Fundusze Europejskie Polska Cyfrowa | Rzeczpospolita Polska | Unia Europejska Europejski Fundusz Rozwoju Regionalnego

„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)", projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20

Graniczna Analiza Danych II – rozszerzenia

1 Wstęp

Graniczna analiza danych w klasycznym podejściu (przedstawionym na poprzednich zajęciach) pozwala na podział jednostek na 2 grupy: efektywnych i nieefektywnych. Dodatkowo dla każdej jednostki brany jest pod uwagę tylko jeden, najbardziej korzystny dla niej wektor wag nakładów i efektów. Na tych laboratoriach omówione zostaną podstawowe rozszerzenia metody DEA pozwalającej na porównanie między sobą jednostek efektywnych oraz ocenę jednostek przy użyciu pełnego spektrum wektorów wag.

W dalszej części laboratoriów potrzebne będą dane z drugiego przykładu z poprzednich zajęć. Dlatego w poniższej tabeli przedstawione są dane dotyczące liczby pracowników, czasu pracy oraz produkcji poszczególnych modeli telefonów przez poszczególne fabryki przykładowej firmy X.

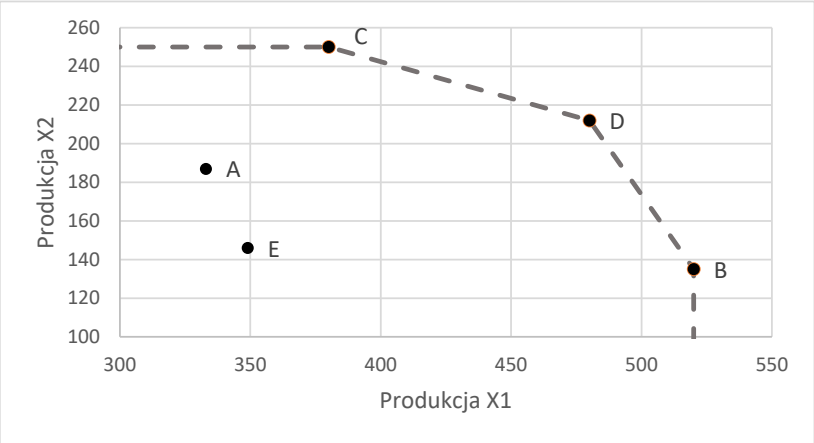
Fabryka	l. pracowników	czas pracy (h)	produkcja X1 (szt/24h)	produkcja X2 (szt/24h)
A	50	12	250	140
B	80	8	416	108
C	15	4	29	19
D	95	6	342	151
E	70	18	550	230

2 Superefektywność

Przykład Firma X produkująca telefony (przypomnij sobie przykład z poprzednich zajęć) postanowiła wyrównać zatrudnienie we wszystkich fabrykach do 100 osób. Czas pracy fabryk jest również taki sam. Mimo tego liczba telefonów (modele X1 oraz X2) produkowanych różni się w zależności od fabryki. Poniższa tabela przestawia dzienną produkcję modeli X1 oraz X2 telefonów oraz wyznaczoną wartość efektywności przy pomocy standardowego modelu CCR granicznej analizy danych.

Fabryka	produkcja X1	produkcja X2	efektywność
A	333	187	0.79
B	520	135	1.00
C	380	250	1.00
D	480	212	1.00
E	349	146	0.72

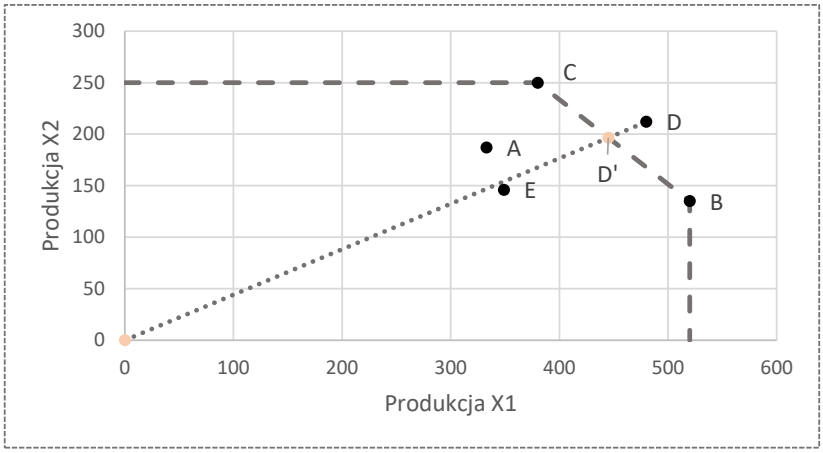
Fabryki z analizowanego przykładu wraz w granicą efektywności przedstawione są na rysunku 1.



