Широбоков И.Г.

МАЭ РАН, Отдел антропологии, Университетская наб., 3, Санкт-Петербург, 199034, Россия

ОБ ОТНОСИТЕЛЬНОЙ ТОЧНОСТИ ОЦЕНКИ ПОЛА ПО ЧЕРЕПУ

Введение. *В настоящее время существует не менее двухсот алгоритмов определения пола по черепу, основанных на статистическом анализе дискретных, линейных, угловых признаков и их комбинаций. И все же многие антропологи предпочитают опираться на личные визуальные наблюдения. Задачи настоящего исследования состоят в изучении возможных причин предпочтения визуального подхода, а также в анализе сравнительной эффективности визуальных и статистических оценок пола по черепу.*

Материалы и методы. *Исследование опирается на анализ публикаций, посвященных методам оценки пола по черепу, вышедшим за последние 70 лет. Сопоставление оценок точности проводилось при помощи непараметрических тестов с учетом различий в статистических методах, подходах к валидации результатов (без валидации, перекрестная проверка, независимое тестирование) и системах фиксации признаков (балловые признаки, краниометрия, геометрическая морфометрия).*

Результаты. *Общие причины недоверия к алгоритмам заключаются в завышенных ожиданиях относительно их возможностей, большей предвзятости к ошибкам, совершаемым моделями, чем совершаемым людьми, отсутствием контроля над классификацией. При этом алгоритмы, как правило, превосходят экспертов в прогнозировании целевой переменной. Средняя точность визуальных оценок пола по черепу несколько ниже оценок статистических моделей и отличается заметной вариативностью. Точность оценок опытных антропологов близка к средним показателей таковой у моделей. Однако эффективность алгоритмов заметно снижается в случае их применения к сериям, отличающихся по своему происхождению от обучающей выборки, особенно при работе с краниометрическими показателями. В значительной части исследований размер обучающих выборок недостаточен для надежной оценки эффективности моделей, а соотношение полов искажено в пользу мужских черепов, что приводит к некоторому завышению точности их определений. Эффективность моделей может также снизится из-за погрешностей при фиксации балловых признаков, причем оценка межисследовательских расхождений не позволяет определить их влияние на точность модели.*

Заключение. *Несмотря на обширную библиографию, сегодня по-прежнему сохраняется дефицит данных как о точности визуального подхода к оценке пола, так и о надежности моделей с заявленной высокой эффективностью. Внедрение гибких методик, позволяющих исследователям самостоятельно контролировать как отбор признаков, так и состав обучающей выборки, поможет преодолеть неприятие алгоритмов и повысить качество определений.*

Ключевые слова: морфология черепа; определение пола; краниометрия; дискретные признаки; недоверие к алгоритмам

DOI: 10.32521/2074-8132.2023.4.069-080

Введение

Определение биологического пола является одним из первых этапов исследования каждого антрополога, работающего со скелетными останками, независимо от поставленных целей исследования. Приоритетное значение тазовых костей в оценке пола не вызывает сомнений у современных исследователей. Точность оценок, вынесенных по результатам изучения морфологических особенностей черепа, несколько уступает точности определений, основанных на характеристиках таза и других костей посткраниального скелета [Spradley, Jantz, 2011], однако череп по-прежнему остается одним из наиболее популярных объектов для анализа межполовых различий в антропологии. Значительную часть коллекций антропологических музеев и институтов в России и за рубежом составляют именно краниологические серии, и череп является единственным источником информации о половозрастной характеристике индивида. Кроме того, кости черепа хорошо поддаются идентификации в случаях, когда посткраниальный скелет отличается плохой сохранностью или фрагментирован, а многие полодифференцирующие признаки могут быть визуально определены без привлечения измерительных инструментов.

Мужские и женские черепа различаются как по общим размерам, так и по форме отдельных элементов. По сравнению с женщинами у мужчин, как правило, сильнее развиты места прикрепления мышц (затылочный гребень, сосцевидные отростки, височные линии, скуловые отростки, углы нижней челюсти), заметнее выражен рельеф надбровной области и надпереносья, чаще встречается наклонный лоб с притупленным верхним краем орбит [Алексеев, Дебец, 1964; Stewart, 1979; Krogman, İşcan, 1986]. Некоторые исследователи выделяют до 40 различных морфологических признаков, комбинации которых с разной степенью уверенности позволяют разделить черепа взрослых мужчин и женщин [Звягин, 1983].

Поскольку для большинства полодифференцирующих признаков черепа затруднительно определить точное анатомическое положение точек, пригодных для проведения сопоставимых измерений, их оценка чаще всего проводится в бинарной системе или в балловой шкале. С развитием технологий и статистических методов анализа некоторые признаки стали проще поддаваться формализации, а исследования с большими сериями с задокументированным полом, позволили оценить значимость (вес) каждого из них с учетом морфологической характеристики и выраженности полового диморфизма в популяциях разного происхождения. По расчетам автора к настоящему времени опубликовано не менее двухсот различных алгоритмов для оценки пола по черепу. Одни из них опираются на результаты статистического анализа различных наборов признаков, охватывающих весь череп, другие сосредоточены на изучении его отдельных элементов (основании черепа, височных костей или нижней челюсти). Большая часть из них представляет собой модели, в основе которых лежат различные варианты линейного (реже квадратичного) дискриминантного анализа (для метрических показателей) или логистическая регрессия (для дискретного или смешанного набора признаков). Однако с каждым годом число алгоритмов продолжает увеличиваться, постепенно возрастает доля моделей, основанных на различных методах машинного обучения, отличающихся наибольшей гибкостью и в некоторых случаях даже не требующих фиксации признаков самим исследователем.

