

И. Е. Тюгашев

Институт археологии и этнографии СО РАН
Новосибирск, Россия
E-mail: tgshgr@yandex.ru

Парадокс Орлоци в статистическом анализе археологических данных

В статье рассматривается проблема применения евклидовой метрики в многомерном статистическом анализе археологических данных, способная приводить к искажениям, получившим название «парадокс Орлоци». Парадокс Орлоци может проявляться двояко. В первом случае два наблюдения, не имеющие ни одного общего вида или признака, могут казаться более сходными, чем два других наблюдения, у которых списки признаков полностью идентичны. Во втором случае наблюдения со сходной внутренней структурой, но различающиеся по размеру, могут статистически выглядеть более удаленными друг от друга, чем наблюдения примерно одинакового размера, но различающиеся по составу признаков. Вероятность возникновения подобных искажений в археологических исследованиях высока, поскольку археологические выборки часто отличаются по численности, содержат нулевые значения, выбросы и несоразмерные комплексы. На примере гипотетической выборки трех групп памятников, различных по составу, демонстрируется, что использование евклидовой метрики в неметрическом многомерном шкалировании приводит к группировке преимущественно по суммарному количеству артефактов. Последующая проверка различий между группами с помощью теста ANOSIM, рассчитанного на основе метрики Евклида, также не выявляет их статистической обособленности. Напротив, применение метрики хорды позволяет получить ординацию, отражающую различия трех смоделированных групп, а результаты теста ANOSIM подтверждают статистическую значимость различий между группами. Предложено три способа избежать парадокса Орлоци: преобразование исходных данных, использование иных метрик и применение методов, ориентированных на анализ структуры распределения признаков.

Ключевые слова: археология, парадокс Орлоци, многомерная статистика, метрика Евклида, метрика хорды, NMDS, ANOSIM.

I. E. Tyugashev

Institute of Archaeology and Ethnography SB RAS
Novosibirsk, Russia
E-mail: tgshgr@yandex.ru

The Orlóci Paradox in the Statistical Analysis of Archaeological Data

This article addresses the problem of applying the Euclidean distance in multivariate statistical analysis of archaeological data, which may lead to distortions known as the “Orlóci paradox.” The Orlóci paradox may manifest itself in two ways. In the first case, two observations that share no species or attributes may appear more similar than two other observations whose lists of attributes are identical. In the second case, observations with a similar internal structure but differing in size may statistically appear more distant from one another than observations of approximately the same size but differing in their attribute composition. The likelihood of such distortions occurring in archaeological research is high, since archaeological samples often differ in size and contain zero values, outliers, and assemblages of unequal scale. Using a hypothetical sample of three groups of sites differing in composition, the article demonstrates that the use of the Euclidean distance in non-metric multidimensional scaling leads to a grouping primarily determined by the total number of artifacts. A subsequent test of between-group differences using ANOSIM based on the Euclidean distance likewise fails to reveal their statistical separation. By contrast, the application of the chord distance produces an ordination that reflects the differences among the three simulated groups, while the results of ANOSIM confirm the statistical significance of the between-group differences. Three ways of avoiding the Orlóci paradox are proposed: transformation of the original data, use of alternative metrics, and application of methods focused on analyzing the structure of attribute distributions.

Keywords: archaeology; Orlóci’s paradox; multivariate statistics; Euclidean distance; chord distance; NMDS; ANOSIM.

Таблица 1. Гипотетические данные классического примера, иллюстрирующего парадокс Орлоци

Наблюдения/Виды	A	B	C
O1	1	0	0
O2	0	1	1
O3	0	4	4

В археологии для сопоставления и анализа массива сложных археологических объектов, характеризующихся большим количеством признаков, используются методы многомерной статистики. Эти методы, как правило, основаны на мерах расстояния, позволяющих количественно оценивать степень сходства или различия между наблюдениями. В большинстве случаев в качестве такой меры применяется евклидово расстояние, что обусловлено сложившимися практиками использования соответствующих статистических процедур и стандартными протоколами анализа. Кроме того, ряд методов концептуально и математически основан именно на евклидовом пространстве (как, например, метод главных компонент), вследствие чего использование иных мер расстояния в их рамках оказывается невозможным. В результате метрика Евклида нередко используется исследователями по умолчанию – как стандартный и программно доступный вариант, не всегда сопровождаемый обсуждением ее применимости к специфике конкретных археологических данных [Baxter, 1994, p. 156–157; Tratebas, 2012; Peterson, Drennan, Bartel, 2016; Тюгашев, 2023a].

