УДК 004.272.43+004.439R DOI: 10.31673/2412-4338.2024.039411

Майборода Максим Володимирович
Державний університет інформаційно-комунікаційних технологій, м. Київ
ORCID 0009-0006-7467-2426

Бажан Тетяна Олександрівна
Державний університет інформаційно-комунікаційних технологій, м. Київ
ORCID 0009-0007-6594-1695

Жебка Сергій Валентинович
Державний університет інформаційно-комунікаційних технологій, Київ
ORCID 0009-0007-4620-9888

Колодюк Андрій Васильович
Державний університет інформаційно-комунікаційних технологій, м. Київ
ORCID 0009-0001-1724-7531

МЕТОДИ МЕТРИЧНОГО ТА НЕМЕТРИЧНОГО БАГАТОВИМІРНОГО ШКАЛЮВАННЯ ДЛЯ ДОВІЛЬНОЇ МАТРИЦІ ДАНИХ В МОВІ R

***Анотація**. У статті розглядаються два підходи до багатовимірного шкалювання — метричне та неметричне — в контексті їхньої реалізації на мові програмування R для аналізу довільних матриць даних. Багатовимірне шкалювання (MDS) є потужним інструментом для візуалізації та інтерпретації складних багатовимірних даних. В сучасних умовах, з урахуванням стрімкого зростання обсягів інформації, ефективні методи аналізу великих даних набувають все більшої актуальності. Метою цієї роботи є порівняння ефективності та точності метричних та неметричних підходів до MDS, визначення їхніх переваг і недоліків, а також надання практичних рекомендацій щодо їх використання для вирішення різноманітних завдань в різних сферах, таких як соціологія, маркетинг, політологія, психологія тощо.*

У статті проведено огляд теоретичних основ багатовимірного шкалювання, описано алгоритми, що лежать в основі реалізації MDS в R, а також проаналізовано особливості застосування кожного методу. Метричний підхід до MDS ґрунтується на припущенні про лінійну залежність між відстанями у вихідних даних і результатами шкалювання, що дозволяє отримувати точні результати для структурованих даних. Неметричний підхід, навпаки, є гнучкішим і здатен працювати з більш абстрактними даними, зокрема у випадках, коли відстані між об'єктами важко піддаються чіткому кількісному опису.

Для оцінки ефективності обох методів було розроблено приклади з використанням реальних і симуляційних даних, на яких було показано, як обидва підходи поводять себе в різних ситуаціях. Метричне шкалювання продемонструвало кращі результати при роботі з даними, що відповідають припущенням про лінійність, тоді як неметричне MDS виявилося більш адаптивним для даних із нелінійними зв'язками. Результати експериментів показали, що для даних з різною структурою і розміром обидва підходи можуть бути корисними, але вибір методу залежить від конкретних вимог до аналізу.

Важливим результатом роботи є розробка набору рекомендацій для вибору методу багатовимірного шкалювання в залежності від типу даних, що аналізуються. Наприклад, для чітко структурованих даних, таких як географічні або демографічні, перевагу слід надавати метричному підходу, тоді як для більш складних та неструктурованих даних, як у психологічних дослідженнях, більш доцільним буде застосування неметричного шкалювання.

У статті також надано приклади коду на мові R для реалізації обох підходів, що можуть бути використані для подальших досліджень та практичної роботи з багатовимірним шкалюванням. Розглянуті приклади демонструють, як обидва методи можуть бути інтегровані у процес аналізу даних для виявлення скритих закономірностей та побудови візуальних моделей на основі

багатовимірних даних. Окрім цього, автори вказують на перспективи подальших досліджень у галузі застосування багатовимірного шкалювання в аналізі великих даних та розробки нових методологій для обробки гетерогенних інформаційних масивів.