На практике большинство алгоритмов пока не получили того распространения, на которое могли бы претендовать по своему прямому назначению и заявленной эффективности. Строго формализованные методы оценки пола преимущественно привлекают внимание специалистов в области судебной антропологии и криминалистов. Антропологи, работающие с материалами из археологических раскопок, чаще ориентируются на общее визуальное впечатление, как правило, пренебрегая статистическими моделями даже в тех случаях, когда фиксируют степень выраженности отдельных признаков в балловой шкале. Это избегание строгой оценки особенно заметно при знакомстве с работами отечественных краниологов, в которых нередко использованные методы оценки пола не указываются вообще или информация о них ограничивается ссылкой на «принятые в современной антропологии методы» или общие методические руководства. Возможно, та-

кая ситуация объясняется уверенностью антропологов в том, что визуальная оценка обеспечивает достаточно высокий уровень точности. Потребность в использовании дополнительных статистических методов в большинстве случаев просто не воспринимается как актуальная, а при работе с фрагментированным материалом зачастую и не может быть удовлетворена из-за невозможности оценить весь необходимый для выполнения алгоритма комплекс признаков.

Задачи настоящего исследования заключаются в том, чтобы (1) проанализировать наиболее вероятные причины игнорирования антропологами строгих методов оценки пола по черепу; (2) сравнить среднюю точность визуального и различных статистических подходов к оценке пола.

Материалы и методы

Материалами настоящего исследования послужили литературные данные. Автором были отобраны и проанализированы публикации, посвященные как новым методам оценки пола по черепу, так и результатам тестирования ранее предложенных подходов. Всего в исследовании были учтены данные 130 публикаций, вышедших за последние 70 лет (и преимущественно относящихся к последним двум десятилетиям). Анализировались следующие показатели: 1) метод классификации (алгоритм); 2) тип оцениваемых признаков (балловые, линейные, угловые); 3) точность оценки пола (доля индивидов с корректно установленным полом относительно числа всех индивидов в выборке); 4) метод валидации достигнутой точности; 5) размер обучающей выборки.

При сравнении эффективности подходов рассчитывались средние значения точности и ее стандартная ошибка для совокупностей моделей, сгруппированных по перечисленным выше показателям. Различия между оценками, полученными в рамках разных подходов, тестировались при помощи непараметрического критерия Манна-Уитни, различия между оценками точности, рассчитанными отдельно для мужских и женских черепов – при помощи Т-критерия Уилкоксона.

В тех случаях, когда в публикации приводятся результаты тестирования разных наборов признаков, рассмотренных в рамках одного метода, при расчетах учитывались максимальные из достигнутых исследователем показатели. В случае, если автор использовал в анализе несколько различных статистических методов, в расчет принимались результаты каждого из них. Если публикация посвящена результатам тестирования нескольких ранее предложенных моделей, полученных при помощи одного статистического метода, но основанных на разных наборах признаков, учитывались оценки точности тех из них, которые демонстрировали максимальную эффективность по результатам исходного анализа. Суммарно были учтены показатели точности 176 моделей. Полный список проанализированных публикаций представлен на сайте https://www.academia.edu/105810838/Skull_sex_estimation_accuracy.

Результаты и обсуждение

Общие причины недоверия к алгоритмам

Прежде чем обратиться к сравнительному анализу эффективности методов, необходимо рассмотреть вопрос о наиболее распространенных причинах недоверия людей к алгоритмам вообще, т.е. к любым статистическим методам оценки некоторой целевой переменной (в нашем случае пола), опирающимся на анализ строго формализованного набора признаков.

В специальных исследованиях было показано, что оценки целевой переменной, полученные при помощи алгоритмов, оказываются практически всегда точнее заключений экспертов. В частности, превосходство алгоритмов над людьми продемонстрировано на примере исследований в сфере медицины, судебной экспертизы, психиатрии, рекрутинга и некоторых других прикладных областях человеческой деятельности (см. обзор в: [Канеман с соавт., 2021]). Причины относительно низкой эффективности экспертных оценок заключаются в их неустойчивости к шуму: люди чрезвычайно восприимчивы к внешним по отношению к интересующей их проблеме факторам. При решении задачи мы склонны переоценивать значимость отдельных признаков, нередко излишнее внимание уделяя деталям, которые представляются нам принципиальными, но не являются таковыми в действительности. Как следствие, оценки людей

отличает от оценок алгоритмов сравнительно низкая воспроизводимость. Высока вероятность, что два человека, изучивших один и тот же набор данных, придут к разным заключениям. Причем различия могут обнаружится даже между оценками одного человека, разделенными некоторым промежутком времени [Канеман с соавт., 2021; Dawes et al.,1989].

Несмотря на доказанную относительно высокую эффективность даже простых статистических моделей, люди не склонны полагаться на их точность. Неприятие может выражаться как в нежелании полагаться на алгоритм, в оценке которого пользователь обнаружил ошибку, так и в недоверии к статистическим моделям вообще, не связанным с личным опытом их применения [Berger et al., 2021]. При этом люди менее терпимы к ошибкам, допущенным алгоритмом, чем к ошибкам людей, даже в тех случаях, когда последние ошибаются чаще [Dietvorst et al., 2015; Renier et al., 2021]. Причины неприятия алгоритмов могут быть разными, включая ложные ожидания относительно возможностей и производительности алгоритмов, отсутствие контроля над процессом вынесения решения (непрозрачность моделей), расхождения между интуитивным решением задачи человеком и оценками модели, а также некоторые другие [Burton et al., 2020].

