UDC 004.42
DOI https://doi.org/10.32782/2663-5941/2025.5.2/38

Statyvka Yu.I.
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

Nedashkivskiy O.L.
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

Mingjun Z.
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»

EVALUATING THE LEVEL OF INTERNATIONALIZATION OF SCIENTIFIC INSTITUTIONS IN THE PRESENCE OF MISSING DATA

The article is devoted to a detailed analysis and solution of the problem of evaluating the level of internationalization of scientific institutions in situations where primary statistical data are incomplete or partially unavailable, which significantly complicates the construction of objective and correct ratings.

The level of internationalization is defined as a complex multidimensional, composite indicator formed by hierarchical aggregation of numerous indicators, the calculation of which in conditions where data is incomplete should remain logical and motivated. The authors propose a clear distinction between two types of unknown information: missing data, when the indicator can be determined in principle, and unavailable, when due to the specifics of the activity, the object is not comparable with others by a certain indicator (one or more). A formalized apparatus for describing the data structure is introduced using special notations for complete indicators, indicators with missing values, with unavailable values, and mixed cases. A system of requirements in the form of predicates has been formulated that guarantee the correctness of the restoration of omissions, the exclusion of the possibility of a rating shift, and the preservation or improvement of the positions of non-comparable institutions after taking into account indicators – sources of incomparability. Analysis of possible cases of incompleteness of the source data allowed us to identify nine patterns relevant to the rating task, which became the basis for the classification of tasks and methods for constructing a rating, covering both cases of complete data ("Trivial") and situations with missing data ("Multiple Regression", "Multiple Imputation"), with non-comparable objects ("Ordinal Statistics", "Merge Partial Ratings"), as well as generalized methods of step-by-step reduction of the task to simpler ones ("Inc-Reduction", "Mix1-Reduction", "Mix-Reduction", "Gen-Reduction"). As a result, a conclusion is made about the sufficiency of the developed conceptual apparatus for formalizing the rating task in conditions of incomplete data, the features of data recovery procedures when solving rating tasks and their difference from similar tasks when studying the general population are determined, clear requirements for methods in the form of predicates are formulated, and a systematic classification of methods is provided.

Key words: *internationalization of scientific institutions, evaluation of the level of internationalization, data pattern, missing and unavailable data, composite indicators, architecture of the software system, software engineering.*

Formulation of the problem. Evaluating the level of internationalization of scientific institutions is a well-known problem [1, 2] due to the complexity and comprehensiveness of the concept of internationalization, which makes it difficult and ambiguous to choose a system of indicators. The large number of indicators needed to calculate the final evaluation of the level of internationalization causes the construction of a hierarchical structure. That is, the level of internationalization is calculated as a composite indicator by aggregating, most often as a weighted sum, dimensions (categories, policies). Each of the dimensions is also calculated as a weighted sum of either lower-level dimensions (subcategories), or directly primary indicators, which, in turn, are obtained on the basis of normalized primary (raw) data [3].

Analysis of recent research and publications. The practice of performing statistical research is faced with the standard problem of missing data – the absence of values of some variables (factors, indicators). Typification by the reason for missing data, presented in [4] and described in detail, in particular, in [5], involves missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). The specified mechanisms of missing data describe the relationships between the measured variables and the probability of missing data and function as assumptions for the analysis of missing data. In particular, MNAR also involves intentional missing data, well known from the tasks of designing fractional factorial experiments. However, all of the listed types of missing data assume that missing data, although not actually known, could, in principle, be measured. Or, in other words, all objects in the sample are comparable.

However, there are known precedents when a ranking is built on the basis of composite indicators when evaluating scientific institutions, despite the incomparability of some of them with the rest. Perhaps the most famous example is the incomparability of universities such as the London School of Economics with the rest of universities according to the N&S indicator of the Shanghai University Ranking, where missing values (Nothing/NA) are interpreted as a value of 0, and the weight of the indicator is distributed between the weights of other primary indicators [6].