Таким чином, стаття робить вагомий внесок у розвиток сучасних підходів до аналізу даних в мовах програмування та відкриває нові перспективи для застосування багатовимірного шкалювання у різних сферах науки і бізнесу. Пропоновані рекомендації щодо вибору методу шкалювання можуть бути корисними для дослідників та практиків, що працюють з великими обсягами даних та прагнуть використовувати новітні методи для їх аналізу.

Ключові слова: *багатовимірне шкалювання, метричне MDS, неметричне MDS, аналіз даних, R, великі дані, візуалізація, евклідова метрика, статистична обробка, Big Data.*

Maiboroda Maksym
State university of information and communication technologies, Kyiv
ORCID 0009-0006-7467-2426

Bazhan Tetiana
State university of information and communication technologies, Kyiv
ORCID 0009-0007-6594-1695

Zhebka Serhii
State university of information and communication technologies, Kyiv
ORCID 0009-0007-4620-9888

Kolodiuk Andrii
State university of information and communication technologies, Kyiv
ORCID 0009-0001-1724-7531

METHODS OF METRIC AND NON-METRIC MULTIDIMENSIONAL SCALING FOR ARBITRARY DATA MATRICES IN R

Abstract*. This article explores two approaches to multidimensional scaling (MDS)—metric and non-metric—in the context of their implementation in the R programming language for analyzing arbitrary data matrices. Multidimensional scaling is a powerful tool for visualizing and interpreting complex multidimensional data. Given the rapid growth in information volumes, efficient methods for analyzing big data have become increasingly relevant. The aim of this study is to compare the effectiveness and accuracy of metric and non-metric MDS methods, identify their advantages and disadvantages, and provide practical recommendations for their use in solving various tasks across multiple fields, such as sociology, marketing, political science, psychology, and more.*

The article provides an overview of the theoretical foundations of multidimensional scaling, describes the algorithms underlying MDS implementation in R, and analyzes the specifics of applying each method. The metric MDS approach is based on the assumption of a linear relationship between distances in the input data and the scaling results, allowing for precise results when working with structured data. The non-metric approach, on the other hand, is more flexible and can handle more abstract data, particularly in cases where the distances between objects are not easily quantifiable.

To evaluate the effectiveness of both methods, examples using real and simulated data were developed, demonstrating how each approach behaves in different situations. Metric scaling performed better when working with data that adheres to linear assumptions, while non-metric MDS proved to be more adaptive for data with non-linear relationships. Experimental results show that both approaches can be useful for data with different structures and sizes, but the choice of method depends on the specific requirements of the analysis.

A significant outcome of this study is the development of a set of recommendations for choosing an MDS method depending on the type of data being analyzed. For example, for well-structured data, such as geographic or demographic data, the metric approach is preferable, whereas for more complex and unstructured data, as in psychological studies, non-metric scaling is more suitable.

The article also provides R code examples for implementing both approaches, which can be used for further research and practical work with multidimensional scaling. The presented examples demonstrate how both methods can be integrated into the data analysis process to uncover hidden patterns and build visual models based on multidimensional data. Additionally, the authors highlight prospects for future research in applying multidimensional scaling to big data analysis and developing new methodologies for processing heterogeneous information arrays.

Thus, this article makes a significant contribution to the development of modern approaches to data analysis in programming languages and opens new perspectives for the application of multidimensional scaling in various scientific and business fields. The proposed recommendations for choosing a scaling method can be useful for researchers and practitioners working with large data volumes and aiming to employ the latest methods for their analysis.

