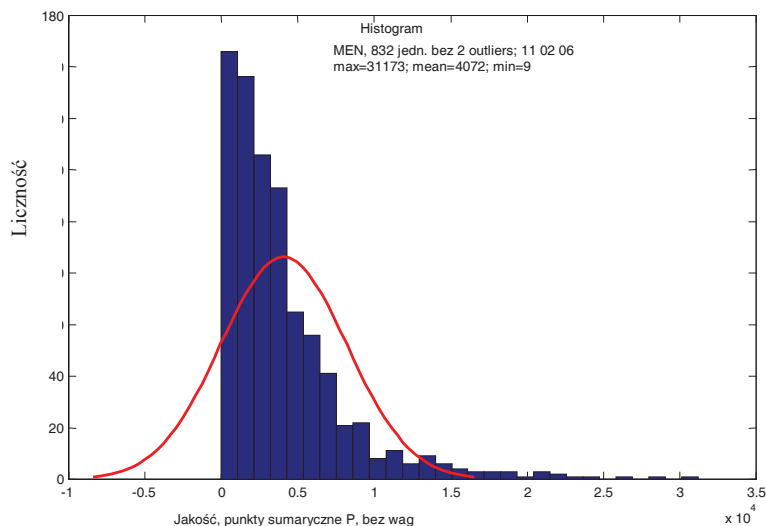
<sup>4</sup> Powyższe przyjęcie równoznaczne jest z założeniem równych jednostkowych wag;  $w_i = 1$ .

zamierzając ustalić dobry podział na kategorie (a na ile?), popatrzmy na statystykę i stowarzyszone rozkłady widziane przez te miary jakości. Wpierw popatrzmy co to za rozkład, a może normalny, to by wiele uprościło. Niestety jak widać z histogramu na rys. 1 jest to rozkład **prawo skośny** (np. *Pareto, Weibull*) a nie Gaussowski i nie możemy skorzystać z jego zasobów, np. kwantyli. Dla czytelności histogramu odrzucono dwa skrajnie daleko odbiegające wyniki (*jak można przypuszczać powstałe z innej interpretacji niejasnych pytań ankiety*). Jak się wydaje w ogóle można by odrzucić ich pięć, aż do liczby punktów 25 tysięcy.

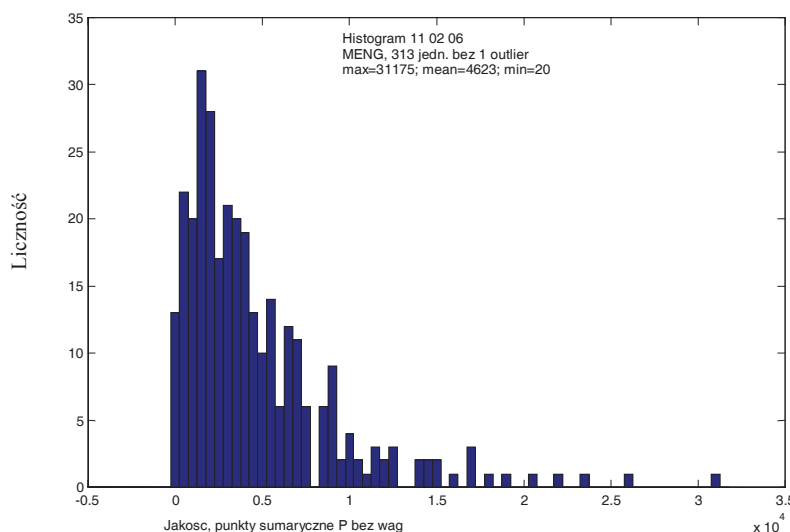
Jak widać z rysunku jest to rozkład prawo skośny - Pareto podobny, z wartością minimalną liczby punktów **P** równą **9**, wartością średnią **4072** i wartością maksymalną **31173**, (*przy odrzuceniu dwu skrajnych wyników z 832*). Tak więc wartość

maksymalna nie przekracza dziesięciokrotnej wartości średniej. Zatem gros ocen jakości jednostek dla symptomu **P** jest skupionych z lewej strony w zakresie niskich wartości, wrócimy do tego problemu jeszcze wielokrotnie.

W tym nurcie rozważań istotne jest pytanie czy ten sam charakter rozkładu mają podzbiory jednostek w **Komisji Nauki** (*jednostki humanistyczne i matematyczno przyrodnicze*) i w **Komisji Gospodarki** (*jednostki nauk technicznych, instytuty i ośrodki badawczo rozwojowe*). Jak się okazuje, obie podgrupy jednostek zachowują ten sam charakter statystyczny rozkładu, a ponieważ podgrupa Komisji Gospodarki jest mniejsza (*313 jednostek*) pokażemy tę własność na jej przykładzie, tak jak na rys. 2.



Rys. 1. Porównanie histogramu rozkładu punktów sumarycznych jednostek **P** z rozkładem Gaussowskim



Rys. 2. Histogram punktów sumarycznych dla podzbioru jednostek naukowo innowacyjnych Komisji Gospodarki RN (313 jedn.)

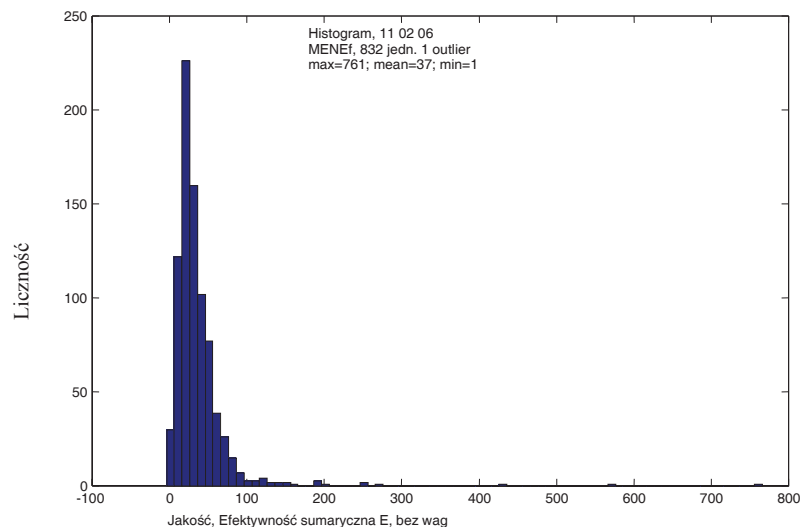
Tak więc w skali punktów sumarycznych **P** mamy zawsze do czynienia z rozkładami prawo skośnymi typu **Pareto** lub Weibulla, z kilkoma jednostkami o anormalnej liczbie punktów (*ourliers*) i po ich odrzuceniu lub zweryfikowaniu, pozostała część rozkładu nadaje się dobrze do ustanowienia jednakowych przedziałów kategoryzacyjnych, co zobaczymy w następnym punkcie.