So, not many works have been devoted to the analysis and solution of the problem of assessing the level of internationalization of scientific institutions in situations where primary statistical data are incomplete or partially unavailable [7, 8]. And even fewer works [9, 10] have been devoted to solving the problems of developing effective software for assessing the level of internationalization of scientific institutions in situations where primary statistical data are incomplete or partially unavailable.

Task statement. The purpose of this article is to define a set of concepts necessary for the formal representation and classification of tasks for constructing a rating by composite indicators in the presence of unknown data, that is, in the presence of missing and/or unavailable data.

Outline of the main material of the study. To achieve the goals, first of all, a specialized notation was developed and proposed. Then, with its help, the task of constructing ratings was formalized. This allowed us to formulate the principles of constructing and aggregating composite indicators. The final stage was the creation of a classification of ranking methods based on data patterns.

Used notation. Unknown data we will classified as either missing or unavailable.

Let the desired composite indicator of the i*-th* institution U_i be calculated as the sum of the values $x_{i,j}$ for n indicators of the previous level:

$$I_i = \sum_{j=1}^{n} x_{i,j} \qquad (1)$$

Without loss of generality, we will consider $x_{i,j}$ as already weighted values. Then the required data can be represented as a matrix $X = (X_1, X_2, \dots, X_n)$, where X_j is a vector of (weighted) values of the *j-th* indicator of the previous level.

To represent data formed by some of the vectors – components of the matrix X, we will use the notation $X^{\langle j_1, j_2, \dots \rangle}$. For example, $X^{\langle 1,3 \rangle}$, $X^{\langle 1-3,6 \rangle}$ mean matrices (X_1, X_3) and (X_1, X_2, X_3, X_6), respectively. All vectors of the matrix X, except for $< j_1, j_2, \dots >$, will be denoted as $X^{-\langle j_1, j_2, \dots \rangle}$. The number of matrix components will be denoted by vertical bars, then $|X| = n$, $\left|X^{\langle 1,3 \rangle}\right| = 2$, $\left|X^{\langle 1-3,6 \rangle}\right| = 4$.

Since some data may be unknown, in the future we will distinguish between missing and unavailable data, indicated in Figure 1, respectively, missing – with the symbol '?', unavailable – with the symbol '*'.

Missing data are those $x_{k,j}$, $k \in \{i_i, i_2, \dots\}$ whose values are unknown, but it is known that institutions U_k perform (can perform) the activity assessed by the *j-th* indicator. Denote the j*-th* indicator, some of the values of which are missing, by X_j^{miss}.

Unavailable data are those $x_{k,j}$, $k \in \{i_i, i_2, \dots\}$ whose values are unknown, but it is known that institutions U_k, due to their specificity, cannot perform the activity assessed by the j*-th* indicator. Then such institutions are incomparable among themselves and with other institutions by the j*-th* indicator. The set of such institutions is denoted by $U^{inc,j}$, and the j*-th* indicator, some of the values of which are unavailable, by X_j^{inc}. The set of incomparable institutions is denoted by U^{inc}, and comparable institutions by U^{comp}.

Some indicators may contain unknown data, some of which are missing data, and some are unavailable. We will denote such indicators with the superscript mix-X_j^{mix}.

We will denote each j*-th* indicator for which all data are known as X_j^{com} and call it complete.

Figure 1 shows the general data pattern for n indicators, among which two are complete X_1^{com} and X_2^{com}, one with missing data X_3^{miss}, one with unavailable data X_{n-1}^{inc} and one, X_n^{mix}, in which some of the data is unknown, some is unavailable. In the general case $\left|X_j^{com}\right| \geq 0$, $\left|X_j^{miss}\right| \geq 0$, $\left|X_j^{inc}\right| \geq 0$, $\left|X_j^{mix}\right| \geq 0$, i.e. the number of complete indicators, indicators with missing and unavailable data, and indicators with missing and unavailable data can be arbitrary.

