Metodyka obliczania rankingu w ramach systemu Blender Danych za rok 2021 i 2022 (opis procedur statystycznych w ramach algorytmu)

1. Operacje wstępne

Wyjściowym zbiorem danych jest macierz (tablica) danych charakteryzujących zawody

$$X = [x_{ij}]$$

gdzie:

- i – numer wiersza (przypadku) indeks numeru zawodu (i=1,…,m);
- j – indeks (numer) zmiennej diagnostycznej charakteryzującej zawody (j=1,…,n)

Wartość zmiennej pochodzi bezpośrednio ze zbioru, który jest importowany przez administratora do systemu po przekazaniu przez instytucję dysponującą danymi (np. liczba osób zarejestrowanych jako bezrobotne) lub też wyliczona w ramach systemu na ich podstawie (np. wyliczenie wartości procentowej w danym powiecie).

Wybór zmiennych do rankingu z bazy (tablicy) X: przy wybieraniu zmiennej w systemie zostaje ona przypisana do obszaru tematycznego jako:

- charakteryzująca rynek pod względem popytowym (np. pojawiające się oferty pracy, wzrost liczby zatrudnień) – do rankingu dostępności ofert (miejsc) pracy;
- pod względem podażowym (np. absolwenci uczelni, szkół średnich) – do rankingu dostępności zasobów pracy;
- pod względem charakterystyki miejsc pracy i ofert – do rankingu atrakcyjności ofert pracy.

Każda ze zmiennych w bazie (tablicy) X w momencie doboru do danego rankingu ma przypisywany charakteru (flagę):

- ***stymulanta*** (im wyższe wartości tej zmiennej dla danego przypadku, tym dany przypadek jest lepszy w badaniu (rankingu) ze względu tylko na tę zmienną);
- ***destymulanta*** (im niższe wartości tej zmiennej dla danego przypadku, tym dany przypadek jest lepszy w badaniu (rankingu) ze względu na tę zmienną).[1]

Jednocześnie administrator ustala wagi (ważności) dla zmiennych:

- takie same wagi dla wszystkich zmiennych: $\omega_j = \frac{1}{n}$
- ewentualnie przypisania odpowiednich wag dla każdej indywidualnej wybranej zmiennej diagnostycznej zgodnie z przyjętą metodyką[2,3]:

[1] W zestawieniu wskaźników w ramach systemu „Blender Danych" brak jest wska źników o charakterze nominant.

[2] Wagi przypisane wskaźnikom w ramach poszczególnych rankingów dostępne w dokumencie „Opis wskaźników użytych do obliczenia rankingów za rok 2021".

[3] Domyślnie rankingi w ramach systemu Blender Danych są tworzone na podstawie wszystkich wskaźników (składowych) Blendera. Użytkownik w ramach pracy w systemie ma jednak możliwość wyboru jedynie części wskaźników (spośród dobranych do systemu) do zbudowania własnych rankingów. Wagi przyjęte dla wskaźników w takim przypadku będą zachowywały proporcję wobec siebie.

$$0 < \omega_j < 1; \sum_{j=1}^{n} \omega_i = 1$$

Na tym etapie definiowania rankingu administrator określa dodatkowo sposób wyliczania wartości wspólnej dla więcej niż jednego powiatu (w zależności od zapytania użytkownika zewnętrznego) – tak dla całego województwa, jak i dowolnej kombinacji wybranych powiatów; w celu umożliwienia wyliczenia wartości wspólnej dla wybranego obszaru w odniesieniu do wszystkich dobranych typów zmiennych – z tego względu możliwe są do wyboru określenia następujące schematy postępowania:

- sumowanie wartości wskaźnika wyliczonych dla poszczególnych powiatów (np. łączna liczba pojawiających się ofert pracy);
- wyliczanie wartości procentowej (np. odsetek osób długotrwale bezrobotnych pochodzący z dzielenia łącznej liczby osób długotrwale bezrobotnych w stosunku do łącznej liczby osób bezrobotnych na wybranym obszarze);
- wyliczanie wartości średniej na podstawie średniej ważonej (np. wyliczanie wartości średniego wynagrodzenia na podstawie średnich powiatowych, z uwzględnieniem liczebności pracowników w poszczególnych powiatach).