Zanim do tego dojdziemy zobaczymy czy ten sam wniosek da się wyprowadzić jeśli jakość badań w jednostkach będziemy mierzyli w skali **efektywności** sumarycznej **E**. Sytuację tę obrazuje kolejny rys. 3, gdzie przy jednym odrzuceniu anormalnej efektywności parametry rozkładu efektywności są: minimum - **1**, średnia - **37**, maksimum - **761**. Tutaj więc wartość maksymalna skali jakości **E** przekracza **20** krotnie wartość średnią, inaczej i znacznie więcej niż w skali punktów **P**.

Podobnie dzieje się dla podgrup nauki (**n**) czy też gospodarki (**g**). Zatem efektywnościowa miara jakości ma większe tendencje grupowania większości jednostek dla małych wartości **E**, a nie narusza anormalności jednostek skrajnych. Wiąże się to prawdopodobnie normalizacją do liczby zatrudnionych  $N_i$ , inną dla każdej jednostki. Wyobraźmy sobie zatem jak będzie wyglądał sytuacja w rozkładzie **E** unormowanym do wartości maksymalnej w zbiorze (*tak jak to nakazuje wspomniane zarządzenie Ministerstwa!*) W skali liniowej będzie to jedna duża plama w okolicy zera i kilka punktów na drodze do jedynki. Jedyna wtedy rada na rozróżnienie jednostek to używanie **logarytmicznej** skali jakości **E**.

### 3. ROZKŁADY LICZNOŚCI JEDNOSTEK W PRZEDZIAŁACH JAKOŚCI I PERCENTYLE

Zmierzając do uchwycenia istoty podziału jednostek na kategorie jakości uporządkujemy



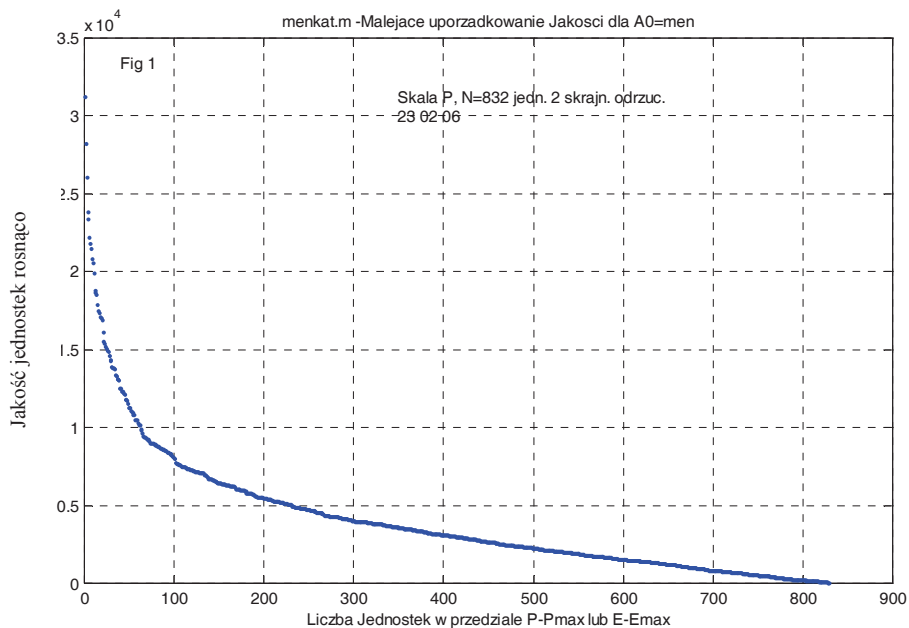
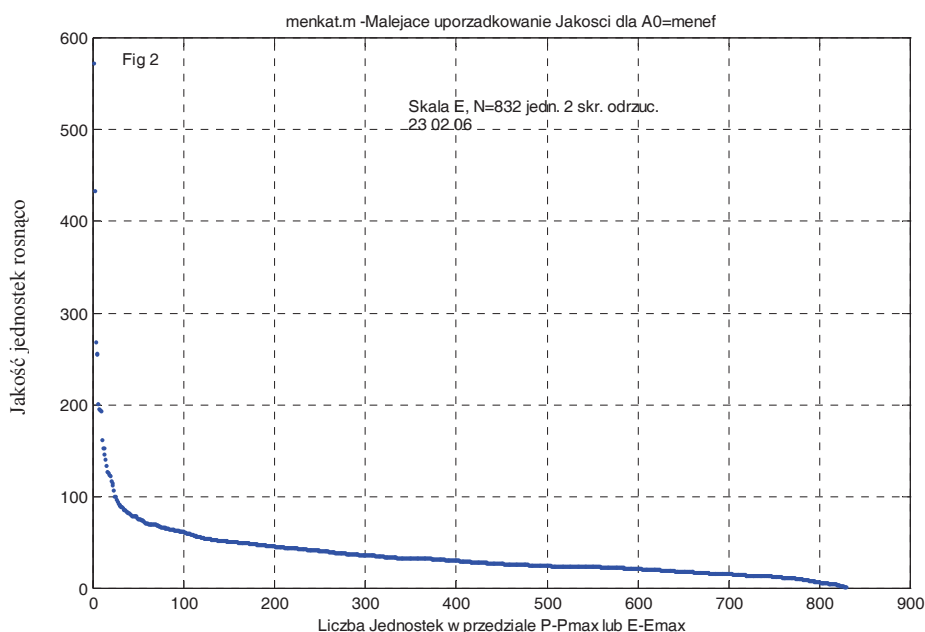
Rys. 3. Histogram łącznej efektywności badań i innowacji **E** dla całego zbioru jednostek z odrzuceniem jednego skrajnego wyniku

jednostki malejąco wg posiadanych punktów sumarycznych **P** a także efektywności **E**. W ten sposób uzyskamy łatwy do uchwycenia rozkład liczości jednostek w zadanym przedziale  $P-P_{max}$  lub  $E-E_{max}$ , tak jak na rys. 4 i 5. Jak widać z rysunku na ogólną liczbę jednostek **832** ujętych w wykazie **100** z nich ma więcej niż około 7 tysięcy punktów **P**. Podobnie dla efektywności **E** jako miary jakości, **100** jednostek ma efektywność sumaryczną większą niż **55**. Porównując rys. 5 i 6 widać również, iż rozkład **E** jest bardziej stromy i podział na kategorie jednostek może być trudniejszy niż dla punktów **P**.