We will also say that the data X form the set $M_X = M_X{}^{obs} \cup M_X{}^{miss} \cup M_X{}^{inc}$, where $M_X{}^{obs}, M_X{}^{miss}, M_X{}^{inc}, M_X{}^{imp}$ are subsets for the observed, missing, unavailable and imputed (recovered) data. The set of data that is unknown for any reason will be denoted $M_X{}^{unob}$ (unobserved) with the obvious equality $M_X{}^{unob} = M_X{}^{miss} \cup M_X{}^{inc}$. Let $M_X{}^{comp}$ be the set of data by which the institutions are comparable, i.e. excludir $X = (X_1, X_2, \ldots, X_n)$ in X_n^{inc}.

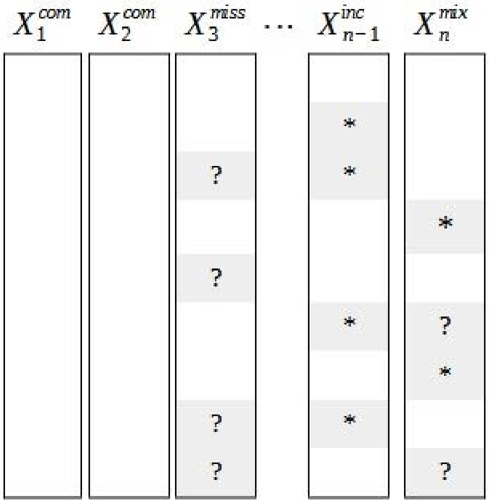

Fig. 1. Data pattern

The task of constructing a rating for N institutions means ordering them according to a certain rule, that is, assigning an integer number (ranking place) to each institution U_i for all n indicators, taking into account the conventional principles and assumptions:

$$Rank : \left((x_{i,1}, x_{i,2}, \ldots, x_{i,n}), P\right) \to \{1, 2, \ldots, m\}, \qquad (2)$$

where $P = \{p_1, p_2, ..., p_r\}$ is a set of conventional principles in the form of predicates.

Note that in the case where all data are known, $M_X{}^{unob} = \emptyset$ (i.e. $M_X{}^{miss} = \emptyset$, $M_X{}^{mix} = \emptyset$ and $M_X{}^{inc} = \emptyset$), the value of the top-level composite indicator I_i is trivially calculated by (1), and the ranking:

$$Rank : \left((x_{i,1}, x_{i,2}, \ldots, x_{i,n}), P\right) = Rank : (I_i, \_) \to \{1, 2, \ldots, m\}, \quad (3)$$

is reduced to ordering m real numbers in descending order by the usual comparison relation > (or ≥) of real numbers with the assignment of a rating of *1* to the largest value of I_i. The use of the underscore symbol instead of P is due to the unnecessary need for a set of predicates in this case.

We will call such a method (rule) of calculating the rating trivial, since its application does not require either data imputation or the use of any specific conventions.

Used principles to construct a rating by aggregating composite indicators. Let a given sample of m institutions U_i, $i \in [1, m]$ whose activity is represented by a vector random variable $X = (X_1, X_2, \ldots, X_n)$ with a distribution density $f(\theta) = (f_1(\theta_1), f_2(\theta_2), \ldots, f_n(\theta_n))$. If $U^{inc} \neq \emptyset$, then indicators with indices, for example, $k = \{j_1, j_2\}$ according to which there are incomparable institutions in the sample, will be considered random variables with a conditional density $f_k(\theta_k)$ of the distribution of only comparable institutions.

To build a rating, in general, there is no need to determine the nature of the distribution and find an unbiased and effective estimate $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$, since the task of studying the general population of institutions is not set.

Instead, to build a rating, the following tasks are solved:

1) restoring M^{unob} using the statistical model M^{obs} built on observed data:

$$\forall\, x_{i,j} \in M_X{}^{unob} \;:\; x_{i,j} = \varphi\left(M_X{}^{obs}, B, P\right), \qquad (4)$$

where φ – statistical model, $B = (\beta_1, \beta_2, \ldots, \beta_p)$ – parameters of the statistical model, P – set of predicates;

2) determining the rating itself based on (2).