***Keywords**: multidimensional scaling, metric MDS, non-metric MDS, data analysis, R, big data, visualization, Euclidean metric, statistical processing, Big Data.*

Постановка проблеми. Завдяки стрімкому розвитку обчислювальної техніки стрімко зросла потреба в обробці великих масивів даних для вирішення поставлених задач у різноманітних сферах діяльності. Часто в дослідах ми зацікавленні не тільки в порівнянні одновимірних дескрипторів спільнот, як наприклад смертність, а й у виявленні, як смертність змінюється від однієї спільноти до іншої і виявлені факторів впливу. В таких ситуаціях для систематизації і аналізу певних даних спеціалісти прибігають до сукупності методів, що називаються **багатовимірне шкалювання.** В цій роботі будуть розглянуті та реалізовані два підходи до багатовимірного шкалювання – метричне та неметричне. Реалізація буде проведена за допомогою мови для статистичної обробки даних R. На основі отриманих даних буде дано оцінку обом методам, а також дано висновок, який з методів доцільніше використовувати при роботі з мовою R.

Аналіз останніх досліджень і публікацій. Багатовимірне шкалювання дозволяє представити великий масив даних про відмінності об’єктів в доступному для інтерпретації вигляді. Методи багатовимірного шкалювання використовуються на практиці для дослідження складних явищ і процесів, що не піддаються опису чи моделюванню а також прямому трактуванню і інтерпретації. Гарним прикладом області використання багатовимірного шкалювання є психологія, а саме, наприклад, оцінка суб’єктивного сприйняття людиною емоцій в залежності від виховання, соціуму, соціально-економічного фактору, релігії, місцевості проживання тощо. Також БШ виявилось корисним в таких областях як антропологія, географія, історія, педагогіка. Дуже широкого застосування багатовимірне шкалювання зазнало в маркетингу, адже ефективність реклами прямо впливає на дохід підприємства, а тому оцінка і виявлення факторів що шкодять її ефективності є обов’язковим. Як факторний та кластерний аналіз, БШ використовується для опису структури даних. Вхідні припущення багатовимірного шкалювання відрізняються від вхідних припущень факторного та кластерного аналізу, а отже, опис і інтерпретація оброблених даних буде відрізнятися. Ще одною областю використання БШ можна назвати соціологію та політологію. Оцінка сприйняття одного або іншого кандидату, партії, вивчення політичних нахилів населення, тощо є однією з основних задач цих наук.

В широкому сенсі, багатовимірне шкалювання є одним із інструментів такого соціально-економічного феномену як Big data analysis.

БШ можна розділити на два напрямки: метричне і неметричне.

Мета і задачі дослідження. Метою даної статті є дослідження та порівняння методів метричного та неметричного багатовимірного шкалювання (MDS) для аналізу довільної матриці даних в мові програмування R. Вивчення особливостей застосування кожного методу, їхніх переваг та недоліків, а також надання практичних рекомендацій щодо їх використання для аналізу реальних даних.

Завдання

1. Огляд теоретичних основ багатовимірного шкалювання, включаючи метричний та неметричний підходи.
2. Опис основних алгоритмів та методів реалізації багатовимірного шкалювання в мові R.
3. Порівняння різних підходів до багатовимірного шкалювання на прикладах реальних та симуляційних даних.
4. Аналіз ефективності та точності метричних та неметричних методів у контексті різних типів даних.
5. Надання рекомендацій щодо вибору методу багатовимірного шкалювання залежно від специфіки та структури даних.
6. Розробка та презентація прикладів коду на R для реалізації метричного та неметричного MDS.
7. Обговорення потенційних напрямків для подальших досліджень у сфері багатовимірного шкалювання в мові R.

Результати дослідження.

1)Реалізація метричного шкалювання

1. Створення матриці даних.

Для реалізації двох методів багатовимірного шкалювання , була задана випадкова матриця data.matrix розмірністю 100x10.

Наступним кроком ми перетворюємо задану матрицю на матрицю відстаней, транспонуємо, центруємо, масштабуємо та задаємо метрику. Виконується ця операція за допомоги функції dist().

```
distance.matrix <- dist(scale(t(data.matrix), center=TRUE, scale=TRUE),
                        method="euclidean")
```

(t(data.matrix), -транспонування матриці.

