BLENDER DANYCH

Data Ranking Algorithm

Table of Contents

1. OCCUPATION VITALITY RANKING

The Data Blender system provides the capability to generate an occupation vitality ranking, which:

- allows occupations to be viewed from the perspective of several factors simultaneously — ranking them on the basis of a set of indicators that allow for their assessment in the context of the vacancy market situation, inflow of employed persons, registered unemployment, and wages paid;
- is prepared at the level of voivodeships and the country;
- is created using a data ranking algorithm based on the TOPSIS method;
- occupations are ranked as a result of their comparison on the basis of selected indicators — described in file 5_RANKING KONDYCJI ZAWODÓW – Wskaźniki (in the Data Blender Methodology tab)

2. METHODOLOGICAL DESCRIPTION OF THE DATA RANKING ALGORITHM

a) Algorithm assumptions

The data processing and ranking algorithm model in the Data Blender system was designed with the following assumptions:

- the output of the algorithm is to be an occupation ranking prepared on the basis of multiple variables. It should reflect both the supply-side and demand-side dimensions of the position of a given occupation, and the assessment resulting from this comparison. It should be possible to present both partial rankings (voivodeship level) and aggregated reports (level of several combined voivodeships, up to national scale);
- the algorithm should allow a ranking to be built regardless of the type of input data — it must therefore include normalisation procedures;
- the algorithm should enable the assignment of weights to individual variables included in the ranking, taking into account data normalisation methods appropriate for various procedures;
- the algorithm should enable the use of various methodological procedures for calculating the ranking, ensuring the possibility of a multi-method approach, i.e. one based on different ranking procedures.

The algorithm should be as automated as possible. This means embedding the analytical procedures of the algorithm within the software. Individual procedures should be accessible

at minimum from the level of the Data Blender system operator in the form of a simple selection field. Selecting a given procedure should lead to the generation of the entire analysis on the basis of the input database provided.
The algorithm should enable the analysis of the influence of socio-economic factors on the situation of a given occupation. These factors may be of various kinds (relating to, inter alia, the level of unemployment and the level of employment). The assessment of occupations in terms of their relationship to the influence of these factors provides the potential for additional assessment, and may thus constitute additional, important management information.

b) Input operations

The input data set is a matrix (table) of data characterising occupations

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

where:
- *i - row (case) number, index of the occupation number (i=1,…,m);*
- *j - index (number) of the diagnostic variable characterising occupations (j=1,…,n)*

Each variable in the database (table) X should be assignable a character (flag):
- **stimulant** (the higher the values of this variable for a given case, the better the case in the study — ranking — with respect to this variable alone)
- **destimulant** (the lower the values of this variable for a given case, the better the case in the study — ranking — with respect to this variable)
- **nominant** (the closer the values of the variable to a certain nominal point c_{0j} value or to a value within a certain interval $[c_{1j}, c_{2j}]$, the better the case in the study — ranking — with respect to this variable)

For nominant-type variables in the database, it should be possible to establish and assign nominal values in the form of a closed set.

Selection of variables for the ranking from the database (table) X: (when selecting variables, the programme should inform about the character of the variable and provide any nominal values)

- from the demand perspective (employers, workplaces, companies),
- from the supply perspective (e.g. graduates of universities and secondary schools),
- all available variables from the database.

Setting weights (importance) for variables – the following options should be available:

- assigning equal weights to all variables: $\omega_j = \frac{1}{n}$,
- in the variant where weights are equal – they may be omitted,
- optionally, assigning appropriate weights to each individual selected diagnostic variable

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

3. IMPLEMENTATION OF THE TOPSIS RANKING PROCEDURE

3.1. Determining the character of variables

The TOPSIS method assumes that all selected diagnostic variables **should have the character of stimulants or destimulants**. Therefore, variables that are nominants must be converted into corresponding variables with the character of stimulants. One of the following transformations may be applied for the conversion:

- ratio formula (only for nominants measured on a ratio scale with values that are ratio indices with values >=0) – converts to stimulants with values in the interval [0,1]

$$x_{ij} = \frac{\min_i \{c_{oj}; x_{ij}^N\}}{\max_i \{c_{oj}; x_{ij}^N\}} \tag{2}$$

i=1,…,m; j=1,…,n; where c_{0j} – *are the nominal best values of nominant variables* ,x_{ij}^N.