Однако применение евклидовой метрики к археологическим данным может быть сопряжено с существенными рисками, обусловленными спецификой самого археологического материала. Археологические выборки часто значительно различаются по объему, содержат выбросы, представленные аномально крупными или, напротив, крайне малочисленными комплексами и нередко характеризуются высокой долей нулевых значений [Тюгашев, 2023б; Колобова и др., 2024, 2026].

В подобных условиях различия между наблюдениями могут определяться преимущественно абсолютным количеством объектов, а не их структурным составом. Это создает риск искажения результатов статистического анализа и затрудняет интерпретацию выявленных закономерностей – реальные группы могут разрываться и создаваться новые кластеры.

В археологических исследованиях данная проблема получает сравнительно ограниченное внимание [Cardillo, Alberti, 2013; Carlson, 2017, p. 299–300; Gehlen et al., 2020], особенно в русскоязычных работах [Тюгашев, 2023б; Колобова и др., 2026]. Однако в экологических и биологических исследованиях риски, связанные с применением евклидовой метрики, обсуждаются на протяжении длительного времени [Orlóci, 1975; Clarke, Somerfield, Chapman, 2006; Legendre, De Cáceres, 2013; Ricotta, 2019, 2021].

Искажения, возникающие при использовании метрики Евклида для анализа несоразмерных данных, получили известность как парадокс Орлоци (Orlóci paradox). Для преодоления данной проблемы и повышения корректности анализа многомерных данных были предложены различные подходы. В настоящей работе рассматривается механизм возникновения парадокса Орлоци, а также обсуждаются возможные способы его устранения.

Одним из первых, обозначивших данную проблему, был венгерский исследователь Ласло Орлоци (László Orlóci), обративший на нее внимание в 1975 г. В классическом виде сформулированный им парадокс может быть представлен следующим образом: два наблюдения, не имеющие ни одного общего вида, могут казаться более сходными, чем два других наблюдения, у которых списки видов полностью идентичны [Orlóci, 1975, p. 28].

Разберем классический пример данного парадокса [Orlóci, 1975, p. 28; Ricotta, 2021]. Представим, что у вас есть следующая выборка из трех наблюдений с тремя видами (табл. 1). У наблюдения O1 нет общих видов с O2 и O3. Наблюдения O2 и O3 имеют общие виды с одинаковым соотношением 1 : 1. Исходя из базового сравнения можно сказать, что наблюдения O2 и O3 наиболее близки, а O1 отличается от них. Однако при применении статистических методов на основе метрики Евклида будет показана иная ситуация. O1 будет ближе к O2, чем O2 к O3.

Происходит это по следующей причине. Формула Евклидова расстояния (E) между O1 и O2:

$$E(O1, O2) = \sqrt{\sum_{j=1}^n (X_{O1j} - X_{O2j})^2},$$

где j – это переменная (в нашем случае – виды: A, B и C), n – число переменных, X_{O1j} – значение переменной j в наблюдении O1, X_{O2j} – значение переменной j в наблюдении O2. В таком случае дистанция $E(O1, O2) = 1,732$, $E(O2, O3) = 4,243$, $E(O1, O3) = 5,745$. Получается, что расстояние между наблюдениями O2 и O3 больше, чем между O1 и O2. В таком случае будет сделан неверный статистический вывод о том, что наблюдения O2 и O3 в меньшей степени похожи друг на друга, чем O2 и O1.