(scale(t(data.matrix), center=TRUE, scale=TRUE) – ми центруємо та масштабуємо значення для кожного рядка, що тепер став стовпцем.

method="euclidean") – ми «кажемо» функції dist(), що ми хочемо утворити матрицю використовуючи евклідову метрику.

2. Реалізація метричного шкалювання.

В мові R метричне шкалювання реалізовано в декількох пакетах даних, тут буде використовуватися пакет (stats). В пакеті stats метричне шкалювання запропоновано у вигляді команди cmdscale (Classical Multi-Dimensional Scaling).

Використані в функції cmd.scale аргументи:

Eig=TRUE – ми кажемо cmdscale(), щоб вона повертала власні значення. Ми використовуємо це для обчислення кількості варіацій на кожну вісь в матриці відстаней.

x.ret=TRUE – ми кажемо cmdscale() щоб вона повернула двічі центровану (рядки і стовпці центровані) версію матриці відстаней.

Загальний вигляд функції

```
cmdscale(d, k=2, eig=FALSE, add=FALSE, x.ret=FALSE,
            list.=eig || add ||x.ret)
```

Детальніше з кожним аргументом функції можна ознайомитись на офіційному ресурсі мови R.

3. Вивід даних після метричного шкалювання.

Для виводу даних після метричного шкалювання було використано функцію “round”. Round розраховує кількість варіацій кожної вісі. В подальшому це можна використовувати для побудови графіку за допомоги функції ggplot(). За допомогою команди data.frame формуємо дата фрейм отриманої матриці. Виводимо дані фрейма на екран.

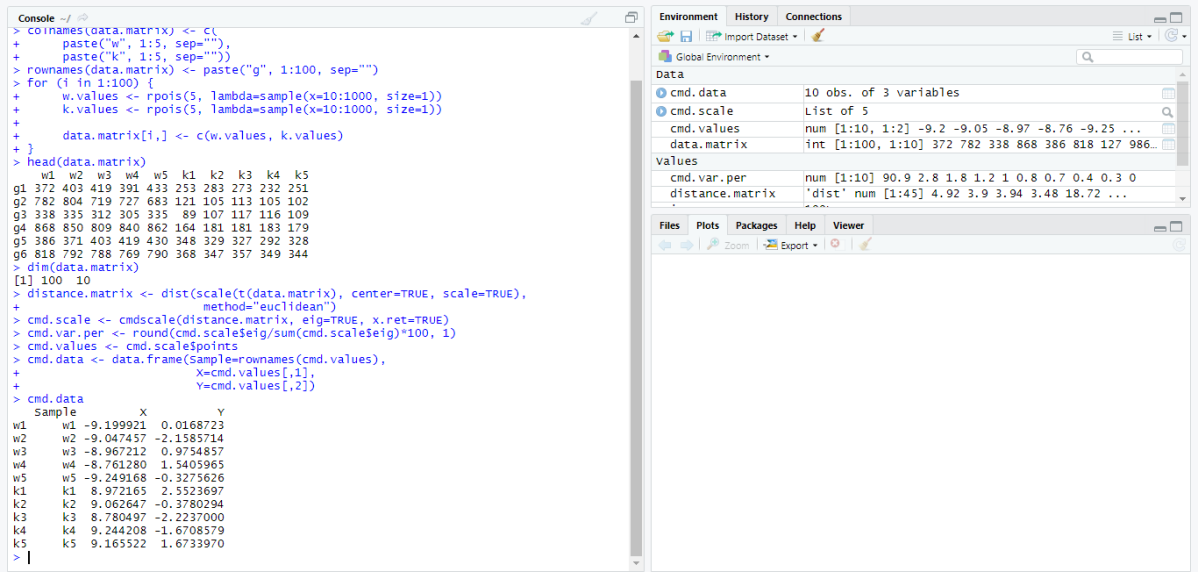

Рис. 1. Вивід даних після метричного шкалювання

2)Реалізація неметричного шкалювання.

