

УДК 574.9

ПРЕИМУЩЕСТВА И ОГРАНИЧЕНИЯ МЕТОДОВ ЭКОЛОГИЧЕСКОГО МОДЕЛИРОВАНИЯ АРЕАЛОВ. 2. MAXENT

© 2020 г. А. А. Лисовский¹, *, С. В. Дудов², **

¹Зоологический музей Московского государственного университета им. М.В. Ломоносова
Большая Никитская ул., 2, Москва, 125009 Россия

²Московский государственный университет им. М.В. Ломоносова, биологический факультет, кафедра геоботаники
Ленинские горы, 1, стр. 12, Москва, 119234 Россия

*E-mail: andlis@zmmu.msu.ru

**E-mail: serg.dudov@gmail.com

Поступила в редакцию 05.06.2019 г.

После доработки 22.11.2019 г.

Принята к публикации 17.12.2019 г.

Среди методов экологического моделирования наибольшую популярность приобрел метод максимальной энтропии, реализованный в программе MaxEnt. Несмотря на частоту его применения, многие исследователи недооценивают влияние исходных параметров и естественного смещения выборки данных о локализации видов на результаты анализа. В то же время выбор типа функций предикторов и параметра сложности существенно влияет на пространственную “компактность” или “сглаженность” модели. Неслучайное распределение точек регистрации организмов в географическом пространстве требует обязательного введения поправок. В случае если планируется изучать влияние факторов среды на формирование ареала, необходимо минимизировать корреляции предикторов.

DOI: 10.31857/S0044459620020049

Из множества используемых алгоритмов моделирования пространственного распределения живых организмов чаще всего используется метод максимальной энтропии, реализованный в программе MaxEnt (Phillips et al., 2006; Phillips, Dudik, 2008). MaxEnt — это алгоритм машинного обучения, который предсказывает присутствие вида в географическом пространстве, основываясь только на точках регистрации видов (presence-only), без учета мест документированного отсутствия. Описание принципов работы MaxEnt приведено в ряде работ авторов алгоритма (Phillips et al., 2004, 2006, 2017), а также в инструкции пользования программой (Phillips et al., 2019). Существует также несколько обзорных статей, где обсуждаются как сами принципы расчетов MaxEnt, так и влияние входящих параметров на результат моделирования (Elith et al., 2010, 2011; Merow et al., 2013; Guillera-Arroita et al., 2015).

В феврале 2019 г. на запрос “species distribution modeling MaxEnt” сервис “Google Академия” показывал 18.6 тыс. ссылок, а на ресурсе “Web of Science” содержалось 2318 статей, из которых 370 опубликовано в 2018 г. Среди причин популярности алгоритма, выраженной в стремительном росте публикаций, — удобный графический интерфейс, относительная простота использования и

наглядное представление результата. Однако удобство имеет и обратную сторону. Используя MaxEnt с установленными по умолчанию настройками, мы можем получить “картинку”, пригодную для иллюстрирования статей. В результате многими пользователями алгоритм воспринимается как “волшебная палочка”, после взмаха которой магическим образом “появляется” карта ареала. Однако неоднократно высказывалось мнение, что предустановленные параметры не всегда позволяют получить оптимальную модель (Shcheglovitova, Anderson, 2013; Syfert et al., 2013; Radosavljevic, Anderson, 2014; Morales et al., 2017). Понимание принципов и ограничений алгоритма, его применимости для решения конкретных проблем является обязательным условием для грамотной постановки научной задачи. В этой работе мы постараемся обратить внимание на влияние изменяемых параметров MaxEnt на результат моделирования и дадим рекомендации по возможным путям решения основных проблем, возникающих на разных стадиях исследования.