Дополнительная проблема заключается в том, что зачастую у людей отсутствует возможность верифицировать свои оценки, и потребность в повышении их точности не воспринимается как актуальная. Это замечание нередко оказывается справедливым по отношению к заключениям физических антропологов, работающим с материалами из археологических раскопок. Часто мы не можем проверить корректность своих половозрастных определений, и точность вынесенных оценок не подвергается проверке на прочность. С другой стороны, статистические модели обучаются и тестируются на материалах с задокументированным полом и возрастом. Допускаемые ими ошибки очевидны, тогда как сам алгоритм классификации напротив, кажется чрезмерно жестким и непрозрачным.

При этом исследователями неоднократно высказывались сомнения относительно целесообразности использования при оценке пола скелета статистических моделей. В частности, было показано, что линейные дискриминантные функции нередко уступают в эффективности опытным экспертам, ориентирующимся на общее визуальное впечатление от черепа [Stewart, 1954; Henke, 1977; Dereli et al., 2018; Lewis, Garvin, 2016]. Наиболее существенная проблема, по-видимому, заключается в том, что в отличие от тазовых костей, черепа обладают заметной межгрупповой изменчивостью, в т.ч. по степени выраженности межполовых различий. Использование неподходящего алгоритма может катастрофически снизить точность оценки, поэтому большинство исследователей подчеркивает, что при оценке пола важно использовать модели, разработанные для конкретной популяции. Обзор публикаций показывает, что в большинстве случаев авторы предлагают именно регионально-специфические методы оценки пола. Недавно опубликованные оптимистичные результаты тестирования нескольких моделей, претендующих на универсальность, требуют проведения независимых исследований [Del Bove, Veneziano, 2022; Kelley, Tallmann, 2022].

Об эффективности визуальной оценки пола по черепу

Несмотря на широкое распространение среди антропологов субъективно-визуального подхода к оценке пола, публикации, в которых анализируются или даже просто указываются оценки его точности, немногочисленны. В некоторых случаях они носят явно умозрительный характер, тогда как в других недостаточно полно описаны условия тестирования. А. Хрдличка указывал, что пол по черепу с нижней челюстью может быть верно установлен в 90% случаев и в 80% случаев, если нижняя челюсть отсутствует [Hrdlicka, 1939], но не приводит никаких обоснований для этой оценки. Близкие оценки были получены У. Крогманом, работавшим с коллекцией с задокументированным полом, в которой, однако, полностью преобладали мужские черепа [Krogman, 1986]. Недавний обзор результатов сопоставления оценок судебных антропологов и тестов ДНК также показал эффективность оценок экспертов (92%) [Thomas et al., 2016]. В последнем случае, однако, объединены результаты, в которых использовались как визуальные методы оценки, так и методы, основанные на морфометрии. В некоторых других исследованиях точность оценки

пола заметно ниже. Ученик А. Хрдлички Т. Стюарт сообщает о 77% случаев корректных определений пола по черепу [Stewart, 1979], и такой же точности удалось добиться при работе с задокументированными коллекциями Ф. Кампсу [Camps et al., 1968].

Точность определений, по всей вероятности, отчасти зависит от опыта антрополога. Рассчитанная по литературным данным средняя точность классификаций по полу составляет 85.9% для визуальных оценок опытных антропологов и 83.5% для оценок без учета опыта. Начинающие исследователи заметно чаще ошибаются, опираясь на визуальную оценку пола по черепу, чем их старшие коллеги [Đurić et al., 2005; Berg, Tersigni-Tarrant, 2014; Lewis, Garvin, 2016]. При этом возможность работать как с черепом, так и тазовыми костями, по всей видимости, минимизирует влияние опыта при визуальной оценке [Đurić et al., 2005].

Влияние практики на точность оценки отчасти подтверждают результаты научно-практического семинара, прошедшего в МАЭ РАН в 2015 году в рамках конференции «Палеоантропологические и биоархеологические исследования: традиции и новые методики». Участникам семинара предлагалось определить половозрастные характеристики 15 случайно отобранных черепов из коллекции с задокументированным полом и возрастом (МАЭ РАН №1830). Всего тесте приняли участие 17 человек, средняя точность оценок пола составила 74%.[1] В тех случаях, когда можно было установить авторство определений (по условию теста допускались и анонимные оценки), точность в среднем оказалась выше в старшей группе исследователей по сравнению с аспирантами и молодыми исследователями (77% и 67% соответственно). В то же время изменчивость оценок возраста не обнаружила никакой связи с опытом исследователей: средние колебания интервальных оценок относительно реального возраста индивидов у старших коллег имели такую же величину, как и у молодых антропологов. Причина разной роли опыта, вероятно, заключается в том, что на практике антропологи чаще всего сопоставляют полодифференцирующие признаки на отдельных элементах скелета, что позволяет корректировать оценки значимости отдельных признаков в конкретных выборках. Задачу облегчает бинарность классификации, помогающая при работе с сериями черепов. В то же время оценки возраста взрослых индивидов чаще всего не поддаются качественной корректировке, поскольку она возможна лишь при работе с коллекциями с задокументированным возрастом смерти.

При этом оценки даже хорошо подготовленных исследователей, имеющих большой опыт работы со скелетными останками, могут расходиться между собой [Meindl et al. 1985; Walrath et al. 2004]. На их вариативность и точность серьезное влияние может оказывать контекстная информация, особенно в тех случаях, когда кости скелета не имеют выраженных мужских или женских признаков, или их комбинация представляется исследователю противоречивой [Nakhaeizadeh et al. 2020].

Об эффективности статистических моделей

Но действительно ли алгоритмы эффективнее людей в оценке пола по черепу?

Результаты анализа литературных данных не позволяют однозначно согласиться с этим утверждением. Рассчитанные автором оценки средней точности статистических моделей и визуальных определений приведены в таблице 1. Для сравнения моделей были отобраны оценки, полученные по результатам кросс-валидации и/или применения обученных моделей к тестовым выборкам, отобранным авторами предлагаемых методов.