Однако искажения, связанные с парадоксом Орлоци, не ограничиваются лишь крайним случаем полного отсутствия общих видов или переменных между наблюдениями. Подобный эффект может проявляться и в менее очевидных ситуациях, когда сравниваемые объекты обладают несоразмерными суммарными значениями признаков. В таких условиях евклидова метрика оказывается чувствительной к общему масштабу наблюдений: различия в объеме данных начинают вносить непропорционально большой вклад в итоговую дистанцию между объектами [Orlóci, 1975, p. 28; Clarke, Somerfield, Chapman, 2006; Carlson, 2017,

р. 299–300]. В результате наблюдения, сохраняющие сходную внутреннюю структуру и сопоставимые соотношения признаков, могут интерпретироваться как статистически удаленные исключительно вследствие различий в общей величине выборки.

Для демонстрации этой ошибки рассмотрим более сложный смоделированный пример, приближенный к реалиям археологических данных (табл. 2). У нас имеется выборка из 12 памятников-наблюдений (поселения/стоянки) с 3 переменными-признаками (типы орудий/посуды). Выборка кажется вполне стандартной для археологических данных. Наблюдения нельзя назвать слишком малочисленными, на всех памятниках присутствуют десятки или сотни артефактов, нет нулей в переменных. Выборка была специально смоделирована так, чтобы данные памятники, очевидно, по соотношению имеющихся на них типов артефактов разбивались на 3 группы: 6-3-2, 2-3-6, 1-1-1 (приблизительно).

Далее выполним ординацию гипотетической выборки методом неметрического многомерного шкалирования (NMDS) в двумерном пространстве (рис. 1; уровень стресса: 0,016) с последующей проверкой различий между группами с помощью теста ANOSIM. В обоих случаях расчеты выполнялись на основе евклидовой матрицы расстояний. Все расчеты проведены в программе PAST v. 4.0 [Hammer, Harper, Ryan, 2001].

Полученная ординация не отражает исходное разделение наблюдений на три равные группы памятников, сформированные на основании сходства соотношений типов артефактов. Напротив, расположение объектов в пространстве ординации определяется преимущественно общей суммой артефактов, представленных в каждом памятнике. Вследствие этого памятники O1, O5 и O9 оказываются пространственно близки друг к другу. Аналогичным образом близкими оказываются O2, O6 и O10, хотя в несколько меньшей степени, несмотря на различия в структуре их состава. Результаты теста

Таблица 2. Гипотетическая выборка, смоделированная для демонстрации парадокса Орлоци, обусловленного различиями в масштабе наблюдений

Наблюдения/Типы	A	B	C	Сумма	Группы
O1	30	15	10	55	6-3-2
O2	56	29	18	103	6-3-2
O3	101	52	31	184	6-3-2
O4	158	78	51	287	6-3-2
O5	10	15	30	55	2-3-6
O6	18	29	56	103	2-3-6
O7	31	52	101	184	2-3-6
O8	51	78	158	287	2-3-6
O9	22	23	22	67	1-1-1
O10	34	35	34	103	1-1-1
O11	61	62	61	184	1-1-1
O12	95	96	96	287	1-1-1