Do celów klasyfikacji jednostek bardzo przydatna byłaby prezentacja miary jakości **P** lub **E** w skali procentu ogólnej liczby jednostek w zadanym przedziale jakości. Można to uzyskać jeśli z histogramów (*np. rys. 2*) policzymy sumy kumulacyjne w przedziale i zaprezentujemy je w znormalizowanej skali procentowej liczości jednostek, tak jak na rys. 6a i 6b.

Jak widać oba rys. 6 przedstawiają tę samą możliwość podziału jednostek na równo liczne grupy, np. 166 jednostek co stanowiłoby około **20 procentowe** przedziały w liczości jednostek, co na rysunkach zaznaczyliśmy jako **20%** => **166**. w tym układzie pierwsze **20%** najlepszych jednostek (*ca 166*) leży w przedziale punktowym powyżej ca **6 tys.** punktów **P** (*rys. 6a*), natomiast w skali efektywności zakres ten rozciąga się powyżej  $E=45$  (*rys. 6b*). Dodać trzeba, że znowu wykres percentyli efektywności **E** jest bardzo stromy co może utrudniać klasyfikacje (*mala rozróżnialność jednostek*), odwrotnie niż w skali punktów **P**.

Interesująca będzie odpowiedź na pytanie, jak te same charakterystyki percentyli przedstawiają się w podzbiorach Komisji Nauki (**n**) i Komisji

Rys. 4. Rozkład licznosci wszystkich jednostek badawczych (832) w zadanym przedziale P-P<sub>max</sub>Rys. 5. Rozkład licznosci wszystkich jednostek badawczych (832) w zadanym przedziale E-E<sub>max</sub>

Gospodarki (g). Jak wynika z podobnych obliczeń 104 najlepsze jednostki Komisji Nauki (20% z ogólnej liczby 518) posiada powyżej 5 tysięcy punktów P, natomiast w skali efektywności E pierwsze 20% ma efektywność E większą od około 40.

To samo pytanie zadane danym Komisji Gospodarki daje podobną odpowiedź. Tutaj liczba jednostek jest mniejsza (318) a przedziały zmienności miar są inne niż poprzednio. Jednostek. I tak 20% najlepszych jednostek (63 jedn.) posiada więcej niż 7 tysięcy punktów P. Prawdopodobnie liczącym się powodem tego wzrostu granicy punktów są wdrożenia w trzecim strumieniu aktywności. Podobnie pierwsze 20% (63jedn.)

najlepszych jednostek uszeregowane w skali efektywności E zaczynają się od liczby E=60.

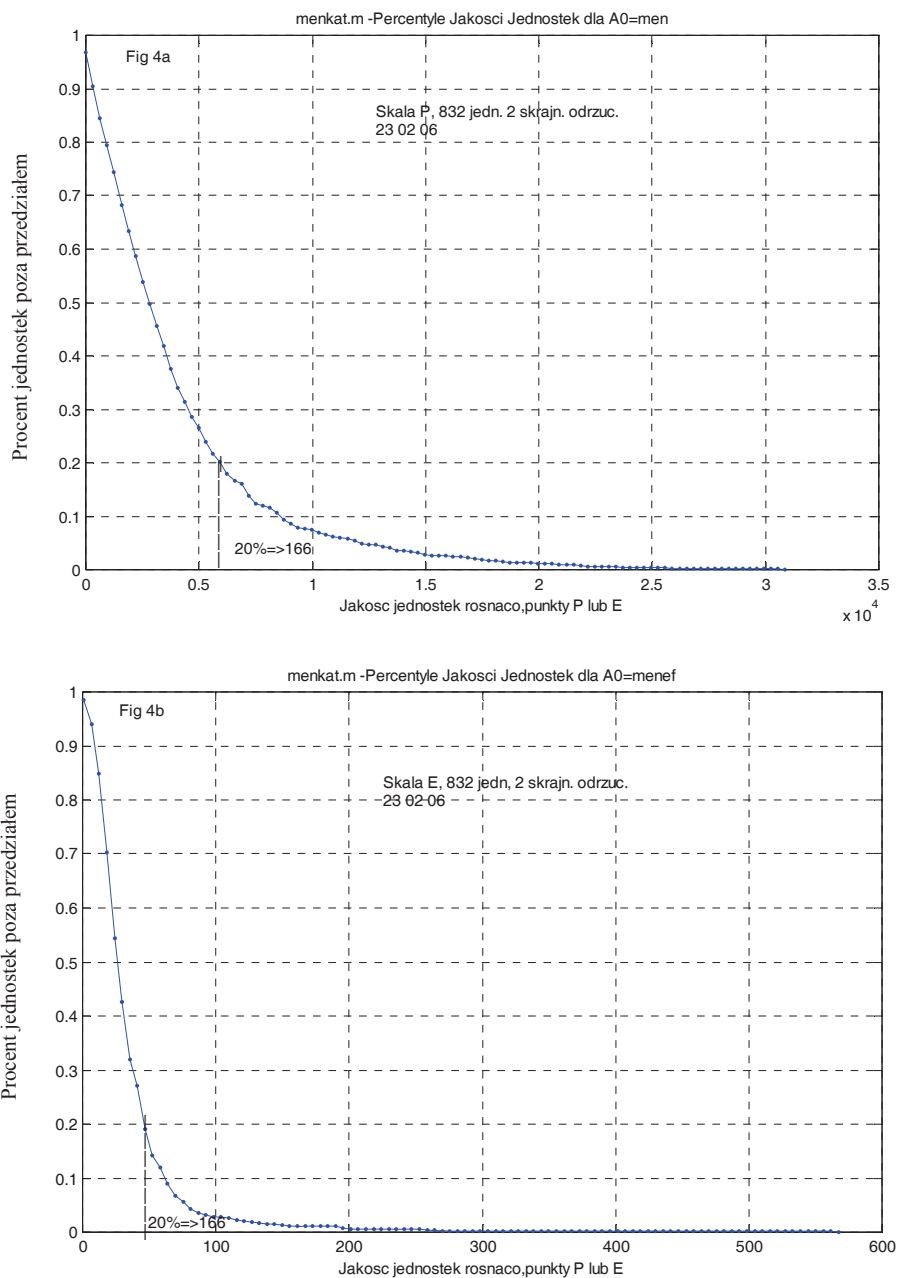
#### 4. KONCEPCJA KATEGORYZACJI JEDNOSTEK

Jak widać z powyższego miary P i E zachowują te same prawidłowości rozkładów kumulacyjnych, a także podobieństwo cech obu komisji zależnie od przyjętej symptomów jakości P lub E jest zachowane. Tak więc dla punktów sumarycznych P krzywe rozkładów percentyli są bardziej prostoliniowe dla małej liczby punktów, co umożliwi lepszy podział na kategorie jakości. Percentyle łącznej efektywności w obu podgrupach

nauki i gospodarki są podobne, niestety bardziej strome dla małych  $E$ , tam gdzie jest duże zagęszczenie jednostek. Utrudni to rozróżnienie jednostek, a także zmniejszy siłę racjonalnych argumentów na rzecz takiego rozróżnienia kategorii dla bliskich ocenowo jednostek.