Такой отбор обусловлен необходимостью нивелировать влияние метода валидации на заявленную точность метода. Авторы некоторых работ приводят оценки точности, полученные для обучающей выборки, или не указывают способ валидации. Такие оценки теоретически могут оказаться излишне оптимистичными (завышенными), поскольку в первую очередь отражают способность модели подстраиваться под изменчивость признаков в конкретной серии черепов (при таком подходе многие методы машинного обучения позволяют добиться точности в

[1] Поскольку была исследована относительно небольшая серия черепов, полученные результаты не учитывались при сравнении средних оценок точности визуального подхода и статистических моделей.

100% случаев). В других исследованиях выборки разбиваются на обучающую и тестовую часть, и точность оценивается только для последней. Еще более эффективный (и чаще используемый) подход предполагает, что выборка должна быть случайным образом разбита на подгруппы. Затем проводится серия испытаний, в ходе которых каждая из подгрупп поочередно играет роль тестовой, в то время как остальные используются для обучения модели. В большинстве публикаций, посвященных оценке пола по черепу, использован вариант перекрестной проверки, при котором число подгрупп равно числу индивидов в выборке (Leave-One-Out Cross Validation). В этом случае каждый череп поочередно используется как тестовый набор, а остальные черепа используются для обучения модели, а затем рассчитывается средняя доля правильных классификаций по полу. Наконец, еще один подход предполагает проведение анализа эффективности ранее предложенных методов независимой группой исследователей. Часто для тестирования привлекаются выборки иного происхождения, нежели были использованы в исходном исследовании, что, как правило, приводит к снижению исходной точности. Большинство таких независимых исследований посвящено тестированию эффективности дискриминантных функций, предложенных Ю. Джайлсом и О. Эллиотом на основе ряда краниометрических параметров [Giles, Elliot, 1963], а также уравнений логистической регрессии для комбинаций из пяти балловых признаков, рассчитанных Ф. Уокером [Walker, 2008]. При этом большинство опубликованных моделей никогда не подвергалось независимой оценке.

Средняя точность оценок пола по черепу, рассчитанная по результатам кросс-валидации, приблизительно одинакова при использовании дискриминантных функций (часто используемой для анализа линейных и угловых размеров) и логистической регрессии (используемой при анализе как метрических, так и дискретных признаков) и составляет около 87–88%. Наибольшей точностью обладают модели, обученные при помощи методов машинного обучения (около 90%) – отличия от других статистических методов имеют значимую величину (U-критерий Манна Уитни, p=0.007). При этом непараметрические критерии не обнаруживают различий между медианными значениями точности оценок, рассчитанными при помощи статистических моделей и установленными визуально. Отчасти это объясняется недостаточным объемом опубликованных данных о точности визуальной оценки, отчасти ее высокой изменчивостью, заметно превышающей изменчивость оценок статистических моделей.

На первый взгляд при визуальной оценке также выше различия в доле точных оценок мужских и женских черепов, но это смещение не выходит за пределы случайного. Вопреки прежним оценкам, согласно которым чаще встречаются ошибки в определении пола по черепу у мужчин [Meindl et al., 1985; Weiss, 1972], систематические различия в точности оценок между полами в большинстве случаев не наблюдаются или их величина не выходит за пределы случайных вариаций. Исключение составляют оценки, полученные при помощи дискриминантного анализа. В этом случае доля точных оценок мужчин

Таблица 1. Средняя точность и стандартная ошибка визуальной оценки и статистических моделей, предназначенных для определения пола по черепу
Table 1. Average accuracy and standard error of visual assessment and statistical models designed to determine sex from the skull

Подход	Средняя точность оценки (%)	Средняя точность оценки мужчин (%)	Средняя точность оценки женщин (%)
Визуальная оценка	83.3±2.7 (16)	87.7±5.2 (6)	85.4±3.8 (6)
Визуальная оценка (опытные антропологи)	85.9±2.1 (14)	92.3±3.1 (5)	85.3±4.7 (5)
Дискриминантный анализ	87.2±0.6 (103)	86.3±0.7 (78)	88.2±0.8 (78)
Логистическая регрессия	87.9±1.0 (27)	88.2±1.2 (22)	87.8±1.4 (22)
Методы машинного обучения	90.4±0.9 (29)	90.6±1.0 (22)	91.8±1.3 (22)

Примечания. В скобках указано число моделей (в случае визуальной оценки – исследователей)
Notes. The number of models is given in parentheses (in the case of visual estimation, the number of researchers)

оказывается действительно несколько ниже, чем оценок женщин (W-критерий Уилкоксона, n=78, p=0.0004). Причины, по которым заключение о средней равной точности определений пола у мужчин и женщин может в действительности оказаться некорректным, будут рассмотрены в следующем разделе.

Если в расчет принять средние визуальные оценки антропологов, в том числе тех, кто знаком с методами определения пола по черепу, но не имеет большего практического опыта работы со скелетами, то точность алгоритмов несомненно превзойдет точность визуальных оценок – независимо от того будем ли мы группировать методы по типу алгоритма или способу первичной фиксации признаков (баллы, линейные признаки, совокупность точек с четкой анатомической локализацией и полуточек). Однако это наблюдение будет справедливо только в рамках оценок точности моделей, рассчитанных их авторами. Поскольку точность таких оценок привязана к популяционной изменчивости показателей, эффективность методов может заметно снизиться в случае их применения к другим популяциям (табл. 2). При этом если точность моделей, основанных на оценке дискретных и балловых признаков, оказывается в среднем на 9% ниже исходной, для моделей, опирающихся на краниометрические данные, средняя разница составляет уже 16%. На практике снижение эффективности имеет даже более катастрофический эффект, о чем позволяет судить сравнение оценок точности отдельно мужских и женских черепов. Смещение точности оценки по полу заметно возрастает при использовании неподходящей популяционной модели, лишая смысла ее применение. К сожалению, из-за отсутствия независимых исследований невозможно оценить уровень снижения эффективности моделей, в основе которой лежат методы геометрической морфометрии, но можно ожидать, что оно также будет превышать наблюдаемую разницу для балловых признаков.