1. ОБЩИЕ ПРИНЦИПЫ РАБОТЫ

Суть метода сводится к поиску закономерности распределения значений факторов среды в

точках, где доказано обитание вида. В качестве входных данных используются координаты точек регистрации вида (ТР; samples — здесь и далее англоязычные термины в скобках соответствуют названию параметров MaxEnt) и предикторы (растровые географические данные; о растровых данных см. Дубинин, Костикова, 2008), описывающие пространственную изменчивость факторов среды на всей территории исследования (один растровый слой — один фактор среды; environmental layers). В начале анализа производится случайный отбор ограниченного числа (max number of background points) ячеек раstra (фоновые точки — ФТ; background points) со всей анализируемой территории. Значения факторов среды в ФТ будут использованы в дальнейшем анализе. При настройках по умолчанию наличие или отсутствие ТР в ячейке игнорируется при случайном выборе ФТ — все ячейки имеют равный вес при отборе (add samples to background). Хотя отбор ФТ случаен, его можно регулировать использованием растрового файла коррекции (bias file), отношение значений в ячейках которого показывает, насколько вероятность отбора одной ячейки будет больше, чем другой. Значения факторов среды в ТР и ФТ используются для обучения модели.

Выборки значений факторов среды в ФТ и ТР трансформируются в “функции предикторов” (ФП; features), которые используются в дальнейшем анализе в качестве переменных. Это простые математические преобразования исходных значений факторов среды: линейная (непреобразованные факторы среды; linear), квадратичная (квадрат значений факторов среды; quadratic), множественная (произведение двух факторов среды; product), бинарная пороговая (бинарная функция от фактора среды со значением 0, если значения фактора ниже определенного порога, и 1, если выше; threshold) и линейная пороговая (непреобразованные факторы среды, но все значения ниже определенного порога заданы константой; hinge) (Elith et al., 2011; Merow et al., 2013). Количество ФП, которые будут использованы в анализе, зависит от имеющегося объема материала (ТР). При объеме материала меньше 10 ТР могут быть использованы только линейные ФП; от 10 до 15 — линейные и квадратичные; от 15 до 80 — линейные, квадратичные и линейные пороговые; больше 80 — все типы ФП (Phillips, Dudik, 2008; Merow et al., 2013).

Итогом работы MaxEnt является расчет нормализованной экспоненциальной функции, в которую в качестве аргумента входит сумма всех ФП, каждый из которых имеет свой коэффициент λ (Elith et al., 2011; Merow et al., 2013; Phillips et al., 2017). Эти коэффициенты рассчитываются в процессе работы алгоритма (чем больше λ , тем больше “вклад ФП в модель”). Суть алгоритма описыва-

ют несколькими разными способами (Elith et al., 2011; Merow et al., 2013). С одной стороны, производится максимизация энтропии в пространстве, т.е. проводится поиск наиболее равномерного географического распределения предсказанного присутствия вида (Phillips, Dudik, 2008; Merow et al., 2013). С другой — максимизируется сходство между ожидаемым (prior) и предсказанным (prediction) распределением относительной вероятности присутствия вида в пространстве ФП (Elith et al., 2011; Merow et al., 2013). Статистическое объяснение алгоритма было неоднократно опубликовано (Elith et al., 2011; Merow et al., 2013; Phillips et al., 2017). Для “интуитивного” понимания сути происходящего, вероятно, нужно добавить, что на ФП накладываются дополнительные ограничения, зависящие от их типа. Так, для линейной ФП среднее значение в обучающей выборке должно быть максимально близким к среднему значению этой ФП в предсказанном диапазоне распространения, для квадратичной ФП ограничивающим параметром является дисперсия, для множественной ФП — ковариация.

Для решения задачи в MaxEnt ищется максимум функции прироста (gain function), состоящей из уменьшаемого (оценка предсказанных относительных вероятностей присутствия в ТР; неизвестны коэффициенты λ при ФП) и двух вычитаемых: оценки предсказанных относительных вероятностей присутствия в ФТ (неизвестны коэффициенты λ при ФП) и контроля сложности модели (неизвестны коэффициенты λ при ФП и параметр сложности β ; см. ниже).