Dla pokazania samej istoty proponowanej kategoryzacji jakości jednostek weźmy zbiór wszystkich jednostek w skali jakości punktów  $P$ , ułatwiających rozróżnienie, jak to już pokazaliśmy niejednokrotnie. Podzielmy cały zakres jakości  $0 - P_{\max}$  na pięć 20% przedziałów, co nam da kategorie po około 165 jednostek, tak jak na rys. 7.

Jak widać z rys. 7, prezentującego przedziały jakości w skali sumarycznych punktów  $P$ , koncepcja ta jest do przeprowadzenia, czyli jakość jednostek jest do rozróżnienia. Oczywiście będzie to bardziej wykonalne jeśli liczba kategorii zmniejszy się np. do czterech, czyli do 25% przedziałów po około 208 jednostek. Wymaga to jednak zastanowienia i symulacji w mniejszych grupach jednorodnych Komisji Gospodarki oraz Komisji Nauki.



Rys. 6. Procent całej liczby jednostek (830) jako funkcja przedziału punktów  $P-P_{\max}$  (a), oraz przedziały efektywności  $E-E_{\max}$  (b)



## 5. SKALOWANIE ROZKŁADÓW

Wspomnieliśmy już kilkakrotnie, że zarządzenie Ministra z sierpnia 2005 wymaga normalizacji efektywności  $E$  do  $E_{max}$  w danej grupie jednorodnej, oczywiście po stosownym obliczeniu efektywności **ważonej**. Popatrzmy zatem jak się może zmienić nasza rozróżnialność jednostek zależnie od zastosowanego sposobu normalizacji w danej grupie zakładając, że wykonujemy to dla efektywności  $E$  i jak zwykle z wagą  $w=1$ . Kolejny rys. 8 pokazuje trzy histogramy tego samego zbioru danych (**menef.txt**) z czterema odrzuconymi skrajnymi jednostkami, przedstawione w trzech różnych skalach; oryginalna skala efektywności  $E$  (*obrazek a*), z normalizacją do wartości maksymalnej w zbiorze  $E_{max}=268$  (*obrazek b*), oraz z normalizacją zgodną z wymogami statystyki, czyli do wartości średniej  $E_m=35,75$  (*obrazek c*). Przy sporządzaniu rysunków cały zakres  $0-E_{max}$  podzieliliśmy na **100** przedziałów.

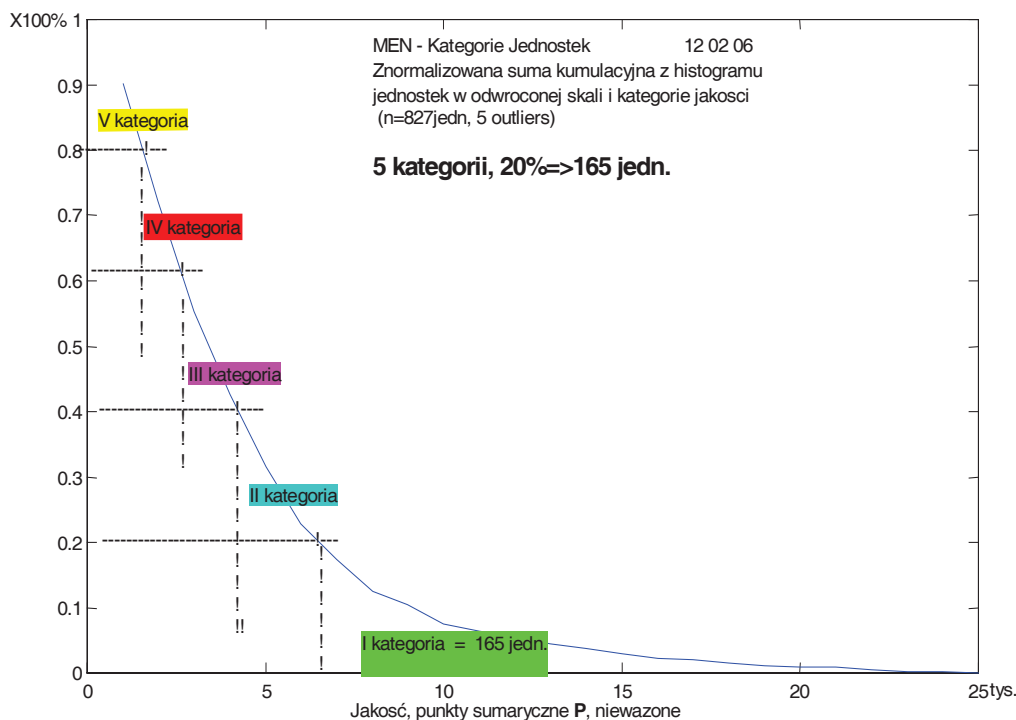
Komentując rys. 8 z punktu widzenia ostatecznego celu, czyli kategoryzacji, zwróćmy uwagę na **szczyt histogramu** z lewej zawierający prawie **80** jednostek w przedziale. Te **80** jednostek na

rysunku 8a (*bez normalizacji*), rozmieszczone jest w przedziale  $E_{max}/100 = 2,68$  punktu skali efektywności. Natomiast na rys. 8b, przy normalizacji do wartości maksymalnej  $E_{max} = 268$ , ta sama liczba jednostek rozmieszczona jest w przedziale **0,01**. Z kolei zaś dla normalizacji zgodnej z wymogami statystyki, czyli do wartości średniej zbioru, te **80** jednostek mieści się w przedziale **0,08** unormowanej skali efektywności.

Zatem jaką metodę normalizacji wybrać by najlepiej rozróżnić jednostki?