В R неметричне шкалювання реалізовано у вигляді функції isoMDS() з пакету {MASS}.

Загальний вигляд функції:

isoMDS(d, y = cmdscale(d, k), k = 2, maxit = 50, trace = TRUE,

tol = 1e-3, p = 2)

Бачимо, що isoMDS використовує cmdscale для розрахунків. З цього можна зробити висновок що неметричне шкалювання використовує метричне як проміжкове обчислення.

d – матриця відмінностей, відношення або відстаней.

k – максимальна розмірність простору, в котрому мають бути представлені дані, це значення предається в cmdscale для розрухунків.

Опираючись на ті самі дані що були отримані в п.1. «Створення матриці даних» реалізуємо неметричне шкалювання.

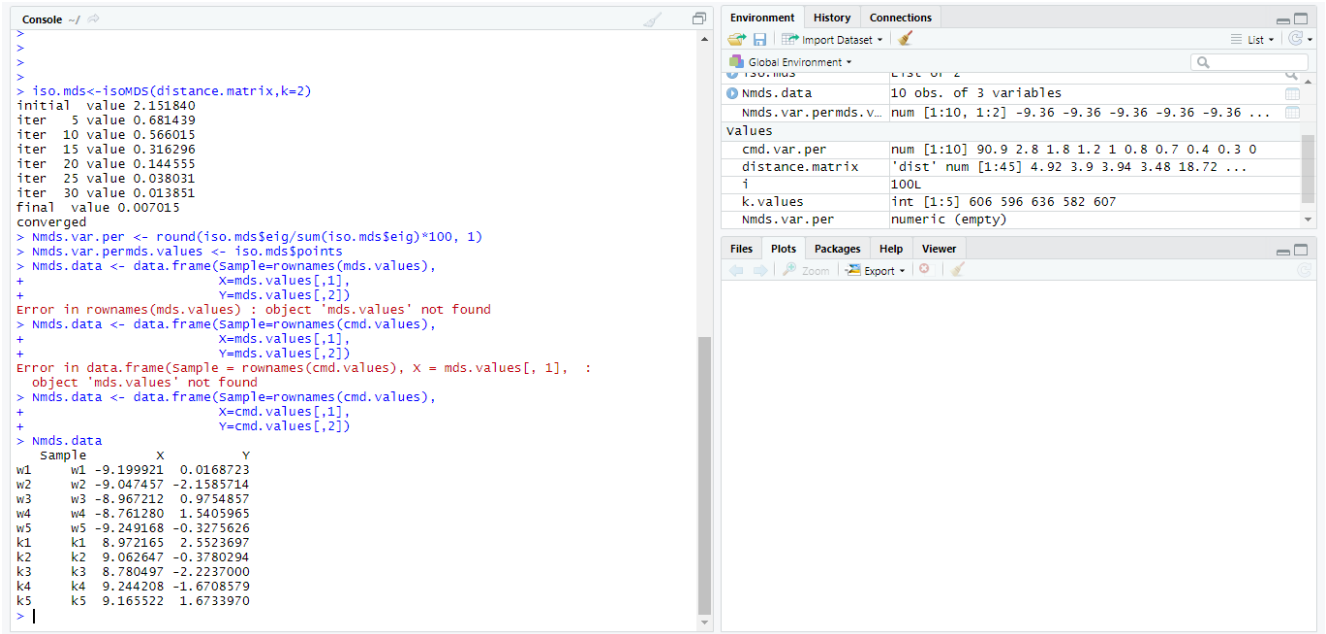

Рис. 2. Створення матриці даних для неметричного шкалювання

2)Порівняння вихідних даних

Як бачимо, дата фрейми для метричного і неметричного шкалювання співпадають по всіх координатах.

3)Висновок:

Методи метричного і неметричного шкалювання для довільної матриці даних не дають ніяких розбіжностей.