Выбор оптимальной модели производится пошагово, число шагов (maximum iterations) по умолчанию установлено равным 500. Это значение чаще всего подходит только для простых моделей или для оценочного анализа. Для сложных моделей с множеством факторов значение параметра требуется повышать. Очевидно, что мы не способствуем поиску оптимальной модели, прерывая анализ до его завершения.

Полученная модель (упомянутая выше нормализованная экспоненциальная функция) может быть применена (спроецирована; project) к любым территориям или временным периодам при условии наличия для них тех же самых предикторов, естественно, с теми же названиями файлов (projection layers).

2. ВХОДНЫЕ ПАРАМЕТРЫ

Существуют несколько параметров, существенно влияющих на ход или результат моделирования, значения которых могут быть изменены при запуске MaxEnt. Так, по умолчанию MaxEnt предлагает автоматически выбрать набор используемых ФП. Тип ФП существенно влияет на ре-

зультаты анализа, поскольку позволяет учитывать или не учитывать нелинейные взаимоотношения между распространением вида и параметрами среды (Austin, 2002). В некоторых исследовательских программах тип ФП может быть содержательно важен (Elith et al., 2011; Merow et al., 2013). Кроме того, использование разных типов ФП по-разному отражается на времени вычислений. Например, при использовании в качестве предикторов 19 слоев климатической модели WorldClim (Hijmans et al., 2005) для 100 ТР рассчитываются по 19 линейных и квадратичных, 171 множественная, 5643 бинарных и линейных пороговых ФП (Merow et al., 2013).

Контроль сложности оказывает непосредственное влияние на выбор оптимальной модели. По умолчанию каждый тип ФП имеет свой коэффициент (λ) для разных объемов выборки ТР (Phillips, Dudik, 2008; Phillips et al., 2019). Например, линейная ФП для 10 ТР имеет $\lambda = 1$, а квадратичная для 100 ТР — $\lambda = 0.05$. Эти коэффициенты выбраны авторами MaxEnt эмпирически. Соответственно, благодаря разным коэффициентам разные ФП вносят разный вклад в искомую функцию. Чем меньше в среднем коэффициенты λ , тем больше локальных “островов” может появиться на результирующей карте за счет локального вклада единичных ФП. Такие “сложные” модели имеют тенденцию очерчивать существующие ТР, а не искать общие закономерности пространственного распределения. Чем больше коэффициент, тем сильнее некая ФП оказывает влияние на модель, что приводит к ее большей пространственной сглаженности. В настройках MaxEnt существует параметр сложности β (regularization multiplier), на который умножается каждый коэффициент λ при расчете второго вычитаемого в gain функции. По умолчанию значение β установлено равным единице; такое значение эмпирически подобрано авторами MaxEnt при тестировании моделей на разных группах организмов (Elith et al., 2011; Merow et al., 2013; Phillips et al., 2019). Уменьшение значения параметра сложности приводит к усложнению модели, увеличение — к упрощению за счет изменения веса параметризации в gain функции. По мнению многих исследователей, имеет смысл опытным путем подбирать параметр сложности в ходе исследования (Anderson, Gonzalez, 2011; Elith et al., 2011; Warren, Seifert, 2011; Merow et al., 2013; Shcheglovitova, Anderson, 2013; Radosavljevic, Anderson, 2014; Phillips et al., 2019), выбирая модель с наилучшим качеством (см. раздел 4) среди построенных с разными значениями этого параметра.

3. РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

Результатом работы алгоритма является растровый слой с итоговым показателем “присут-

ствия” вида в каждой ячейке (файл вид.asc); файл с основными результатами расчетов (maxentResults.csv); файл со значениями коэффициентов λ и минимальных/максимальных значений всех ФП, использованных в анализе (вид.lambdas); файл, описывающий форму кривых распределения предсказанных значений по площади раstra и долю неверно предсказанных присутствий по предсказанным значениям (вид_omission.csv) и ряд визуализированных графиков и т.п., в том числе “карта ареала”, имеющая наглядное представление с градиентной цветовой заливкой. Одним из результатов является набор рассчитанных пороговых значений (thresholds), необходимых для перевода вероятностной модели распространения вида в бинарный вид со значениями: 1 — вид присутствует и 0 — вид отсутствует.