Rysunek 1: Przykładowe fabryki telefonów firmy X wraz z granicą efektywności

Jak widać w powyższej tabeli oraz na wykresie z 5 fabryk aż 3 są ocenione jako efektywne. W wielu rzeczywistych sytuacjach istnieje potrzeba porównania ze sobą jednostek efektywnych (lub sporządzenia pełnego rankingu tych jednostek. Jedną z metod pozwalających na ocenę porównawczą jednostek efektywnych między sobą jest miara **superefektywności** (ang. super-efficiency). Opiera się ona na ocenie jaka byłaby efektywność badanej jednostki jeśli usunąć ograniczenie

o maksymalnej względnej efektywności dla tej jednostki. Ilustracja wyznaczania miary superefektywności przedstawiona jest na rysunku 2.



Rysunek 2: Superefektywność fabryki D

Na tym rysunku pokazany został przebieg granicy efektywności w przypadku usunięcia badanej jednostki (fabryki D) z analizowanego zbioru. Superefektywność mierzona jest odległością badanej jednostki od granicy efektywności wyznaczonej bez tej jednostki. W przykładzie z rysunku superefektywność fabryki D określona jest wzorem:

$$SE_D = \frac{OD}{OD'}$$

Aby wyznaczyć superefektywność jednostki DMU_o należy rozwiązać model programowania matematycznego bardzo podobny do standardowego modelu w przestrzeni efektywności, jednak usuwając z niego ograniczenie o maksymalnej względnej efektywności badanej jednostki. Przykładowo dla modelu CCR zorientowanego na nakłady model matematyczny wygląda następująco:

$$
\begin{aligned}
\max \quad & \sum_{n=1}^{N} \mu_n \cdot y_{no} \\
\text{p.o.} \quad & \sum_{m=1}^{M} \nu_n \cdot x_{mo} = 1 \\
& \sum_{n=1}^{N} \mu_n \cdot y_{nk} \leq \sum_{m=1}^{M} \nu_n \cdot x_{mk}, \qquad k = 1, 2, \ldots, K, k \neq o \\
& \mu_n, \nu_n \geq 0 \qquad\qquad m = 1, 2, \ldots M, n = 1, 2, \ldots N
\end{aligned}
$$

Analogicznie wyglądałyby modele superefektywności w przypadku orientacji na efekty oraz w modelu BCC.

Dla jednostek efektywnych superefektywność jest zawsze większa lub równa 1. Dla jesnostek nieefektywnych jest równa wartości klasycznej efektywności.

Na podstawie wartości superefektywności można zbudować pełen ranking jednostek (sortując je w kolejności malejącej superefektywności).

3 Efektywność krzyżowa

Kolejną miarą pozwalającą porównać jednostki efektywne jest **efektywność krzyżowa** (ang. cross-efficiency). Polega ona na stworzeniu macierzy K x K wartości efektywności krzyżowych. Wartość w komórce CE_{ij} macierzy efektywności krzyżowych wyznacza się jako miarę efektywności jednostki i przy użyciu wektora wag najbardziej korzystnego dla jednostki j. W tym celu należy najpierw rozwiązać standardowy model granicznej analizy danych (odpowiednio CCR lub BCC) dla każdej jednostki, aby wyznaczyć najbardziej korzystny wektor wag wejść i wyjść dla każdej z jednostek. Następnie korzystając z danego wektora wag należy obliczyć efektywność wszystkich jednostek zgodnie ze wzorem:

$$
CE_{ij} = \frac{\sum_{n=1}^{N} \mu_n^{(j)} \cdot y_{ni}}{\sum_{m=1}^{M} \nu_m^{(j)} \cdot x_{mi}},
$$

gdzie $(\mu_n^{(j)}, \nu_m^{(j)})$ to najkorzystniejszy wektor wag uzyskany dla jednostki j. Wartości na głównej przekątnej macierzy efektywności krzyżowej odpowiadają klasycznym wartościom efektywności jednostek.