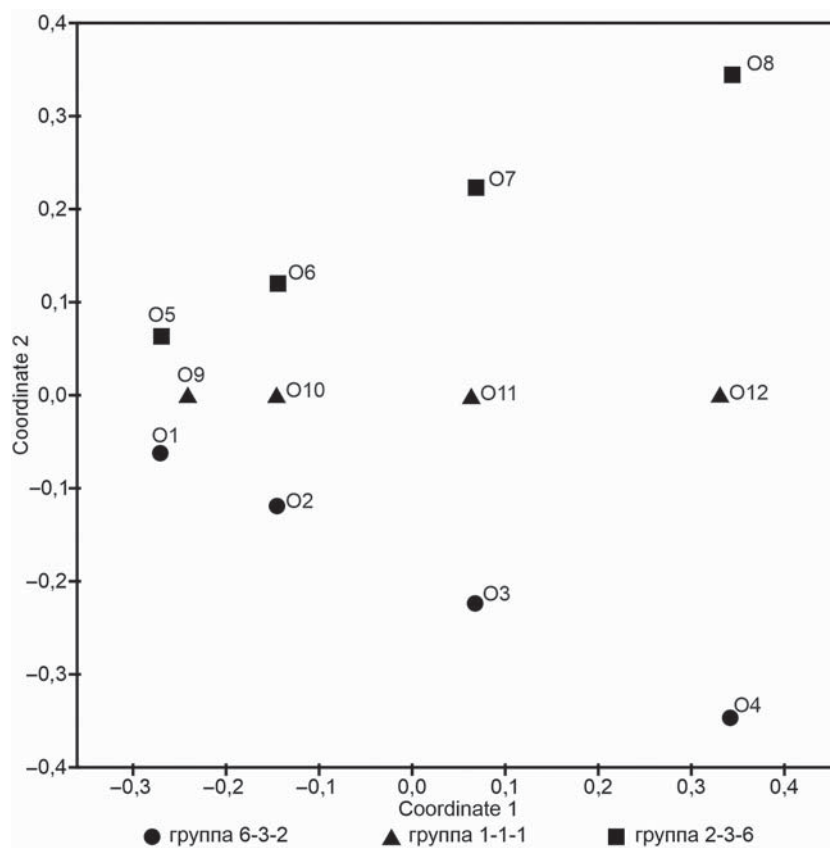


Рис. 1. Неметрическое многомерное шкалирование (NMDS) в двумерном пространстве гипотетической выборки памятников O1–O12 с использованием метрики Евклида.

ANOSIM также не выявляют статистически значимых различий между предложенными группами ($p = 0,16$): различия между памятниками внутри групп оказываются сопоставимыми с различиями между памятниками, принадлежащими к разным группам ($R = 0,11$).

Особенностью рассматриваемой исследовательской задачи является то, что первостепенное значение

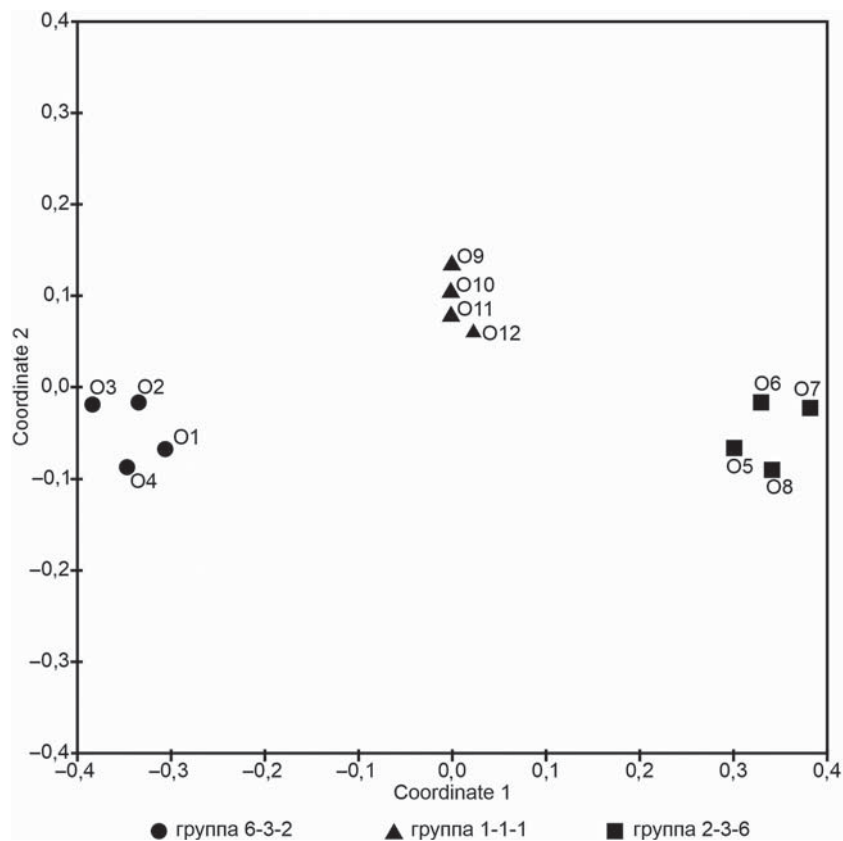


Рис. 2. Неметрическое многомерное шкалирование (NMDS) в двумерном пространстве гипотетической выборки памятников O1–O12 с использованием метрики хорды.