Further consideration will be carried out on the basis of the principles, which we will present in the form of a set of predicates $P = \{p_1, p_2, p_3, p_4, p_5\}$:

p_1: vectors X_j are numerical, i.e. each value $x_{i,j}$ is a real number;

p_2: data $M_X{}^{obs}$ satisfy the requirements necessary for the application of statistical data recovery procedures, see e.g. [4, 5];

p_3: data recovery is a deterministic process. That is, if for any two institutions U_r and U_s the observed data coincide, then the recovered data will also coincide:

$$\forall\, k \,:\; x_{r,k} \in M_X{}^{imp} \,\&\, x_{s,k} \in M_X{}^{imp} \quad \text{і}$$

$$\forall\, \pi \,:\; x_{r,\pi} \in M_X{}^{obs} \,\&\, x_{s,k} \in M_X{}^{obs}, \text{ то}$$

$$x_{r,\pi} = x_{s,\pi} \Rightarrow x_{r,k} = x_{s,k}$$

p_4: the weighted sum of each of the incomparable institutions U_i^{inc} should not deteriorate due to the evaluation by the indicators X_j^{inc} and X_j^{mix}:

$$\forall\, U \in U^{inc}\; \forall\, j \;:\; \left(X_j \neq X^{inc} \,\&\, X_j \neq X^{mix}\right)$$

$$\Rightarrow \left(0 \leq Rank\left(\left[X_{imp}{}^{<j>}\right]\right) - Rank(X_{imp}) < \varepsilon\right)$$

where $\varepsilon \geq 0$, $\left[X_{imp}{}^{\langle j \rangle}\right]$ is a matrix with vectors by which all institutions are comparable;

p_5: the set of incomparable institutions does not contain subsets with respect to comparability. In other words, all incomparable institutions are incomparable by the same list of indicators:

$$\forall U_i \in U^{inc} \ \forall j \ : \ X_j = X_j{}^{inc} \vee X_j = X_j{}^{mix} \quad \Rightarrow \ U_i = U_i{}^{inc,j}$$

The first two predicates p_1 and p_2 are non-specific and provide the usual requirements for statistical procedures.

Predicate p_3 – is specific for rating construction, since it makes it impossible for two institutions with identical sets of observed data to receive a higher rating than the other. That is, its purpose is to prevent rating bias. Since the rating is built on composite indicators with data recovery, the predicate is formulated in terms of the imputation process.

The purpose of the specific predicate p_4 – is to compensate for losses in the assessment of incomparable institutions due to the use of indicators $X_j{}^{inc}$ and $X_j{}^{mix}$, which causes incomparability.

Predicate p_5 provides a rational application of p_4, because if $\left(U^{inc} = U^{inc1} \cup U^{inc2}\right) \& \left(U^{inc1} \neq U^{inc2}\right)$, then compensation for losses ceases to be obvious and justified, or even possible.

Classification of ranking methods based on data patterns. As noted above, determining the rating involves restoration (imputation), the determination of the rating involves restoring (imputation) the $M_X{}^{unob}$ data by (4) and calculating the rating according to (2). Since the restoration (imputation) of the data depends significantly on the data patterns, the methods for calculating the rating can also be classified according to the pattern corresponding to the data X.

Table 1 lists the data templates and names of methods that are proposed and suitable for use.

As already mentioned, in the case where all institutions are comparable ($M_X{}^{inc} = \oslash$, $M_X{}^{mix} = \oslash$) and there is no missing data ($M_X{}^{miss} = \oslash$), the values of the top-level composite indicator I_i are trivially calculated by (1), and the ranking by (3). Therefore, we call this method "Trivial", see No. 1 in Table 1. Its application does not require either data imputation or the use of specific conventions p_3, p_4 and p_5.

In the case when all institutions are comparable, but there is missing data, the "Multiple Regression" or "Multiple Imputation" methods are used, see methods No. 2 and No. 3 in Table 1, depending on the number of indicators with missing data.