2. Metoda obliczeniowa do rankingów – metoda TOPSIS

- **Unormowanie (normalizacja) wybranych zmiennych diagnostycznych**

W pierwszym kroku następuje przekształcenie zmiennych na wartości unormowane na skali przedziałowej poprzez **standaryzację**:

$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$; i=1,…,m; j=1…,n

gdzie:
$\bar{x}_j = \frac{\sum_{i=1}^{m} x_{ij}}{m}$ – jest średnią wartością j-tej zmiennej diagnostycznej x_j; m – liczba przypadków (wartości zmiennej); n- liczba wybranych zmiennych do analizy
$S_j = \sqrt{\frac{\sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2}{m}}$ – jest odchyleniem standardowym j-tej zmiennej diagnostycznej x_j; m – liczba przypadków (wartości zmiennej)

- **Wybór odległości pomiędzy obiektami w przestrzeni wielowymiarowej potrzebnej**

Każdy obiekt badany (zawód) charakteryzowany wybranymi zmiennymi diagnostycznymi jest interpretowany jako punkt z przestrzeni wielowymiarowej R^n. Zachodzi zatem konieczność wyboru na potrzeby rankingu odpowiedniej metryki odległości, za pomocą której będą porównywane obiekty.

Zastosowano jako miarę odległości odległość euklidesową:

- **odległość euklidesowa w przypadku wag jednakowych (lub braku wag)** pomiędzy k-tym oraz p-tym obiektem punktem w przestrzeni wielowymiarowej.

$$d_{k,p} = \sqrt{\left(\frac{1}{n}\right)^2 \sum_{j=1}^{n} (z_{kj} - z_{pj})^2} \quad \text{lub} \quad d_{k,p} = \sqrt{\sum_{j=1}^{n} (z_{kj} - z_{pj})^2}$$

- **odległość euklidesowa w przypadku wag zmiennych** pomiędzy k-tym oraz p-tym obiektem punktem w przestrzeni wielowymiarowej:

$$d_{k,p} = \sqrt{\sum_{j=1}^{n} \omega_j^2 \cdot \left(z_{kj} - z_{pj}\right)^2}$$

gdzie: $0 < \omega_j < 1, \sum_{j=1}^{n} \omega_j = 1$ – ustalony system wag dla j-tej zmiennej

- **Budowa rankingu**

W pierwszej kolejności wyznacza się abstrakcyjny obiekt najlepszy (idealny) o najlepszych wartościach zmiennych diagnostycznych (maksymalnych dla stymulant i minimalnych dla destymulant)

$$z_0 = [z_{01} \quad z_{02} \quad \dots \quad z_{0n}]$$

$$z_{0j} = \begin{cases} \max\limits_i z_{ij} & gdy \quad z_j - stymulanta \\ \min\limits_i z_{ij} & gdy \quad z_j - destymulanta \end{cases}$$

Następnie wyznaczany jest drugi obiekt (antyidealny) w przestrzeni wielowymiarowej (antywzorzec), który jest obiektem o najgorszych wartościach zmiennych diagnostycznych (minimalnych dla stymulant oraz maksymalnych dla destymulant).

$$z_{-0} = [z_{01} \quad z_{02} \quad \dots \quad z_{0n}]$$

$$z_{-0j} = \begin{cases} \min\limits_i z_{ij} & gdy \quad z_j - stymulanta \\ \max\limits_i z_{ij} & gdy \quad z_j - destymulanta \end{cases}$$

Następnie badane jest podobieństwo obiektów (badanych zawodów) do abstrakcyjnego obiektu najlepszego wyznaczając odległości obiektów od wzorca $d_{i,0}$ – jako miarę zastosowano odległość euklidesową.

Badane jest także podobieństwo analizowanych obiektów do abstrakcyjnego obiektu najgorszego, poprzez wyznaczenie odległości obiektów od antywzorca $d_{i,-0}$ – jako miarę zastosowano odległość euklidesową.

Wyznaczana jest wartość dla zmiennej agregatowej (syntetycznej) ze wzoru:

$$R_i(TOPSIS) = \frac{d_{i,-0}}{d_{i,-0} + d_{i,0}}$$

- **Wyznaczenie ostatecznych rankingów (posortowanie obiektów nierosnąco) i przypisanie rankingów porządkowych**

Wartości tej zmiennej syntetycznej należą do przedziału [0,1]. Im wyższe jej wartości dla danego obiektu, tym wyższe jego miejsce w rankingu.

Wartość wynikowa rankingu dostępna dla użytkownika zewnętrznego przeliczania jest na skalę (adekwatnie do rankingu: dostępności ofert pracy, dostępności zasobów pracy, atrakcyjności ofert pracy) od 0 do 100 (poprzez wymnożenie wartości zmiennej syntetycznej $R_i(TOPSIS) \cdot 100$).

Skala przyjęta dla prezentacji danych wynikowych dla użytkownika zewnętrznego polega na wydzieleniu pięciu klas wyników rozłożonych równomiernie na skali:

1-20 – wartość bardzo niska
21-40 – wartość niska
41-60 – wartość średnia
61-80 – wartość wysoka
81-100 – wartość bardzo wysoka