Метричне і неметричне багатомірне шкалювання в R з використанням методу Терстоуна

Для порівняння метричного і неметричного шкалювання було взято матрицю що показує частку суб'єктів, що віддали перевагу стимул-стовпець стимулу-рядку з книги Multidimentional scaling за авторством Mark L. Davidson.

Таблиця 1

Матриця для порівняння метричного і неметричного шкалювання

Овочі	**Матриця P**								
	Турнепс	**Капуста**	**Свекла**	**Спаржа**	**Морква**	**Шпинат**	**Фасоль**	**Боби**	**Зерно**
Турнепс	—	0,818	0,770	0,811	0,878	0,892	0,899	0,892	0,926
Капуста	0,182	—	0,601	0,723	0,743	0,736	0,811	0,845	0,850
Свекла	0,230	0,399	—	0,561	0,736	0,676	0,845	0,797	0,818
Спаржа	0,189	0,277	0,439	—	0,561	0,588	0,620	0,709	0,764
Морква	0,122	0,257	0,264	0,439	—	0,493	0,574	0,709	0,764
Шпинат	0,108	0,264	0,324	0,412	0,507	—	0,628	0,526	0,627
Фасоль	0,108	0,155	0,203	0,334	0,426	0,372	—	0,527	0,473
Боби	0,108	0,155	0,203	0,334	0,291	0,318	0,473	—	0,527
Зерно	0,074	0,142	0,182	0,270	0,236	0,372	0,358	0,372	—

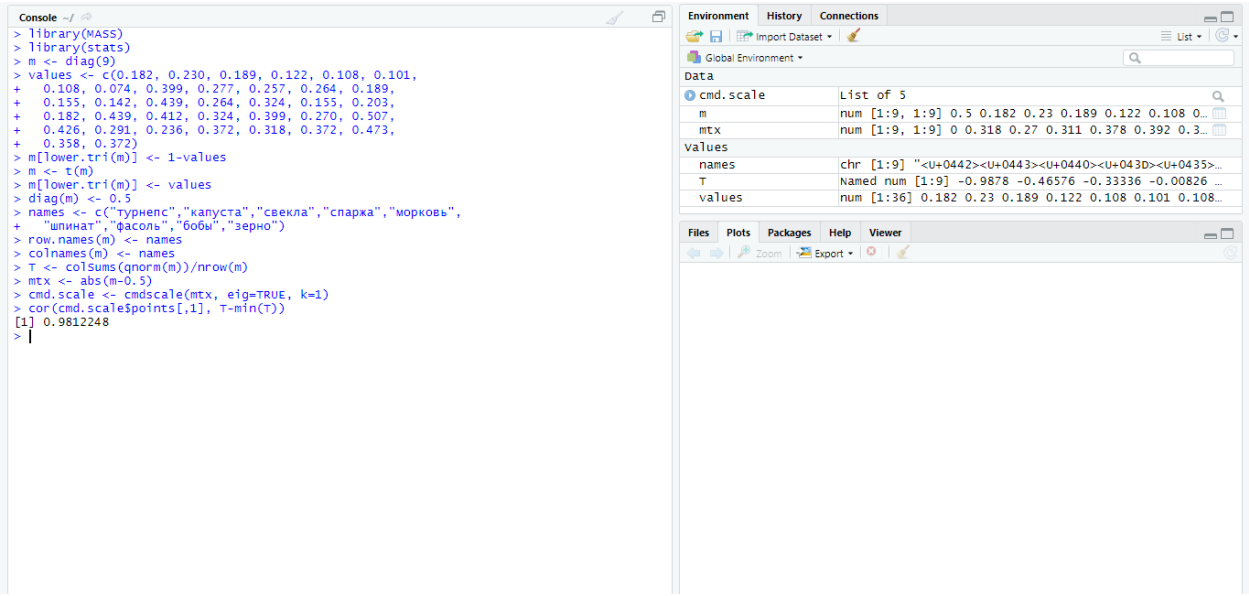

Рис. 3. Обчислення кореляції між шкальними оцінками для метричного шкалювання

1)Реалізація

Будуємо матрицю даних.