- difference formula (the stimulant obtained by this formula is measured on an interval scale)

$$x_{ij} = -\left|x_{ij}^N - c_{0j}\right| \tag{3}$$

i=1,…,m; j=1,…,n; where c_{0j}– *are the nominal best values of nominant variables* x_{ij}^N.

- the following formula may also be applied:

$$x_{ij} = \begin{cases} \frac{-1}{x_{ij}^N - c_{oj} - 1} & gdy \quad x_{ij} < c_{0j} \\ 1 & gdy \quad x_{ij} = c_{0j} \\ \frac{1}{x_{ij}^N - c_{oj} + 1} & gdy \quad x_{ij} > c_{0j} \end{cases} \tag{4}$$

i=1,…,m; j=1…,n

3.2. Normalisation of selected diagnostic variables

To determine the normalised values on an interval scale, the standardisation method was used.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (5)$$

where:

i=1,...,m; j=1...,n

$$\bar{x}_j = \frac{\sum_{i=1}^{m} x_{ij}}{m} \quad (6)$$

is the mean value of the j-th diagnostic variable x_j; m – number of cases (values of the variable); n – number of selected variables for analysis

$$S_j = \sqrt{\frac{\sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2}{m}} \quad (7)$$

is the standard deviation of the j-th diagnostic variable x_j; m - number of cases (values of the variable)

Each object under study (occupation) characterised by selected diagnostic variables will be interpreted as a point in multidimensional space R^n. It is therefore necessary to select an appropriate distance metric for the purposes of the ranking, using which the objects will be compared.

3.3. Selection of the distance between objects in multidimensional space required to determine the ranking.

The computational procedure employs the classical Euclidean distance measure:

- **Euclidean distance (equal weights — no weights)** between the k-th and p-th object point in multidimensional space.

$$d_{k,p} = \sqrt{\left(\frac{1}{n}\right)^2 \sum_{j=1}^{n} \left(z_{kj} - z_{pj}\right)^2} \quad \text{or} \quad d_{k,p} = \sqrt{\sum_{j=1}^{n} \left(z_{kj} - z_{pj}\right)^2} \quad (8)$$

- **Euclidean distance (variable weights)** between the k-th and p-th object point in multidimensional space.

$$d_{k,p} = \sqrt{\sum_{j=1}^{n} \omega_j^2 \cdot \left(z_{kj} - z_{pj}\right)^2} \quad (9)$$

where: $0 < \omega_j < 1, \sum_{j=1}^{n} \omega_j = 1$ *– the established weight system for the j-th variable*

3.4. Building the ranking

First, an abstract ideal object is identified, with the best values of the diagnostic variables (maximum for stimulants and minimum for destimulants).

$$z_0 = [z_{01} \quad z_{02} \quad \dots \quad z_{0n}] \quad (10)$$

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

Next, a second anti-ideal object in multidimensional space is identified (the anti-pattern), which is an object with the worst values of the diagnostic variables (minimum for stimulants and maximum for destimulants).

$$z_{-0} = [z_{01} \quad z_{02} \quad \dots \quad z_{0n}] \tag{11}$$

$$z_{-0j} = \begin{cases} \min_i z_{ij} & gdy \quad z_j - stymulanta \\ \max_i z_{ij} & gdy \quad z_j - destymulanta \end{cases}$$

The similarity of objects (occupations under study) to the abstract ideal object is then examined by determining the distances of objects from the pattern $d_{i,0}$ — the previously adopted Euclidean distance measure is to be used as the measure.

The similarity of the analysed objects to the abstract worst object is also examined, by determining the distances of objects from the anti-pattern $d_{i,-0}$ — the previously adopted Euclidean distance measure is likewise to be used as the measure.

The values for the aggregate (synthetic) variable are then determined using the formula:

$$R_i(TOPSIS) = \frac{d_{i,-0}}{d_{i,-0} + d_{i,0}} \tag{12}$$

The values of this synthetic variable fall within the interval [0,1]. The higher the values for a given object, the higher its position in the ranking.

3.5. Determination of the final rankings (sorting objects in non-increasing order) and assignment of ordinal ranks.