Building a rating by method No. 2 and all subsequent ones, since they involve restoration, requires checking the truth of the predicate p_3.

If there are incomparable institutions among the institutions, but there is no missing data, then methods No. 4-6 are used, depending on the number of indicators $X_j{}^{inc}$ and the accepted value in the predicate p_4.

If the number of indicators for which there are incomparable institutions is more than one, then requirement p_5 must be met.

The "Ordinal Statistics" method guarantees, according to the predicate p_4, that the rating of non-comparable institutions can remain unchanged, or increase after taking into account the indicator $X_j{}^{inc}$. In contrast, the "Merge Partial Ratings" method guarantees that taking into account the indicator $X_j{}^{inc}$ will not change the rating of non-comparable institutions.

The name of the "Inc-Reduction" method is used if the number of indicators $X_j{}^{inc}$ is more than one and indicates that it can be reduced to methods No. 4 or No. 5.

Similarly, the "Mix1-Reduction" and "Mix-Reduction" methods are used if all missing and unavailable data are in the indicators $X_j{}^{mix}$. As the name implies, they can be reduced to the previously mentioned methods.

Table 1

Classification of rating methods

№	Data pattern			Method
1	$M^{unob} = \oslash$			Trivial
2	$M^{miss} \neq \oslash M^{inc} = \oslash M^{mix} = \oslash$	$\lvert X_j{}^{inc} \rvert = 1$		Multiple Regression
3		$\lvert X_j{}^{miss} \rvert > 1$		Multiple Imputation
4	$M^{miss} = \oslash \ M^{inc} \neq \oslash M^{mix} = \oslash$	$\lvert X_j{}^{inc} \rvert > 1$	$\varepsilon \geq 0$	Ordinal Statistics
5			$\varepsilon = 0$	Merging Partial Ratings
6		$\lvert X_j{}^{inc} \rvert > 1$		Inc-Reduction
7	$M^{miss} = \oslash M^{inc} = \oslash M^{mix} \neq \oslash$	$\lvert X_j{}^{mix} \rvert = 1$		Mix1-Reduction
8		$\lvert X_j{}^{mix} \rvert > 1$		Mix-Reduction
9	$M^{miss} \neq \oslash M^{inc} \neq \oslash M^{mix} \neq \oslash$			Gen-Reduction

Finally, the "Gen-Reduction" method provides for the construction of a rating in the most general case by sequential reduction to the already mentioned methods.

Conclusions. The article is devoted to a detailed analysis and solution of the problem of assessing the level of internationalization of scientific institutions in situations where primary statistical data are incomplete or partially unavailable.

The proposed set of concepts is sufficient for a formal representation of the problem of constructing a rating by component indicators in the presence of unknown data.

Differences in solving data recovery problems are identified for the cases of: a) constructing a rating; b) studying the general population using sample data.

Requirements for rating construction methods in the form of predicates are formulated.

A classification of rating construction methods based on a data template is carried out.

The results obtained in total will allow creating an effective software architecture that will help solve the problem of assessing the level of internationalization of scientific institutions in situations where primary statistical data are incomplete or partially unavailable.

Bibliography:

1. Brandenburg, U., and G. Federkeil. How to Measure Internationality and Internationalisation of Higher Education Institutions. Indicators and Key Figures. 2007. URL: https://www.che.de/en/download/how_to_measure_internationality_ap_92-pdf.

2. Bas M.C., Boquera M., Carot. Measuring internationalization performance of higher education institutions through composite indicators, *INTED2017 Proceedings*, 2017, pp. 3149–3156. DOI: 10.21125/inted.2017.0815.

3. Nardo, M. et al. Handbook on Constructing Composite Indicators: Methodology and User Guide. *OECD Statistics Working Papers*, No. 2005/03, OECD Publishing, Paris. 2005. DOI: 10.1787/533411815016.

4. RUBIN, Donald B. Inference and missing data. *Biometrika*, Vol. 63 Issue 3. 1976. pp. 581-592. URL: http://qwone.com/~jason/trg/papers/rubin-missing-76.pdf.