имеет соотношение значений по признакам, а не их суммарная величина. В связи с этим для устранения или снижения искажений, связанных с парадоксом Орлоуци возможны три подхода.

Первый предполагает предварительное преобразование исходных данных с последующим расчетом евклидовых расстояний [Cardillo, Alberti, 2013; Carlson, 2017, p. 300]. Второй подход основан на использовании мер расстояния, в которых преобразование значений является частью самой процедуры расчета, в которой далее различия между наблюдениями оцениваются в евклидовом пространстве. В качестве подобных мер расстояния предлагаются, в частности, метрики хорды (chord) и Хеллингера (Hellinger) [Orlóci, 1975, p. 28; Ricotta, 2019; Колобова и др., 2024]. Третий подход заключается в использовании иных методов статистического анализа, изначально ориентированных на сопоставление структуры распределения признаков, а не абсолютных различий между наблюдениями. Наиболее распространенный подобный метод в археологии – анализ соответствий (correspondence analysis). Анализ соответствий сравнивает наблюдения не по обычному евклидову расстоянию (как, например, анализ главных компонент), а по хи-квадрат расстоянию между распределениями признаков внутри каждого наблюдения. Поэтому результат анализа отражает различия

наблюдений по относительным частотам признаков, а не по абсолютным количествам. Наиболее многочисленные категории оказывают меньшее влияние на результат, тогда как редкие становятся более заметными [Baxter, 1994, p. 107–114].

В текущей работе будет использован второй подход. Рассмотрим наши группы поселений на графике неметрического многомерного шкалирования (NMDS) в двумерном пространстве и проверим кластеры тестом ANOSIM, используя метрику хорды. На графике (рис. 2; уровень стресса: 0,025) полученная ординация отражает разделение наших памятников на 3 группы по различному соотношению в них типов артефактов. По результатам ANOSIM выделенные группы статистически значимо различаются ($p = 0,0003$; $R = 1$). Наблюдения внутри каждой группы более сходны между собой, чем с наблюдениями из других групп.

Таким образом, сопоставление результатов, полученных с ис-

пользованием евклидовой метрики и метрики хорды, демонстрирует принципиальное значение выбора меры расстояния для интерпретации археологических данных.

В первом случае были отражены преимущественно различия в суммарном количестве артефактов, вследствие чего изначально заданные группы, основанные на сходстве соотношений типов, не были распознаны как статистически обособленные.

Во втором случае применение метрики хорды позволило снизить влияние общего масштаба наблюдений и выявить различия в структуре комплексов. Вместе с тем при значительном количестве переменных с нулевыми значениями определенный риск искажений может сохраняться даже при использовании метрик хорды и Хеллингера [Cardillo, Alberti, 2013; Ricotta, 2021].

Рассмотренные примеры действия парадокса Орлоуци являются гипотетическими и идеальными. Реальные археологические выборки куда более сложные и разнородные по количеству наблюдений, переменных и их значений. В связи с этим применение евклидовой метрики к исходным данным в ряде ситуаций способно приводить к статистически убедительным, но содержательно ошибочным результатам, которые выявить и зафиксировать будет гораздо сложнее [Тюгашев, 2023б; Колобова и др., 2026].

Таким образом, выбор меры расстояния не следует рассматривать как формальный или технический этап статистического анализа. При применении методов многомерной статистики желательно предварительно определить исследовательскую задачу, оценить полноту, структуру и характер исходных данных и, исходя из этого, выбирать подходящие методы анализа. В ряде случаев для получения более корректных результатов может потребоваться предварительное преобразование данных либо использование специализированной метрики, например, менее чувствительной к различиям в масштабе наблюдений. Поэтому при описании статистического анализа также стоит указывать, какая именно мера расстояния использовалась, если применяемый метод допускает выбор между несколькими метриками.

Финансирование

Исследование выполнено в рамках проекта НИР ИАЭТ СО РАН № FWZG-2025-0007 «Применение цифровых технологий при анализе археологических источников и реконструкции истории древних сообществ».