Вероятно, перекрестная проверка не всегда позволяет избежать завышения оценок точности даже в случаях, учитывающих популяционные требования к их применению. Средние оценки точности для дискриминантных функций и логистической регрессии, приведенные в таблице 1, совпадают с соответствующими показателями, полученными исследователями, которые вообще не прибегали к валидации данных. В ходе работы исследователи могут тестировать различные комбинации признаков, часть из которых оказываются более эффективными для дифференциации выборки, чем другие (и рекомендуются к применению в дальнейшем). Как было указано выше, при расчете средних оценок точности автором учитывались максимальные из достигнутых показатели, при обобщении результатов независимого тестирования – оценки тех моделей, которые продемонстрировали наибольшую точность в исходных исследованиях. Однако последние нередко оказывались не самыми эффективными, уступая в точности другим тестируемым моделям. Различия между оценками точности, полученными для разных комбинаций

Таблица 2. Средняя точность и стандартная ошибка оценки пола по разным системам признаков черепа
Table 2. Average accuracy and standard error of sex estimation by different cranial trait systems

Подход	По результатам кросс-валидации и тестов			По результатам независимого тестирования		
Система признаков	средняя точность оценки (%)	точность оценки мужчин (%)	точность оценки женщин (%)	средняя точность оценки (%)	точность оценки мужчин (%)	точность оценки женщин (%)
Балловая система	89.6±1.0 (16)	90.3±1.7 (13)	90.0±1.6 (13)	80.9±1.4 (22)	83.7±2.4 (20)	78.3±3.0 (20)
Краниометрия / линейные признаки	87.4 ± 0.5 (74)	87.0 ± 0.6 (52)	88.9 ± 0.7 (52)	71.1±2.9 (17)	80.0±5.3 (14)	60.5±8.0 (14)
Геометрическая морфометрия	88.2 ± 2.1 (16)	87.5 ± 3.0 (11)	87.0 ± 3.1 (11)	Нет данных	Нет данных	Нет данных

Примечания. В скобках указано число моделей.
Notes. The number of models is given in parentheses.

признаков, отчасти обусловлены не их объективной дифференцирующей способностью, а лучшей подстройкой к изменчивости межполовых различий в конкретной выборке. Иными словами, когда исследователь тестирует различные наборы признаков, пытаясь выбрать наиболее эффективную модель, точность наилучшей из них может оказаться завышенной.

Строгое тестирование требует привлечения к анализу независимых данных, которые не использовались при обучении моделей. Приходится признать, что соблюдение этого условия вообще невозможно проследить по публикациям, поскольку авторы, недовольные полученными результатами, могут изменить набор признаков или метод их статистического анализа и вновь запустить перекрестную проверку или провести дополнительное тестирование на «более подходящем» материале до достижения приемлемого результата. Весьма вероятно, что авторы, стремясь к получению оптимальных результатов в ходе этих манипуляций, не осознают, что их действия приводят к завышению реальных показателей эффективности отобранной итоговой модели. Именно поэтому сегодня актуальной задачей является не разработка новой эффективной модели оценки пола по черепу, а проведение независимого тестирования десятков уже предложенных методов с заявленной высокой точностью оценки.

О человеческих недостатках статистических моделей

Причины, по которым гипотеза о преимуществе моделей перед экспертной оценкой не находит однозначного подтверждения в задаче определения пола по черепу, могут быть разными. Одна из них уже указывалась выше: надежных данных об эффективности визуальной оценки пока просто недостаточно. Другая важная причина заключается в роли человеческого фактора, негативно влияющего как на качество исходных данных, так и на качество моделей.

Задача создания надежной модели требует привлечения к анализу больших серий, в которой в равных долях представлены мужские и женские черепа, однако большинству антропологов не удается найти достаточный материал для региональных моделей. Средний размер обучающей выборки составляет 349 черепов, при этом более чем в половине всех публикаций

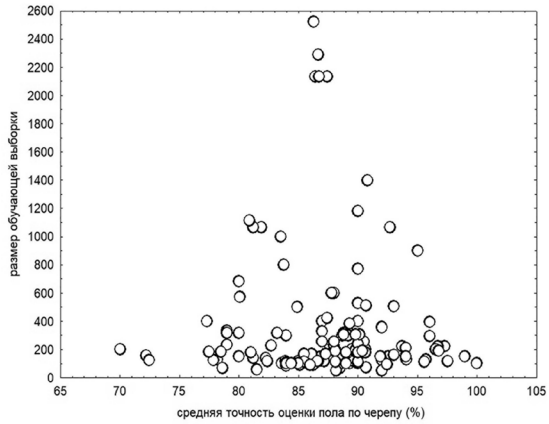

Рисунок 1. Размах изменчивости средней точности в оценке пола в зависимости от размера обучающей выборки (по результатам обзора 130 публикаций)
Figure 1. Variation in the mean accuracy in sex estimation depending on the size of the training sample (based on the results of a review of 130 publications)

на каждый пол приходится менее 100 черепов. Анализ публикаций показывает отчетливую зависимость между размером выборки и размахом изменчивости показателей точности алгоритмов (рис. 1). Чем больше черепов включает в себя обучающая серия, тем меньше дисперсия достигнутой точности вокруг общей средней оценки. Это может означать, что различия между оценками точности отчасти объясняются недостаточными размерами обучающей выборки, а не объективными преимуществами и недостатками конкретных моделей. При этом большие серии, как правило, имеют смешанный состав. Наблюдаемая картина пока не позволяет установить, возможно ли повысить эффективность модели за счет включения в обучающую выборку индивидов разного происхождения, но по крайней мере оно не влечет за собой очевидного снижения точности[2].