Przykład Tabela poniżej przedstawia wektory wag uzyskane dla drugiego przykładu z poprzednich zajęć (2 wejścia oraz 2 wyjścia) w modelu CCR.

	ν_1	ν_2	μ_1	μ_1
A	0.0066	0.0558	0.0000	0.0064
B	0.0084	0.0406	0.0024	0.0000
C	0.0667	0.0000	0.0000	0.0203
D	0.0066	0.0618	0.0018	0.0025
E	0.0064	0.0307	0.0018	0.0000

Bazując na wagach z powyższej tabeli oraz ocenach fabryk można wyznaczyć efektywność krzyżową. Przykładowo efektywność krzyżowa fabryki C przy wykorzystaniu wag optymalnych dla D (CE_{CD}) będzie równa:

$$CE_{CD} = \frac{0.0018 \cdot 29 + 0.0025 \cdot 19}{0.0066 \cdot 15 + 0.0618 \cdot 4} = 0.288$$

W ten sam sposób można wyznaczyć całą macierz efektywności krzyżowych. **Na podstawie średniej efektywności krzyżowej dla danej jednostki (średniej z wiersza) można porównać wszystkie jednostki między sobą tworząc ranking zupełny jednostek.**

4 Badanie odporności

O ile pokazana w pierwszym punkcie miara superefektywności nadal opiera się na jednym (korzystnym) wektorze wag dla jednostki (innym dla każdej jednostki w zbiorze) o tyle efektywność krzyżowa bierze już pod uwagę pewien podzbiór wszystkich wektorów wag i ocenia na ich podstawie wszystkie jednostki. Kolejnym krokiem jest ocena jak zachowuje się miara efektywności jednostek biorąc pod uwagę wszystkie możliwe wektory wag nakładów i efektów. Na to pytanie odpowiada analiza odporności dla granicznej analizy danych. Pozwala ona na określenie zarówno ekstremalnych (minimalnej i maksymalnej możliwej) wartości miary efektywności dla każdej jednostki, jak również wykorzystanie stochastycznych metod (symulacja Monte Carlo) do określenia rozkładu miary efektywności oraz oszacowania oczekiwanej wartości miary efektywności.

4.1 Ekstremalne wartości efektywności

Maksymalna możliwa efektywność jednostki jest równa jej efektywności wyznaczonej przy pomocy standardowego modelu DEA. **Minimalna** możliwa efektywność może zostać wyznaczona przy pomocy następującego modelu:

$$
\begin{aligned}
\min \quad & \sum_{n=1}^{N} \mu_n \cdot y_{no} \\
\text{p.o.} \quad & \sum_{m=1}^{M} \nu_m \cdot x_{mo} = 1 \\
& \sum_{n=1}^{N} \mu_n \cdot y_{nk} \geq \sum_{m=1}^{M} \nu_m \cdot x_{mk} - C \cdot (1 - b_k) && k = 1, 2, \ldots, K \\
& \sum_{k=1}^{K} b_k \geq 1 \\
& \mu_n, \nu_m \geq 0 && n = 1, 2, \ldots, N, m = 1, 2, \ldots, M \\
& b_k \in \{0, 1\} && k = 1, 2, \ldots, K,
\end{aligned}
$$

gdzie C to dowolna duża stała dodatnia.

W powyższym modelu minimalizowana jest miara efektywności badanej jednostki (DMU_o) przy zachowaniu ograniczenia, że dla co najmniej jednej jednostki efektywność jest równa 1. Jest to gwarantowane przez wprowadzenie zmiennych binarnych b_k i wymuszenie, aby suma tych zmiennych była większa lub równa 0. W przypadku, gdy zmienna b_k przyjmuje wartość 0, składnik $-C(1 - b_k)$ przyjmuje bardzo niską wartość gwarantując, że ograniczenie jest zawsze spełnione. W przypadku, gdy b_k jest równe 1 składnik $-C(1 - b_k)$ się zeruje co powoduje, że ograniczenie działa normalnie i wymusza efektywność danej jednostki co najmniej równą 1.

Wartości otrzymane przy użyciu powyższego modelu oraz klasycznej metody DEA dają w rezultacie przedział możliwych wartości efektywności dla danej jednostki, biorąc pod uwagę całe spektrum możliwych wektorów wag.