Список литературы

Колобова К. А., Кривошапкин А. И., Павленок Г. Д., Павленок К. К. Заполняя лауну: средний палеолит Кулбулака в контексте продолжительной хронологии обихватского варианта // *Stratum plus*. – 2026. – № 1. – С. 87–104.

Колобова К. А., Харевич А. В., Мухтаров Г. А., Кривошапкин А. И. Атрибуция и вариабельность многослойных палеолитических комплексов: пример орудийных индустрий грота Оби-Рахмат // *Camera Praehistorica*. – 2024. – № 2 (13). – С. 24–33.

Тюгашев И. Е. Оценка вариабельности майкопской культуры по данным прикладной статистики // 300-летие Российской академии наук – археология и этнография Сибири: традиции, школы и открытия. – Новосибирск: Изд-во ИАЭТ СО РАН, 2023а. – С. 123–125.

Тюгашев И. Е. Использование метода многомерного шкалирования для выделения групп памятников майкопской культуры // Проблемы археологии, этнографии, антропологии Сибири и сопредельных территорий. – Новосибирск: Изд-во ИАЭТ СО РАН, 2023б. – Т. XXIX. – С. 929–936.

Baxter M. J. Exploratory Multivariate Analysis in Archaeology. – Edinburgh: Edinburgh Univ. Press Ltd, 1994. – 307 p.

Cardillo M., Alberty J. Stone Tool Manufacture Strategies and Lithic Raw Material Exploitation in Coastal Patagonia, Argentina: A Multivariate Approach // *J. of Archaeology*. – 2013. – Vol. 2013, N 1. – Article ID 128470. – P. 1–12.

Carlson D. L. Quantitative Methods in Archaeology Using R. – Cambridge: Cambridge Univ. Press, 2017. – 432 p.

Clarke K. R., Somerfield P. J., Chapman M. G. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages // *J. of Experimental Marine Biology and Ecology*. – 2006. – Vol. 330, N 1. – P. 55–80.

Gehlen B., Schneid N., Roth G., Zander A. Typochronology for the Mesolithic between 9000 and 7800 cal BC in Central Europe: a new approach to use constrained correspondence analysis (CCA) of microliths for dating // *From the Early Preboreal to the Subboreal period – Current Mesolithic research in Europe. Studies in honour of Bernhard Gramsch*. – Norderstedt: Welt und Erde Verlag, Kerpen-Loogh, 2020. – P. 315–367.

Hammer Ø., Harper D. A. T., Ryan P. D. PAST: Paleontological Statistics Software Package for Education and Data Analysis // *Palaeontologia Electronica*. – 2001. – Vol. 4, N 1. – URL: https://palaeo-electronica.org/2001_1/past/past.pdf (Accessed: 25.09.2023).

Legendre P., De Cáceres M. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning // *Ecology Letters*. – 2013. – Vol. 16, iss. 8. – P. 951–963.

Orlói L. Multivariate Analysis in Vegetation Research. – Hague: Springer-Science+Business Media, 1975. 1st edition. 276 p.

Peterson C. E., Drennan R. D., Bartel K. L. Comparative Analysis of Neolithic Household Artifact Assemblage Data from Northern China // *J. of Anthropol. Research*. – 2016. – Vol. 72, N 2. – P. 200–225.

Ricotta C. Can we trust the chord (and the Hellinger) distance? // *Community Ecology*. – 2019. – Vol. 20, N 1. – P. 104–106.

Ricotta C. From the euclidean distance to compositional dissimilarity: What is gained and what is lost // *Acta Oecologica*. – 2021. – Iss. 111: 103732. – P. 1–3.

Tratebas A. North American–Siberian Connections: Regional Rock Art Patterning Using Multivariate Statistics // *A Companion to Rock Art*. – Chichester: Blackwell Publ. Ltd, 2012. – P. 143–159.

References

Baxter M. J. Exploratory Multivariate Analysis in Archaeology. Edinburgh: Edinburgh Univ. Press Ltd, 1994. 307 p.