Как в обучающих, так и в тестовых выборках, число мужских черепов, как правило, несколько превышает число женских. Среднее соотноше-

[2] Это заключение может оказаться ложным, если детальный анализ покажет, что такие модели систематически лучше определяют пол в одних популяциях по сравнению с другими (даже в рамках обучающей выборки). Именно отсутствие такого анализа смущает в работах авторов, предлагающих универсальные методы оценки – средняя точность еще не отражает их реальной эффективности на практике.

ние составляет 54:46 и 58:42 соответственно. Очевидно, авторы большинства исследований исходят из того, что незначительное преобладание мужских черепов в обучающей выборке теоретически не должно заметно влиять на относительную точность оценок определения мужских и женских черепов. Однако, вероятно, это не совсем так. В частности, если при расчетах учитывать только данные публикаций, в которых черепа обоего пола представлены приблизительно в равном соотношении (по 49-50%), то средняя точность оценок мужских черепов окажется ниже на 3% (86.8% и 89.7% соответственно), причем наблюдаемые различия имеют статистически значимую величину (T-критерий Уилкоксона, n=22, p=0.008). Эти результаты лучше согласуются с тезисом К. Вайса о недооценке числа мужчин в палеодемографических выборках [Weiss, 1972], нежели средняя оценка точности, не учитывающая соотношение полов в обучающей выборке или оценки, полученные при помощи визуального подхода, отличающиеся чрезвычайной вариабельностью.

Однако состав и размер обучающей выборки сами по себе еще не объясняют исчерпывающе, почему гипотеза о преимуществе алгоритмов перед экспертами не находит подтверждения при решении задачи определения пола. Главное различие между настоящим исследованием и исследованиями, на которые опирается упомянутая гипотеза, заключается в качестве обрабатываемых данных. Во-первых, статистические модели и антропологи, определяющие пол, очевидно опираются на несколько различающиеся наборы признаков. Во-вторых, серьезное влияние на эффективность моделей могут оказывать межисследовательские расхождения в оценке исходных полодифференцирующих признаков, использованных для обучения и тестирования результатов. Алгоритмы превосходят людей при равном доступе и равном качестве информации, которые сложно обеспечить при сравнении визуального и статистического подходов.

Во многих публикациях, посвященных методам оценки пола по черепу, содержится раздел, в котором приводятся результаты оценки расхождений между исследователями в измерениях / визуальной оценке анализируемых признаков. Как правило, наблюдаемые различия рассматриваются как низкие или приемлемые. Однако проблема заключается в том, что ошибка в оценке исходных признаков сама по себе еще не позволяет оценить ее влияние на эффективность моделей. В некоторых случаях (особенно при использовании балловой системы оценки признаков), несмотря на существенное согласие между оценками исследователей, показатели точности моделей, примененных к одной и той же выборке, изученной разными авторами, могут заметно расходиться [Lewis, Garvin, 2016]. Причина этого, вероятно, заключается в разном весе признаков, учитываемых моделью: даже разница в один балл иногда может оказывать существенное влияние на результат. К сожалению, во многих работах влияние возможных внутри- и межисследовательских расхождений в оценке исходных признаков непосредственно на эффективность моделей.

Заключение

Средняя точность моделей для определения пола по черепу не имеет заметных преимуществ перед визуальной оценкой опытных антропологов – по крайней мере, в случае если череп сохранился полностью. Это заключение удовлетворит многих антропологов, и все же оно будет неполным и даже некорректным без следующих примечаний:

1. Несмотря на множество публикаций, посвященных методам оценки пола, сохраняется очевидный дефицит данных о надежности визуального подхода и влиянии на его точность опыта работы с коллекциями, степени сохранности черепа и других факторов изменчивости. Не менее актуальным остается проведение независимого тестирования статистических моделей с высокой заявленной точностью на материалах разного происхождения.

2. Статистические модели разных типов различаются по уровню точности и обладают специфическими недостатками. Качество моделей, основанных на анализе краниометрических признаков, в большей степени страдает из-за межпопуляционной изменчивости. Модели, опирающиеся на балловые признаки, сильнее зависят от межисследовательских расхождений. Методы геометрической морфометрии все еще редко

применяются и их эффективность пока не поддается надежной проверке. Модели, использующие дискриминантные функции и логистическую регрессию, менее эффективны, чем методы машинного обучения. Однако первые в отличие от последних неоднократно становились объектом независимого тестирования. При этом методы машинного обучения могут быть в большей степени склонны к переобучению, а проверка их эффективности (как и практическое применение) часто требует владения специальными техническими навыками со стороны исследователей.

3. Модели имеют более широкое применение в судебной антропологии, что обусловлено спецификой профессии. Судебный антрополог должен уметь показать, что методы, используемые им для определения пола, основаны на надежных принципах и методологии. Если результаты независимой экспертизы (например, анализа ДНК) противоречат его заключениям, это может поставить под сомнение доверие к квалификации исследователя и в дальнейшем использоваться для дискредитации других проведенных им анализов. Применение моделей позволяет строго обосновывать полученные результаты, включая вероятность ошибки и возможные недостатки методики [Williams, Rogers, 2006]. Для антрополога, имеющего дело с материалами археологических раскопок, вопросы обоснованности заключения и персональной ответственности воспринимаются как менее значимые, а случайные ошибки не оказывают существенного влияния на результаты анализа серийного материала.