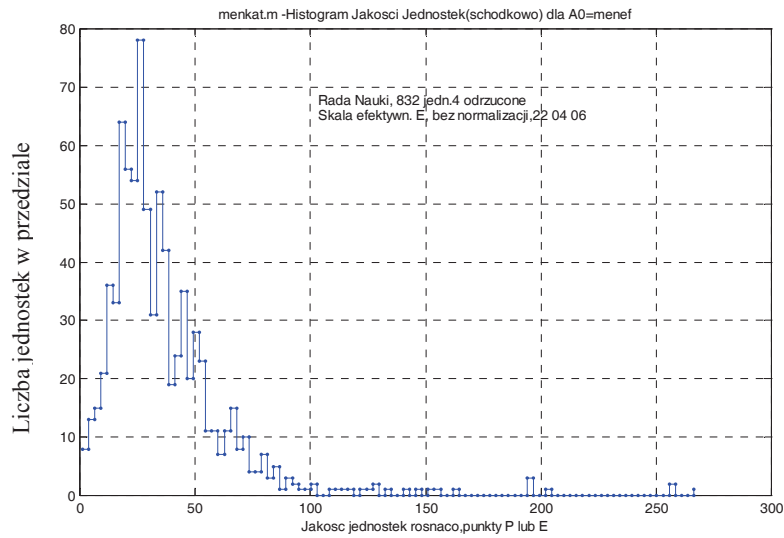
Odpowiedź jest jedna, **nie normalizować** wyniku, a wtedy rozróżnialność jednostek będzie najlepsza i tym samym przynależność do danej kategorii jakości najlepiej i **obiektywnie umotywowana** i rozróżnialna.

A jak by wyglądała tak kategoryzacja przy założonym podziale na 4 kategorie jakości wg efektywności jako symptomu jakości jednostki i bez żadnej normalizacji? Rys. 9 pokazuje taką hipotetyczną możliwość dla całej Rady Nauki (*832 jednostki*).

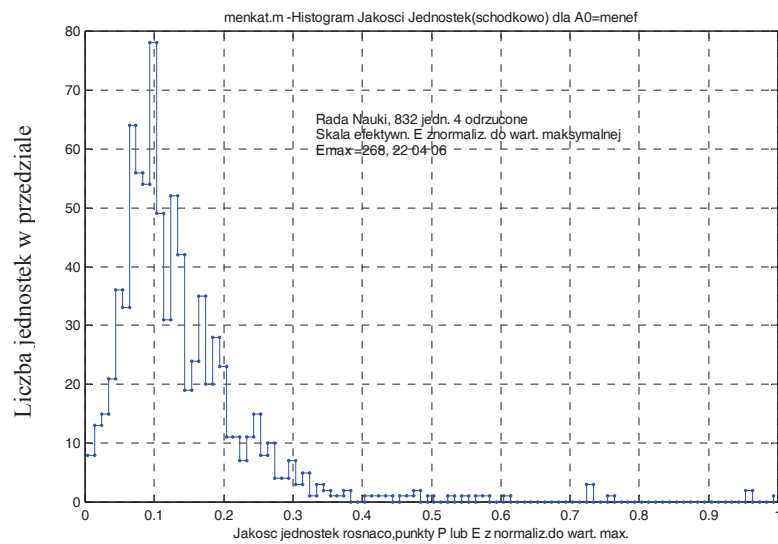


Rys. 7. Propozycja podziału jakości jednostek na pięć równych przedziałów 20% po około 165 jednostek

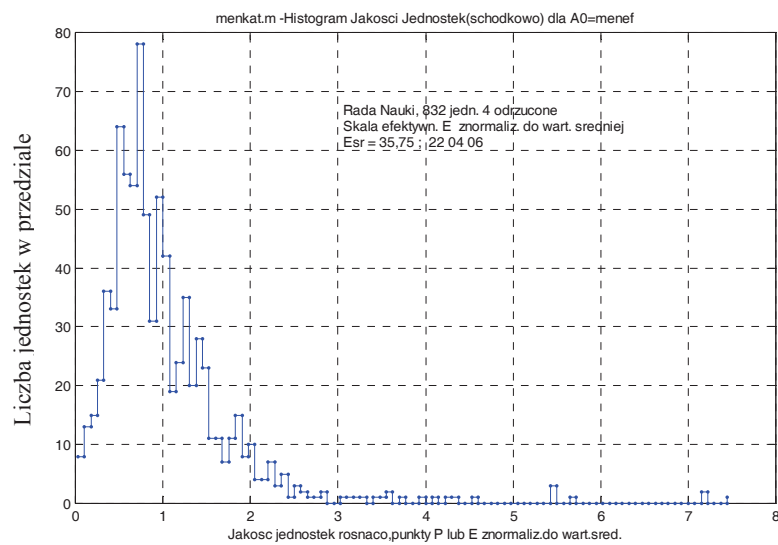
A



B

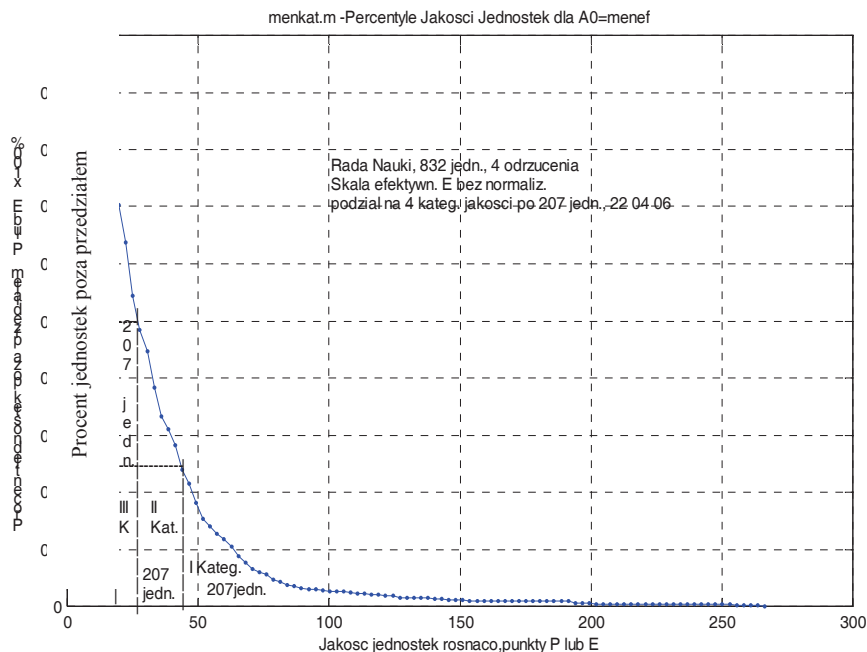


C



Rys. 8. Trzy różne skale prezentacji histogramu efektywności tego samego zbioru jednostek całej Rady Nauki





Rys. 9. Czysto algorytmiczny podział jednostek całej Rady Nauki na 4 kategorie jakości po 25%, czyli po 207 jednostek

Jak widać z rys. 9 zastosowane równe przedziały 25 procentowe generują nierówne przedziały w skali efektywności E, lecz zdrowo rozsądkowa ingerencja przywróci tu z pewnością zamierzony ład.