Cardillo M., Alberty J. Stone Tool Manufacture Strategies and Lithic Raw Material Exploitation in Coastal Patagonia, Argentina: A Multivariate Approach. *J. of Archaeology*, 2013. Vol. 2013, No. 1. Article ID 128470. P. 1–12. doi:10.1155/2013/128470

Carlson D. L. Quantitative Methods in Archaeology Using R. Cambridge: Cambridge Univ. Press, 2017. 432 p. doi:10.1017/9781139628730

Clarke K. R., Somerfield P. J., Chapman M. G. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages. *J. of Experimental Marine Biology and Ecology*, 2006. Vol. 330, No. 1. P. 55–80. doi:10.1016/j.jembe.2005.12.017

Gehlen B., Schneid N., Roth G., Zander A. Typochronology for the Mesolithic between 9000 and 7800 cal BC in Central Europe: a new approach to use constrained correspondence analysis (CCA) of microliths for dating. In *From the Early Preboreal to the Subboreal period – Current Mesolithic research in Europe. Studies in honour of Bernhard Gramsch*. Norderstedt: Welt und Erde Verlag, Kerpen-Loogh, 2020. P. 315–367.

Hammer Ø., Harper D. A. T., Ryan P. D. PAST: Paleontological Statistics Software Package for Education

and Data Analysis. *Palaeontologia Electronica*, 2001. Vol. 4. No. 1. URL: https://palaeo-electronica.org/2001_1/past/past.pdf (Accessed: 25.05.2026).

Kolobova K. A., Kharevich A. V., Muhtarov G. A., Krivoschapkin A. I. Attribution and variability of multilayered paleolithic complexes: the case of the Obi-Rahmat tool kits. *Camera Praehistorica*, 2024. Vol. 2 (13). P. 24–33. (In Russ.). doi:10.31250/2658-3828-2024-2-24-33

Kolobova K. A., Krivoschapkin A. I., Pavlenok G. D., Pavlenok K. K. Filling the Gap: The Middle Palaeolithic of Kulbulak in the Context of the Extended Chronology of the Obi-Rakhmat Variant. *Stratum Plus*, 2026. No. 1. P. 87–104. (In Russ.). doi:10.55086/sp26185102

Legendre P., De Cáceres M. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters*, 2013. Vol. 16, Iss. 8. P. 951–963. doi:10.1111/ele.12141

Orlóci L. Multivariate Analysis in Vegetation Research. Hague: Springer-Science+Business Media, 1975. 1st edition. 276 p. doi:10.1007/978-94-017-5608-2

Peterson C. E., Drennan R. D., Bartel K. L. Comparative Analysis of Neolithic Household Artifact Assemblage Data from Northern China. *J. of Anthropol. Research*, 2016. Vol. 72, No. 2. P. 200–225. doi:10.1086/686312

Ricotta C. Can we trust the chord (and the Hellinger) distance? *Community Ecology*, 2019. Vol. 20, No. 1. P. 104–106. doi:10.1556/168.2019.20.1.11

Ricotta C. From the euclidean distance to compositional dissimilarity: What is gained and what is lost. *Acta Oecologica*, 2021. Iss. 111: 103732. P. 1–3. doi:10.1016/j.actao.2021.103732

Tratebas A. North American–Siberian Connections: Regional Rock Art Patterning Using Multivariate Statistics. In *A Companion to Rock Art*. Chichester: Blackwell Publ. Ltd, 2012. P. 143–159.

Tyugashev I. E. The Re-evaluation of the Variability of the Maykop Culture According to Applied Statistics. In *Proceedings of the LXIII Russian (with International Participation) Archaeological and Ethnographic Conference of Students and Young Scientists*. Novosibirsk: IAET SB RAS Publ., 2023a. P. 123–125. (In Russ.).

Tyugashev I. E. Use of Multidimensional Scaling Method for Identifying Groups of Sites of the Maykop Culture. In *Problems of Archaeology, Ethnography, Anthropology of Siberia and Neighboring Territories*. Novosibirsk: IAET SB RAS Publ., 2023b. Vol. 29. P. 929–936. (In Russ.). doi:10.17746/2658-6193.2023.29.0929-0936

Тюгашев И. Е. <https://orcid.org/0000-0001-5885-1535>

Дата сдачи рукописи: 31.05.2026 г.