И все же представляется, что требования указывать использованные методы, а также описывать признаки, на основе которых было составлено заключение о половозрастной характеристике индивидов, будут справедливыми для всех антропологов. Проблема метода особенно актуальна при изучении фрагментированных останков, где вероятность систематических ошибок особенно велика. В таких случаях читатель должен иметь возможность оценить обоснованность определений, иногда имеющих принципиальное значение для демографической и культурной характеристики населения. Но это требование важно соблюдать и при решении совершенно иных задач. Например, при оценке влияния пола и возраста на изменчивость интересующих исследователя признаков скелета (метрических или дискретных) более корректным представляется опираться не на половозрастные определения как таковые, а на набор формализованных значений признаков, которые легли в основу антропологического заключения.

Нет никаких сомнений, что в будущем роль алгоритмов при решении разнообразных задач в антропологии будет только возрастать. Это не означает, что сколь-нибудь значимая доля опубликованных двух сотен моделей определения пола найдет реальное практическое применение. Более вероятно, что изменится сам подход к возможностям использования статистических моделей. Специальные исследования показывают, что уровень недоверия к алгоритмам заметно снижается при демонстрации их способности обучаться на своих ошибках [Berger et al., 2021]. Именно внедрение гибких методик анализа, позволяющих исследователям самостоятельно контролировать как отбор признаков, так и сравнительные данные, используемые в качестве обучающей выборки, может привести к качественному повышению точности половозрастных определений по костям скелета.

Библиография

Алексеев В.П., Дебец Г.Ф. Краниометрия. Методика антропологических исследований. М.: Наука. 1964. 128 с.

Звягин В.Н. Методика краниоскопической диагностики пола человека // Судебно-медицинская экспертиза, 1983. №3. С.15-17.

Канеман Д., Сибони О., Санстейн К.Р. Шум. Несовершенство человеческих суждений. М.: АСТ. 2021. 544 с.

Информация об авторе

Широбоков Иван Григорьевич, к.и.н.;
ORCID ID: 0000-0002-3555-7509;
ivansmith@bk.ru

Поступила в редакцию 21.08.2023,
принята к публикации 25.08.2023.

Shirobokov I.G.

Peter the Great Museum of Anthropology and Ethnography (Kunstkamera) RAS, Department of Physical Anthropology, Universitetskaya emb., 3, Saint Petersburg, 199034, Russia

ON THE RELATIVE ACCURACY OF THE SKULL SEX ESTIMATION

Introduction. *There are no fewer than two hundred algorithms for sex estimation based on cranial morphology, relying on statistical analysis of non-metric, linear, angular traits, and their combinations. Nevertheless, many physical anthropologists prefer to rely on visual observations. The objectives of this research encompass exploring potential reasons behind the preference for a visual approach and conducting an analysis of the comparative effectiveness of visual and statistical methods for sex estimation.*

Materials and methods. *The study is grounded in an analysis of publications related to methods of sex estimation based on cranial traits, spanning the past 70 years. Comparison of accuracy estimates was conducted using non-parametric tests, considering differences in statistical methods, validation approaches (no validation, cross-validation, independent test), and variable types (non-metric traits, craniometry, geometric morphometrics).*

Results. *General reasons for skepticism towards algorithms include unrealistic expectations regarding their capabilities, greater susceptibility to errors by models compared to humans, lack of control over classification. However, algorithms generally surpass experts in predicting the target variable. The average accuracy of visual sex estimations based on cranial traits is slightly lower than the estimates of statistical models and exhibits noticeable variability. The accuracy of estimations made by experienced anthropologists is comparable to the average performance of models. Nevertheless, the effectiveness of algorithms significantly diminishes when applied to datasets originating from sources other than the training set, particularly when dealing with craniometric traits. In a substantial portion of studies, the size of the training datasets is insufficient for a reliable assessment of model effectiveness, and the sex distribution is skewed towards male skulls, leading to some inflation of the accuracy of their estimates. Model effectiveness can also decline due to errors in the evaluation of non-metric traits, and the assessment of inter-researcher discrepancies does not allow for an evaluation of their impact on model accuracy.*

Conclusion. *Despite an extensive bibliography, there remains a lack of data on both the accuracy of the visual approach to sex estimation and the reliability of models with claimed high effectiveness. The adoption of flexible methodologies enabling researchers to independently control both variable selection and the composition of the training set will help overcome algorithm aversion and enhance the quality of estimates.*

Keywords: skull morphology; sex determination; craniometry; non-metric traits; algorithm aversion

DOI: 10.32521/2074-8132.2023.4.069-080

References

Alekseev V.P., Debets G.F. *Kraniometriya. Metodika antropologicheskikh issledovanii* [Craniometry. Methodology of anthropological research]. Moscow, Nauka Publ., 1964. 128 p. (In Russ).

Zvyagin V.N. Metodika kranioskopicheskoi diagnostiki pola cheloveka [Methodology of cranioscopic diagnostics of human sex]. *Sudebno-meditsinskaya ehkspertiza* [Forensic Medical Expertise / Sudebno-Meditsinskaya Ekspertisa], 1983, 3, pp.15-17. (In Russ.).

Kahneman D., Sibony O., Sunstein C.R. Shum. *Nesovershenstvo chelovecheskikh suzhdenii* [Noise: A flaw in human judgement]. Moscow, AST Publ., 2021. 544 p. (In Russ.)

Berg G.E., Tersigni-Tarrant A. Sex and ancestry determination: assessing the “gestalt”. *Proceedings of the 66th Annual Meeting of the American Academy of Forensic Sciences*; 2014 Feb 17-22; Seattle, WA. Colorado Springs, CO, American Academy of Forensic Sciences, 2014. pp. 414-415.