Використовуємо метод Терстоуна, в R цей метод можна реалізувати функцією colSums.

Перетворюємо матрицю часток на матрицю відстаней командою abs().

Для метричного шкалювання:

Проводимо метричне шкалювання.

Обчислюємо кореляцію між шкальними оцінками, для метричного шкалювання за допомоги функції kor(). Кореляція становить 0.9812248.

Для неметричного шкалювання:

Проводимо неметричне шкалювання. Обчислюємо кореляцію між шкальними оцінками. Для неметричного шкалювання становить 0.995967.

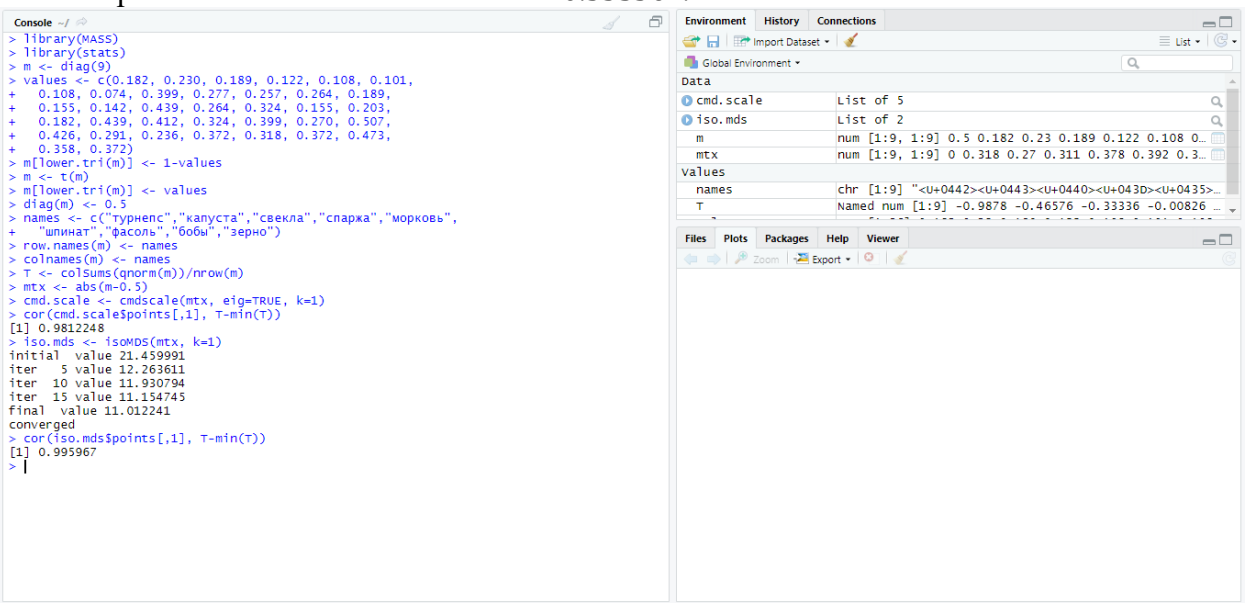

Рис. 4. Обчислення кореляції між шкальними оцінками для неметричного шкалювання

2)Порівняння

Бачимо, що коефіцієнт кореляції для метричного шкалювання менший за коефіцієнт кореляції для неметричного шкалювання.

3)Висновок

Більш оптимальним є неметричне шкалювання, адже коефіцієнт кореляції для неметричного шкалювання вищій, а отже похибка менша.