## 6. OPTIMALIZACJA OPISU ZBIORU JEDNOSTEK

Dążąc do prostoty opisu zagadnienia wzięliśmy wyżej pod uwagę najprostszy wektor opisujący stan jednostek, mianowicie wektor o dwu symptomach (*miarach*) jakości {E, P} a nawet traktowaliśmy te symptomy alternatywnie; rozważając oddzielnie E lub P. Zastanówmy się obecnie nad lepszym opisem tego zbioru jednostek, a więc na wektorze cech o większej liczbie symptomów. A może wśród tej zwiększonej wymiarowości opisu znajdziemy nowy zbiór **cech niezależnych** (*ortogonalnych*) charakteryzujących każdą jednostkę?

Zarówno P jak i E są sumami odnośnych punktów  $P_i$  i efektywności  $E_i$  ( $i=1, 2, 3$ ). Weźmy zatem jako składowe wektora opisującego jednostkę jej punkty  $P_i$  i efektywności składowe  $E_i$  oraz sumaryczne, co maksymalnie da nam wektor symptomów S opisujący jednostkę o ośmiu składowych

$$S_0 = \{P_1, E_1, P_2, E_2, P_3, E_3, P, E\} \quad (1)$$

Dla zbadania istotności tych symptomów jakości i możliwości prostszego optymalnego opisu jednostki zastosujemy koncepcje składowych głównych (*principal component*) znaną w analizie wielowymiarowych populacji (*PCA – Principal Component Analysis*) [5]. Operacja ta staje się prostsza i lepiej interpretowana jeśli dodatkowo

zastosujemy rozkład naszej macierzy obserwacji (*tworzą ją kolejne 832 realizacje wektora cech  $S_0$  dla każdej jednostki*) wg składowych szczególnych, tzw. w nomenklaturze angielskiej *SVD (Singular Value Decomposition)*. Program taki wykorzystałem już do badania populacji ocen ekspertów dotyczących wniosków o wsparcie badań przez **Fundację Nauki Polskiej** [4], stąd też z drobnymi zmianami posłużę się nim tutaj.

Jak wynika z tych badań najlepsze rezultaty w ekstrakcji optymalnego opisu populacji otrzymuje się kiedy symptomową macierz obserwacji (**SOM**) przekształcimy przez centrowanie i normalizację jej kolumn (*symptomów*) do wartości średniej. Podobna transformacja  $SOM^5$  sprawdza się i w naszym przypadku co widać z rys. 10, gdzie kolejne obrazki przedstawiają ważne etapy i rezultaty obliczeń.

I tak, obrazek lewy górny przedstawia w zarysie całą macierz obserwacji z 8mioma symptomami w skali jednostek w liczbie **832**. Jak widać, poza sumaryczną wartością punktów P (*największym z symptomów*), niewiele tu można zauważyć, co najwyżej wyciągnąć wniosek o dużym skomplikowaniu zagadnienia. Obrazek prawy górny pokazuje liczbę niezależnych składowych informacyjnych zawartych w naszej macierzy obserwacji i jak widać mimo 8śmiu składowych mamy tylko sześć niezależnych

<sup>5</sup> W naszym przypadku każda jednostka to wiersz ośmiokolumnowej SOM ze swoimi wartościami wektora cech. Zatem nasza SOM zawiera 832 wiersze i 8 kolumn opisujących jakość jednostki.

źródeł informacji (dwa słupki zerowe a dwa dalsze mniejsze niż 10%). Domyśleć się można z własności ortogonalności rozkładów **SVD** oraz sposobów tworzenia naszych symptomów, że w naszym wektorze (1) sumy punktów **P** i efektywności **E** można by odrzucić jako symptomy zależne, już raz uwzględnione w poprzez **P<sub>i</sub>** oraz **E<sub>i</sub>**.

Obrazek środkowy lewy przedstawia trzy pierwsze składowe główne **SG<sub>i</sub>**, (lub szczególne), przedstawione w skali wszystkich **832** jednostek. Jak widać da się tu z trudem odróżnić trzy kolory, ale liczba jednostek nie ułatwia rozróżnienia. Natomiast te same składowe główne przedstawione są w skali swych składowych czyli pierwotnych symptomów wektora **S<sub>0</sub>** na obrazku prawy środkowym, który jest dobrze rozróżnialny i niezwykle istotny dla naszej analizy. Widać na nim trzy niezależne charakterystyki naszej populacji i sposób ich pozyskania tak jak na obrazku prawym dolnym pokazującym wagi pierwotnych symptomów **P<sub>i</sub>**, **E<sub>i</sub>**. Charakterystykom tym (składowym głównym **SG<sub>i</sub>**) odpowiadają oczywiście proporcjonalne wysokości słupków na rysunku prawym górnym. Tak więc pierwszą niezależną charakterystykę naszej populacji jednostek **SG1** (40% informacji) uzyskamy z prostej ważonej sumy **P<sub>3</sub>** i **E<sub>3</sub>**, jeśli pominiemy **E** i **P** z uwagi na ich zależność od składników cząstkowych, o czym już mówiliśmy. Kolejna 20% charakterystyka **SG2**, jak widać z obrazka prawego dolnego, to liniowa

kombinacja stosownych wag i wszystkich symptomów **P<sub>i</sub>** i **E<sub>i</sub>**. Wreszcie trzecia składowa główna **SG3**, najmniejsza, tworzona jest przede wszystkim jako liniowa kombinacja symptomów **E<sub>1</sub>**, **P<sub>3</sub>** i **E<sub>3</sub>**, bo udziały innych symptomów (obrazek dolny prawy) są prawie zerowe. Co jest ważne, że te same wnioski o składowych głównych można wyciągnąć analizując podzbiory jednostek Komisji Nauki (518 jednostek) oraz Komisji Gospodarki (313 jednostek), czego już nie pokazujemy.