Berger B., Adam M., Rühr A., Benlian A. Watch me improve – algorithm aversion and demonstrating the ability to learn. *Bus. Inf. Syst. Eng.*, 2021, 63, pp. 55-68. DOI: 10.1007/s12599-020-00678-5.

Burton J.W., Stein M., Jensen T.B. A systematic review of algorithm aversion in augmented decision making. *J. Behavioral Decision Making*, 2020, 33 (2), pp. 220-239. DOI: 10.1002/bdm.2155.

Camps F.E. *Gradwohl's legal medicine. Identification by the skeletal structures*. 2nd ed. Bristol, John Wright & Sons Ltd., 1968, pp. 123-140.

Dawes R., Faust D., Meehl P. Clinical versus actuarial judgment. *Science*, 1989, 243 (4899), pp. 1668-1674. DOI: 10.1126/science.2648573.

Del Bove A., Veneziano A. A generalised neural network model to estimate sex from cranial metric traits: a robust training and testing approach. *Applied Sciences*, 2022, 12, 9285. DOI: 10.3390/app12189285.

Dereli A.K., Zeybek V., Sagtas E., Senol H., Ozgul H.A. et al. Sex determination with morphological characteristics of the skull by using 3D modeling techniques in computerized tomography. *Forensic Sci. Med. Pathol.*, 2018, 14, pp. 450-459. DOI: 10.1007/s12024-018-0029-0.

Dietvorst B.J., Simmons J.P., Massey C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J.Exper. Psych.: General*, 2015, 144 (1), pp. 114-126. DOI: 10.1037/xge0000033.

Đurić M., Rakočević Z., Đonić D. The reliability of sex determination of skeletons from forensic context in the Balkans. *Forensic Sci. Int.*, 2005, 147 (2-3), pp. 159-164.

Giles E., Elliot O. Sex determination by discriminant function analysis of crania. *Am. J. Phys. Anthropol.*, 1963, 21 (1), 53-68. DOI: 10.1002/ajpa.1330210108.

Henke W. On the method of discriminant function analysis for sex determination of the skull. *J. Hum. Evol.*, 1977, 6 (2), pp. 95-100. DOI:10.1016/S0047-2484(77)80111-5.

Hrdlička A. *Practical anthropometry.* Philadelphia, The Wistar Institute of Anatomy and Biology, 1939. 231 p.

Kelley S.R., Tallman S.D. Population-inclusive assigned-sex-at-birth estimation from skull computed tomography scans. *Forensic Science*, 2022, 2, pp. 321-348. DOI: 10.3390/forensicsci2020024.

Krogman W.M., İşcan M.Y. *The human skeleton in forensic medicine*. Springfield, IL, Charles C. Thomas, 1986. 576 p.

Lewis C.J., Garvin H.M. Reliability of the Walker cranial nonmetric method and implications for sex estimation. *J. Forensic Sci.*, 2016, 61 (3), pp. 743-751. DOI: 10.1111/1556-4029.13013.

Meindl R.S., Lovejoy C.O., Mensforth R.P., Carlos L.D. Accuracy and direction of error in the sexing of the skeleton: Implications for paleodemography. *Am. J. Phys. Anthropol.*, 1985, 68 (1), pp. 79-85. DOI:10.1002/ajpa.1330680108.

Nakhaeizadeh S., Dror I.E., Morgan R.M. Cognitive bias in sex estimation: The influence of context on forensic decision-making. *Sex Estimation of the Human Skeleton*, 2020, pp. 327-342. DOI:10.1016/b978-0-12-815767-1.00020-1.

Renier L.A., Schmid Mast M., Bekbergenova A. To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, 2021, 124, 106879. DOI: 10.1016/j.chb.2021.106879.

Spradley M.K., Jantz R.L. Sex estimation in forensic anthropology: skull versus postcranial elements. *J. Forensic Sci.*, 2011, 56 (2), pp.289-296. DOI:10.1007/ 978-1-59745-099-7_9.

Stewart T.D. Sex determination of the skeleton by guess and by measurement. *Am. J. Phys. Anthropol.*, 1954, 12 (3), pp. 385-392. DOI: 10.1002/ajpa.1330120312.

Stewart T.D. *Essentials of forensic anthropology*. Springfield IL, Charles C. Thomas, 1979. 317 p.

Thomas R.M., Parks C.L., Richard A.H. Accuracy rates of sex estimation by forensic anthropologists through comparison with DNA typing results in forensic casework. *J. Forensic Sci.*, 2016, 61 (5), pp. 1307-1310. DOI: 10.1111/1556-4029.13137.

Walker P.L. Sexing skulls using discriminant function analysis of visually assessed traits. *Am. J. Phys. Anthropol.*, 2008, 136 (1), 39–50. DOI: 10.1002/ajpa.20776.

Walrath D.E., Turner P., Bruzek J. Reliability test of the visual assessment of cranial traits for sex determination. *Am. J. Phys. Anthropol.*, 2004, 125 (2), pp. 132-137. DOI: 10.1002/ajpa.10373.

Weiss K.M. On the systematic bias in skeletal sexing. *Am. J. Phys. Anthropol.*, 1972, 37 (2), pp. 239-249. DOI: 10.1002/ajpa.1330370208.

Williams B.A., Rogers T.L. Evaluating the accuracy and precision of cranial morphological traits for sex determination. *J. Forensic Sci.*, 2006, 51, pp. 729-735. DOI: 10.1111/j.1556-4029.2006.00177.x.

Information about Author

Shirobokov Ivan Grigorievich, PhD.;
ORCID ID: 0000-0002-3555-7509; ivansmith@bk.ru.