Пространственные выходные данные (outputs) реализованы в нескольких формах, которые имеют разные диапазоны варьирования, могут отличаться визуально, но при этом монотонно соотносятся между собой. Простой (raw) результат содержит в ячейках показатель относительной вероятности присутствия вида (Merow et al., 2013). Сумма значений всех ячеек равна единице, поэтому величины в каждой ячейке очень малы и зависят от числа ячеек. В накопленном (cumulative) представлении результата ячейке присваивается сумма всех “простых” (raw) значений, меньших или равных значению данной ячейки, и пересчитывается в диапазоне от 0 до 100. Логистическое logistic и дважды логарифмическое cloglog представления результата являются нормализацией простого результата, точнее его перевод в более привычный бинарный вид (есть вид/нет вида), с плавным переходом между двумя состояниями (Phillips, Dudik, 2008; Phillips et al., 2017). В нормализации участвует параметр τ (распространенность; default prevalence). Смысл этого параметра сводится к вероятности присутствия вида в каждой ячейке изучаемого раstra. Другими словами, можно описать распространенность как долю ячеек, в которых ожидается наличие вида. Установленное по умолчанию значение τ , равное 0.5, подразумевает, что вид обитает на половине анализируемой территории, что редко соответствует действительности. Используя любое произвольное значение распространенности (в том числе и установленное по умолчанию), мы искажаем “нормализованную” форму вывода; на другие формы вывода это не влияет. Сам MaxEnt, как и все алгоритмы анализа присутствия (Лисовский и др., 2020), не может рассчитать значение τ (Ward et al., 2009; Elith et al., 2011; Merow et al., 2013; Guillera-Aroita et al., 2015).

По мнению ряда авторов, простой результат MaxEnt отражает показатель относительного присутствия вида, поэтому может быть использован при расчете встречаемости или пригодности

местообитаний. Этот тип результата пропорционален вероятности присутствия вида при соблюдении ряда дополнительных условий (выявляемость вида одинакова во всей изучаемой выборке, ТР распределены случайно) (Merow et al., 2013; Guillera-Aroita et al., 2015). Накопленный результат наиболее пригоден при поиске границ распространения видов. Логистический результат, хотя и наиболее заманчив для использования в силу интуитивно интерпретируемого нормализованного распределения, основан на произвольном значении τ , поэтому его не рекомендуется использовать во всех случаях, когда ареал будет использован не в качестве иллюстрации, а для дальнейших вычислений (Elith et al., 2011; Merow et al., 2013). Необходимо заметить, что авторы алгоритма в разных работах трактуют логистический результат как вероятность присутствия вида (Phillips, Dudik, 2008; Phillips et al., 2017), что не может быть верно для моделей, основанных на анализе присутствия (Guillera-Aroita et al., 2015).

Поскольку информация об отсутствии вида недоступна, значения в ячейках простой результирующей модели MaxEnt линейно связаны с вероятностью присутствия вида, но коэффициенты этой линейной регрессии неизвестны (Ward et al., 2009; Elith et al., 2011; Guillera-Aroita et al., 2015). Это порождает задачу поиска точки пересечения этой линейной модели с осью абсцисс, т.е. вычисление порогового значения в результирующей модели, которое соответствует нулевому присутствию (отсутствию) вида в природе. Критике самой возможности предсказать пороговые значения, анализируя только данные о присутствии видов, посвящен ряд статей (Liu et al., 2013; Merow et al., 2013; Guillera-Aroita et al., 2015). Авторы этих работ единодушны в сомнительности использования порогов в исследовательских целях, хотя и находят, что некоторые расчетные пороговые значения оказываются ближе к ожидаемым величинам, чем другие. Параллельно идет дискуссия о прогностическом качестве разных типов пороговых значений, рассчитываемых MaxEnt (Braunisch, Suchant, 2010; Liu et al., 2013; Merow et al., 2013). Вероятно, использование пороговых значений допустимо для решения ограниченного круга задач, например для создания генерализованных иллюстраций с обязательной оговоркой об условности изображаемых границ распространения.