4.2 Rozkład miary efektywności

Oprócz przedziału efektywności, interesujący dla analityka może być też rozkład miary efektywności (jak często efektywność dla danej jednostki jest blisko jej dolnej granicy a jak często raczej przy górnej). Aby oszacować taki rozkład można wykorzystać symulację Monte Carlo. Polega ona na losowym doborze próbki danych i oszacowanie poszukiwanej miary na podstawie tej próby.

W przypadku DEA losowany jest zestaw n próbek wag (nakładów i efektów). Dla każdej próbki wyznaczana jest wartość efektywności wszystkich jednostek na podstawie standardowego wzoru na efektywność. Ze względu na to, że losowe wagi nie zapewniają efektywności z zakresu [0-1] to wszystkie efektywności są normalizowane poprzez podzielenie przez maksimum.

Przykład: Dla przykładu firmy X (z 2 wejściami i 2 wyjściami) wylosowano następujące 3 wektory wag:

próbka	ν_1	ν_2	μ_1	μ_1
1	0.4	0.6	0.2	0.8
2	0.1	0.9	0.4	0.6
3	0.7	0.3	0.8	0.1

Po wygenerowaniu próbek należy wyznaczyć efektywność każdej fabryki dla każdej próbki. Przykładowo efektywność fabryki A, dla próbki 1 wyznacza się następująco:

$$E_A = \frac{0.4 \cdot 50 + 0.6 \cdot 12}{0.2 \cdot 250 + 0.8 \cdot 140}.$$

W ten sposób należy wyznaczyć efektywności wszystkich jednostek. Poniższa tabela przedstawia wyznaczone wartości.

	efektywność			względna efektywność		
fabryka/próbka	1	2	3	1	2	3
A	5.96	11.65	5.54	0.79	0.75	0.65
B	4.61	15.21	5.88	0.61	0.99	0.69
C	2.50	4.51	2.15	0.33	0.29	0.25
D	4.55	15.26	4.23	0.60	0.99	0.50
E	7.58	15.43	8.51	1.00	1.00	1.00

Następnie dane agregowane są do postaci przedziałowego indeksu akceptowalności efektywności (ang. efficiency acceptability interval index, EAII) poprzez zliczenie w jakim procencie przypadków dana jednostka zawierała się w danym zakresie. Zakresy można dowolnie definiować, ale zazwyczaj wykorzystuje się n równych przedziałów. Można też wyznaczyć oczekiwaną (średnią) wartość efektywności dla każdej jednostki

Przykład: Dla przykładowych próbek i efektywności rozkład efektywności wygląda następująco (podział na 4 przedziały):

Fabryka	[0-0.25]	(0.25-0.50]	(0.5 - 0.75]	(0.75-1.00]	wartość oczekiwana
A	0	0	2/3	1/3	0.73
B	0	0	2/3	1/3	0.76
C	1/3	2/3	0	0	0.29
D	0	1/3	1/3	1/3	0.70
E	0	0	0	3/3	1.00

5 Zadanie domowe - część II

Uzupełnij skrypt zadania z poprzednich zajęć tak, aby dla wyznaczana była również **superefektywność** analizowanych lotnisk oraz **średnia efektywność krzyżowa**. Przy pomocy podanego pliku *samples.csv* zawierającego 100 próbek wag wejść i wyjść wyznacz również **rozkład efektywności dla każdej z jednostek** (przy podziale na 5 równych przedziałów) oraz oszacuj **oczekiwaną wartość efektywności jednostek**. Na podstawie superefektywności, średniej efektywności krzyżowej oraz wartości oczekiwanej **wyznacz rankingi jednostek**. Określ czy rankingi uzyskane tymi trzema miarami są zgodne i spróbuj odpowiedzieć na pytanie dlaczego są lub nie są zgodne – co wpływa na podobieństwa i/lub różnice w pozycjach rankingowych poszczególnych lotnisk.

W ramach rozwiązania prześlij skrypt w języku python oraz sprawozdanie zawierające poszczególne wyniki zadań z obu tygodni w formie pliku PDF.

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

„Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)", projekt finansowany ze środków Programu Operacyjnego Polska Cyfrowa POPC.03.02.00-00-0001/20