Висновки та перспективи подальших досліджень. В процесі аналізу R документації, а також інформації з вільних джерел був виявлений пакет (vegan), в якому неметричне шкалювання реалізовано у вигляді функції metaMDS. При більш детальному вивченні пакету, було виявлено ряд функцій, що значно полегшують процес неметричного шкалювання. Наприклад функція decostand(), використовується для трансформації та стандартизації даних. Також в пакеті vegan є функція rankindex(), що визначає найкращу метрику відстані для вхідного набору даних.

З огляду на те, що неметричне шкалювання в R реалізовано з використанням cmdscale , тобто метричного шкалювання, а також на основі порівняння шкальних оцінок з другого прикладу було зроблено висновок, що неметричне шкалювання є більш коректним і точним з точки зору вихідних даних.

Як було вказано в «Ідеї неметричнго шкаювання», для неметричного методу є істотним саме порядок оцінок близькості найчастіше дані надаються в порядковому вигляді, а не

інтервальному. Метричне шкалювання адаптовано до інтервальних даних, неметричне – до інтервальних і порядкових.

На основі всієї інформації що була проаналізована, було зроблено висновок, що неметричне багатовимірне шкалювання є більш уніфікованим методом, а тому рекомендується використовувати саме цей метод.

Список використаних джерел

1. Borg, I., Groenen, P. J. F. "Modern Multidimensional Scaling: Theory and Applications." Springer, 2005
2. Cox, T. F., Cox, M. A. A. "Multidimensional Scaling." Chapman and Hall/CRC, 2001
3. Kruskal, J. B. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." Psychometrika, 1964
4. Torgerson, W. S. "Multidimensional scaling: I. Theory and method." Psychometrika, 1952
5. Venables, W. N., Ripley, B. D. "Modern Applied Statistics with S." Springer, 2002
6. Офіційна документація пакету MASS в R: https://cran.r-project.org/web/packages/MASS/index.html
7. Офіційна документація пакету smacof в R: https://cran.r-project.org/web/packages/smacof/index.html
8. Статті на сайті R-bloggers про багатовимірне шкалювання: https://www.r-bloggers.com/
9. Zhebka V., Skladannyi P., Bazak Y., Bondarchuk A., Storchak K. Methods for Predicting Failures in a Smart Home / CEUR Workshop Proceedings, 2024, 3665, p. 70–78
10. Malinov V., Zhebka V., Kokhan I., Storchak K., Dovzhenko T. Cryptocurrency as a Tool for Attracting Investment and Ensuring the Strategic Development of the Bioenergy Potential of Processing Enterprises in Ukraine / Lecture Notes on Data Engineering and Communications Technologies, 2024, 195, p. 387–405

Reference

1. Borg, I., Groenen, P. J. F. "Modern Multidimensional Scaling: Theory and Applications." Springer, 2005
2. Cox, T. F., Cox, M. A. A. "Multidimensional Scaling." Chapman and Hall/CRC, 2001
3. Kruskal, J. B. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." Psychometrika, 1964
4. Torgerson, W. S. "Multidimensional scaling: I. Theory and method." Psychometrika, 1952
5. Venables, W. N., Ripley, B. D. "Modern Applied Statistics with S." Springer, 2002
6. Official documentation of the MASS package in R: https://cran.r-project.org/web/packages/MASS/index.html
7. Official documentation of the smacof package in R: https://cran.r-project.org/web/packages/smacof/index.html
8. Articles on the R-bloggers website about multidimensional scaling: https://www.r-bloggers.com/
9. Zhebka V., Skladannyi P., Bazak Y., Bondarchuk A., Storchak K. Methods for Predicting Failures in a Smart Home / CEUR Workshop Proceedings, 2024, 3665, p. 70–78
10. Malinov V., Zhebka V., Kokhan I., Storchak K., Dovzhenko T. Cryptocurrency as a Tool for Attracting Investment and Ensuring the Strategic Development of the Bioenergy Potential of Processing Enterprises in Ukraine / Lecture Notes on Data Engineering and Communications Technologies, 2024, 195, p. 387–405