4. ОЦЕНКА КАЧЕСТВА И ПОИСК ОПТИМАЛЬНОЙ МОДЕЛИ

Вероятностный подход к построению ареалов, наличие параметров, значения которых необходимо рассчитать в процессе анализа, разные результаты моделирования при варьировании входящих параметров — все это требует наличия алгоритма оценки качества и выбора оптимальной

модели. Стоит различать эти два понятия. Оценка качества модели предполагает ответ на вопрос “Хороша ли моя модель?” при получении первичных результатов. Выбор оптимальной модели требуется, когда существует набор моделей, построенных по одним и тем же данным, но с различными параметрами.

Первичная оценка качества моделей в большинстве случаев проводится экспертным способом. Хотя эксперт может владеть полной опубликованной информацией о распространении и плотностях распределения вида, общая картина в его голове складывается из отдельных точек регистрации, без информации между ними и далеко за их пределами. Поэтому экспертное мнение в данном случае остается лишь первичным методом оценки. Оптимальным способом оценки качества, вероятно, является полевое тестирование модели, но прежде, чем планировать полевые работы, следует убедиться, что модель достаточно “хороша” с точки зрения возможностей моделирования.

Базовой мерой оценки качества модели в MaxEnt является площадь под ROC-кривой — AUC (area under receiver operating characteristic (ROC) curve). Этот показатель прогностической способности интерпретируется как вероятность того, что случайно выбранные ТР предсказаны лучше, чем случайно выбранные ФТ. Для построения ROC-кривой по оси абсцисс откладывается $(1 - \text{специфичность})$ (specificity, доля верно предсказанных отсутствий), а по оси ординат — чувствительность (sensitivity, доля верно предсказанных присутствий $\text{TP} = 1 - \text{omission rate}$); соответственно обе оси имеют диапазон варьирования от 0 до 1, а площадь под кривой, проходящей в этих осях, не превышает 1. Случайное распределение результатов моделирования соответствует значению $\text{AUC} = 0.5$. Специфичность не может быть корректно оценена в MaxEnt, поскольку информация об отсутствии видов не используется. Набор точек отсутствия заменен случайной выборкой из ФТ (Phillips et al., 2006). По значению AUC качество моделирования можно условно разделить на пять категорий (Araújo et al., 2005): 0.9–1 — “отлично”, 0.8–0.9 — “хорошо”, 0.7–0.8 — “удовлетворительно”, 0.6–0.7 — “плохо”, <0.6 — “очень плохо” (моделирование не удалось).

По результатам экспериментов значение AUC чувствительно к изменению разных входящих данных и параметров (Warren, Seifert, 2011; Fourcade et al., 2018). Например, эта величина оказывается выше у более сложных моделей при прочих равных условиях, т.е. значение AUC зависит от использованных ФП (рис. 1), параметра сложности и т.п. Поэтому в разных работах обсуждаются другие способы оценки качества моделей. Один из способов, реализованных в MaxEnt, заключа-