5. Enders, Craig K. Applied missing data analysis. *Guilford Publications.* 2022. P. 401.

6. ShanghaiRanking's Academic Ranking of World Universities. Methodology 2024. URL: https://www.shanghairanking.com/methodology/arwu/2024.

7. Knight J. Monitoring the Quality and Progress of Internationalization. *Journal of Studies in International Education*, 5(3), 2001. pp. 228–243. DOI: 10.1177/102831530153004.

8. Knight, Jane. Internationalization Remodeled: Definition, Approaches, and Rationales. *Journal of Studies in International Education*. Vol 8. 2004. pp. 5–31. DOI: 10.1177/1028315303260832.

9. Statyvka Y. I., Nedashkivskiy O. L., Mingjun Z. Model of the process for evaluating the level of internationalization of the scientific institution activities. *Connectivity*, No. 3 (175). 2025. pp. 42–51. DOI: 10.31673/2412-9070.2025.020915,

10. Wedyan F., Abufakher S. Impact of design patterns on software quality: a systematic literature review. *IET Software.* Vol. 14, Issue 1. 2019. pp.1–17. DOI: 10.1049/iet-sen.2018.5446.

Стативка Ю.І., Недашківський О.Л., Мінцзюнь Ч. ОЦІНКА РІВНЯ ІНТЕРНАЦІОНАЛІЗАЦІЇ НАУКОВИХ УСТАНОВ ЗА НАЯВНОСТІ ВІДСУТНІХ ДАНИХ

Стаття присвячена детальному аналізу та вирішенню проблеми оцінювання рівня інтернаціоналізації наукових установ у ситуаціях, коли первинні статистичні дані є неповними або частково недоступними, що суттєво ускладнює побудову об'єктивних і коректних рейтингів. Рівень інтернаціоналізації визначається як складний багатовимірний, композитний, показник, що формується за допомогою ієрархічної агрегації численних індикаторів, обчислення якого в умовах, коли дані не повні, має залишатись логічним і вмотивованим. Автори пропонують чітке розмежування двох типів невідомої інформації: відсутні дані (missing), коли показник може бути визначений у принципі, та недоступні (unavailable), коли через специфіку діяльності об'єкт не підлягає порівнянню з іншими за певним індикатором (одним чи більше). Запроваджено формалізований апарат опису структури даних з використанням спеціальних позначень для повних індикаторів, індикаторів з відсутніми значеннями, з недоступними значеннями та змішаних випадків. Сформульовано систему вимог у формі предикатів, що гарантують коректність відновлення пропусків, виключення можливості зміщення рейтингу та збереження або покращення позицій непорівнюваних установ після врахування показників – джерел непорівнянності. Аналіз можливих випадків неповноти вихідних (source) даних дозволив виявити дев'ять релевантних завданню рейтингування патернів, які стали основою класифікації завдань та методів побудови рейтингу, що охоплює як випадки повних даних ("Trivial"), так і ситуації з пропущеними ("Multiple Regression", "Multiple Imputation"), з непорівнюваними об'єктами ("Ordinal Statistics", "Merge Partial

Ratings”), а також узагальнені методи поетапного зведення задачі до простіших (“Inc-Reduction”, “Mix1-Reduction”, “Mix-Reduction”, “Gen-Reduction”). У підсумку зроблено висновок про достатність розробленого понятійного апарату для формалізації задачі рейтингування в умовах неповноти даних, визначено особливості процедур відновлення даних при вирішенні завдань рейтингування та їх відмінність від аналогічних завдань при вивченні генеральної сукупності, сформульовано чіткі вимоги до методів у формі предикатів і наведено системну класифікацію методів.

Ключові слова: *інтернаціоналізація наукових установ, оцінка рівня інтернаціоналізації, структура даних, відсутні та недоступні дані, композитні показники, архітектура програмної системи, програмна інженерія.*

Дата надходження статті: 18.09.2025
Дата прийняття статті: 29.09.2025
Опубліковано: 16.12.2025