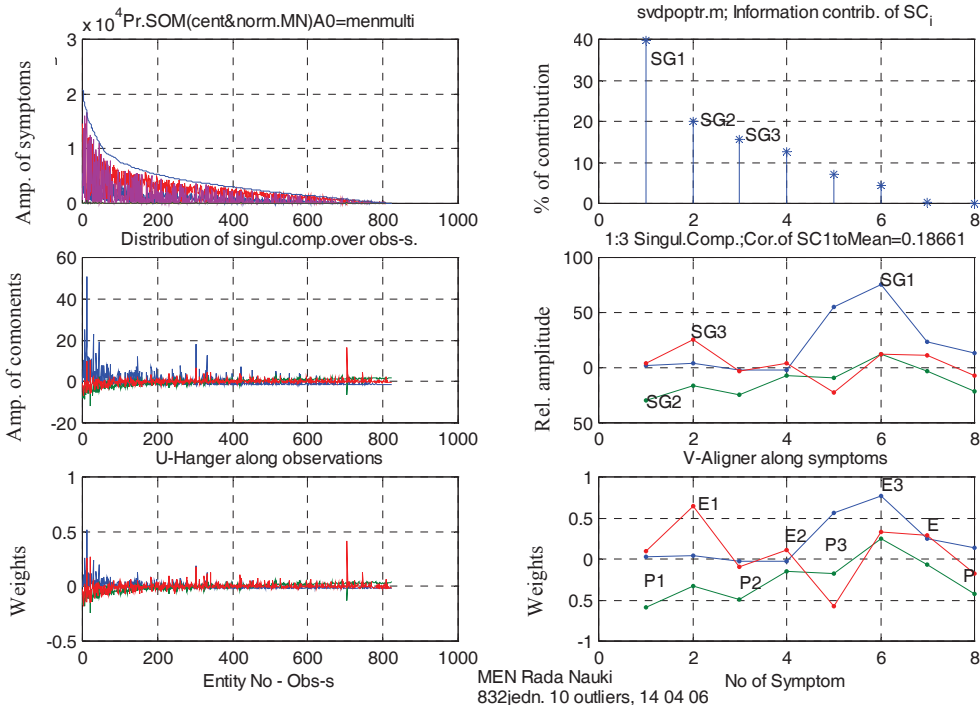
Zatem analiza nasza pokazała, że potrafimy wyróżnić trzy maksymalnie informacyjne cechy (symptomy) naszej populacji jednostek;

**SG1** jako opis aktywności wdrożeniowej (głównie **P3**, **E3**),

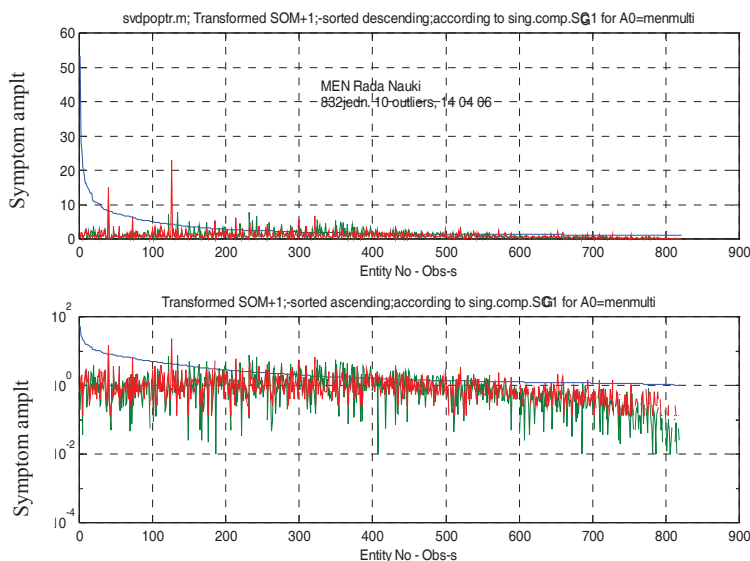
**SG2** jako ważona suma wszystkich strumieni aktywności jednostek (**P1**, **E1**, **P2**, **E2**, **E3**),

**SG3** jako osiągnięcia naukowo wdrożeniowe (**E1**, **P3**, **E3**).

Zachodzi jeszcze pytanie czy potrafimy dokonać kategoryzacji podług tak wyróżnionych składowych optymalnego opisu. To znaczy czy potrafimy pogrupować jednostki zgodnie z wybraną składową główną. Okazuje się że tak, jak to wynika z rysunku 11 gdzie przykładowo pogrupowano wszystkie jednostki zgodnie z malejącą wartością **SG1**.



Rys. 10. Składowe główne opisu całej populacji jednostek przy 8miu symptomach opisujących jakość jednostki przedstawione wzdłuż osi zbioru jednostek (z lewej) i zbioru symptomów (z prawej)



Rys.11. Przykład uszeregowania pierwszych trzech kolumn macierzy obserwacji wszystkich 832 jednostek zgodnie z malejącą wartością pierwszej składowej głównej SG1, w skali liniowej (górze) i logarytmicznej (dół)

Widać z powyższego rysunku, iż różnica w wartościach symptomu uogólnionego **SG1**, klasyfikującego jednostki rozciąga się od **0** do **50**, a skala logarytmiczna ułatwiłaby nam tu zadanie wielce (obrazek dolny).

Jednak w obecnej kategoryzacji jednostek możemy dysponować tym, co mamy i to najlepiej w minimalnym rozmiarze wektora symptomów. Analizując obrazek prawy dolny rys. 10 zauważmy, że najbardziej istotne są tam składowe wektora symptomów  $S = \{E_1, E_2, E_3, P_c\}$ . Wybierzmy zatem taki minimalny opis do dalszej analizy i zamiast rozważać kwestie jego optymalizacji popatrzymy na ułatwienie możliwości klasyfikacji (kategoryzacji jednostek). Zatem w naszym programie obliczeniowym stworzymy sztucznie **najlepszą jednostkę**, taką co ma maksymalne wszystkie cechy, tzn. nasz wzorec jakości będzie:  $S_{max} = \{E_{1max}, E_{2max}, E_{3max}, P_{cmax}\}$  i obliczymy odległości istniejących jednostek od tego idealnego wzorca. Oczywiście zasadą jest, że im bliżej wzorca jakości znajduje się dana jednostka tym lepsza jej kategoria. Taką propozycję kategoryzacji dla jednostek Komisji Gospodarki przedstawia niżej kolejny rys. 12.

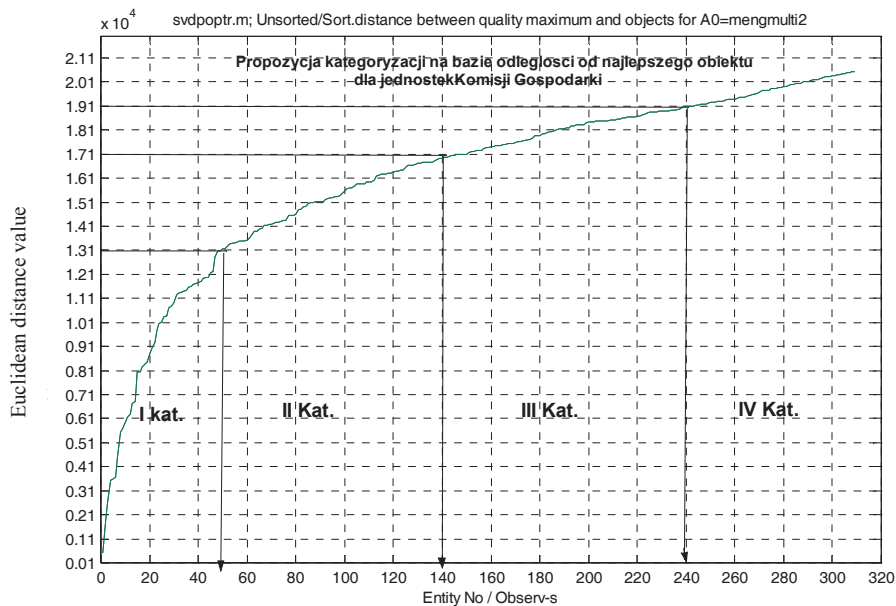
Konieczne dane do takich obliczeń znajdują się wszystkie w arkuszu kalkulacyjnym **Excel** dostarczonym przez **OPI**, można by zatem w jednej dodatkowej kolumnie obliczyć odległość  $D_i$  dla każdej jednostki, wg prostego wzoru na odległość Euklidesową (lub inną);

$$D_i = \{ (E_{1max} - E_{1i})^2 + (E_{2max} - E_{2i})^2 + (E_{3max} - E_{3i})^2 + (P_{cmax} - P_{ci})^2 \}^{1/2}. \quad (2)$$

I to już wszystko, bo **Excel** pogrupuje następnie rosnąco jednostki wg ich odległości od najlepszej, a nam pozostaje tylko zastosować zdrowy rozsądek przy walidacji wprowadzonych danych i korygowaniu granic takich kategorii jakości (patrz rys. 12).

## 7. PODSUMOWANIE

Metodologia kategoryzacji jednostek podjęta wyżej jest frapującym problemem aplikacji statystycznej teorii decyzji. Analizując całość danych, początkowo w kategoriach dwu zmiennych opisujących (symptomów jakości jednostek) pokazaliśmy, że cała populacja jednostek i podgrupy jednostek Komisji Nauki i Komisji Gospodarki mają ten sam charakter **Pareto** podobny rozkładów prawo skośnych. Analizując zaś rozkłady kumulacyjne jednostek pokazaliśmy, że można pogrupować jednostki na pięć kategorii po 20% każda (rys. 7 lub na 4 rys. 9). Poruszyliśmy również problem normalizacji wyniku efektywności każdej jednostki do wartości reprezentatywnej dla danej podgrupy. Okazuje się z tego, że najlepsze rozróżnienie jednostek daje brak normalizacji. Następnie podjęliśmy temat optymalnego opisu populacji jednostek formułując wektor obserwacji o większym wymiarze, ośmiu symptomów składowych:  $P, E, P_i, E_i, i=1, 2, 3$ . Za pomocą analizy składowych głównych pokazaliśmy, że do optymalnej charakterystyki jednostek wystarczy sześć składowych wektora  $\{P_i, E_i\}, i=1,2,3$ . Pokazano także wstępnie, że da się klasyfikować jednostki za pomocą wyłonionych przez **PCA** składowych głównych  $SG_i$ .



Rys. 12. Propozycja kategoryzacji jednostek na podstawie odległości od obiektu najlepszego w czterowymiarowej przestrzeni opisu efektywności cząstkowych i punktów całkowitych  $\{E_1, E_2, E_3, P_c\}$

Nie mając jednak zamiaru robić rewolucji, np. przez propozycję nowych pytań ankiety<sup>6</sup>, zaproponowano prostą metodę klasyfikacji jako **odległość jednostki** od obiektu idealnego dla danego zbioru jednostek, co automatycznie na podstawie obecnych danych (zweryfikowanych) mogłoby wyliczać program Excel.

Jakie zatem praktyczne wnioski wynikają z powyższego dla aktualnie przeprowadzonego procesu kategoryzacji? Jak się wydaje można je ująć w kilku punktach jak niżej.

1. Paretoński charakter populacji implikuje, że kategorie nie powinny być liczniejsze niż **20%** danej grupy, bądź też co najwyżej **25%**.
2. Najlepszą różnorodność jakości jednostek dają nieznormalizowane punkty **P** lub **E**.
3. W większości przypadków analizy rozkładów populacji jak i w **PCA** istotną rolę odgrywały efektywności cząstkowe  $E_i$  jak i punkty cząstkowe  $P_i$ . **Nie da** się więc powiedzieć jednoznacznie, że jakość jednostek opisują tylko efektywności  $E_i$ , bądź ich ważone sumy, trzeba zatem w kategoryzacji używać punktów **P** jak i efektywności **E**.
4. Dążąc do ulepszenia efektów, kategoryzacji można by ostrożnie próbować kategoryzacji według pokazanej koncepcji składowych głównych  $SG_i$ , lecz wymagałoby to dalszych obliczeń i symulacji, oczywiście dla zweryfikowanych już danych wejściowych.

Nie mając obecnie czasu i możliwości zaprojektowania i przeprowadzenia ponownie akcji oceny jednostek (*nowa ankietą*), można by przyjąć klasyfikację jakości jako odległość od zdefiniowanej

jednostki najlepszej, posiadającej najlepsze wartości swych cech opisujących. Jak wynika z obliczeń mogą to być cech opisane wektorem  $S=\{E_1, E_2, E_3, P_c\}$ , a odległość Euklidesowska jednostek w każdej zdefiniowanej subpopulacji jest do policzenia w ramach programu **Excel**, (patrz rys. 12).

## 8. LITERATURA

- [1] Strona internetowa Komitetu Badań Naukowych: <http://www.kbn.gov.pl>, kwiecień 2006.
- [2] Bendyk E.: *Potęga Google*. Polityka, No 13, 1 kwiecień 2006, s78-80.
- [3] Dane zawarte w CD: *System oceny parametrycznej*, 2001-2004, OPI, Warszawa, luty 2006.
- [4] Cempel C., *Rozkład symptomowej macierzy obserwacji populacji jako pomoc w ocenie jakości wniosków i obiektów*. Diagnostyka, 35, 2005, s7-12.
- [5] Internet: [http://en.wikipedia.org/Wiki/Principal Component Analysis](http://en.wikipedia.org/Wiki/Principal%20Component%20Analysis). Kwiecień 2006.
- [6] Internet: [http://en.wikipedia.org/Wiki/Singular Value Decomposition](http://en.wikipedia.org/Wiki/Singular%20Value%20Decomposition). Kwiecień 2006.
- [7] Wachowiak P., (red): *Pomiar Kapitału Intelktualnego Przedsiębiorstwa*. Wyd. SGH, 2005, s 167.

Prof. dr hab. **Czesław CEMPEL** informacja na stronie 25.

<sup>6</sup> Ustanowionych np. wg metod pomiaru kapitału intelektualnego [7, rozdz. 7].