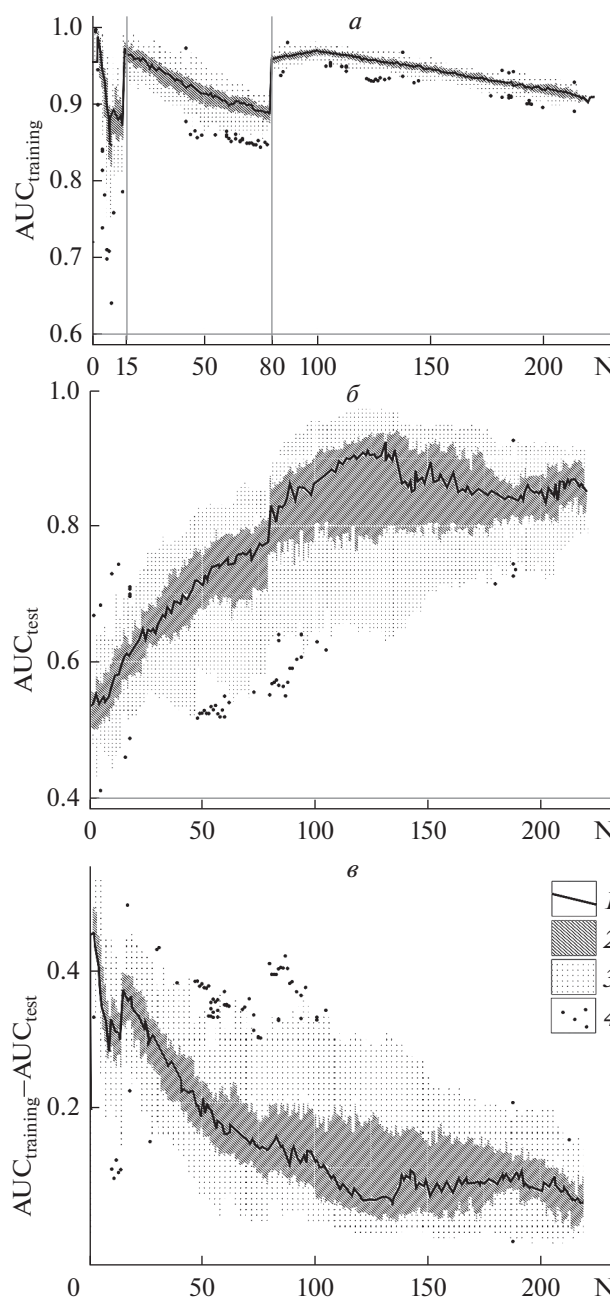


Рис. 1. Изменение разных критериев оценки качества (AUC_{training} , AUC_{test} , $AUC_{\text{training}} - AUC_{\text{test}}$) моделей пригодности местообитаний 12 видов птиц Московского региона с ростом количества точек регистрации (N). Параметры: предикторы, климатическая модель WorldClim (Hijmans et al., 2005), месячные композиты видимого и ближнего инфракрасного диапазона спутникового радиометра MODIS за девять месяцев 2004 г. (19 и 63 растровых слоя соответственно), выборка из 10000 фоновых точек, автоматический выбор ФП, параметр сложности 1, обучающий и тестовый набор данных независимы. Условные обозначения в легенде: 1 – среднее, 2 – нижний и верхний квартили, 3 – минимальное и максимальное значение выборки, 4 – “выскакивающие значения”. *a* – изменение AUC_{training} ; при автоматическом выборе ФП при достижении выборкой ТР объема 15 и 80 происходит повышение сложности модели за счет включения в анализ новых типов ФП, в этих точках происходит скачкообразный рост AUC_{training} . *б* – изменение AUC_{test} ; объем тестовой выборки постоянен; нарастание сложности модели происходит за счет увеличения обучающей выборки; таким образом, график показывает изменение AUC, происходящее только за счет усложнения модели. *в* – изменение $AUC_{\text{training}} - AUC_{\text{test}}$.

ется в сравнении результатов анализа обучающей (AUC_{training}) и тестовой выборки (AUC_{test}). Мера AUC_{test} , рассчитываемая на независимом наборе данных, менее чувствительна к параметризации (Warren, Seifert, 2011). Разница $AUC_{\text{training}} - AUC_{\text{test}}$, таким образом, лучше отражает изменение качества модели, чем AUC_{training} . Существует дополнительная проблема – средствами MaxEnt, используя случайные подвыборки любых типов (replicated run type), крайне сложно получить пространственно независимые обучающую и тестовую выборки. Для получения независимых наборов данных предлагается использовать пространственную кросс-валидацию (k-fold cross validation или spatial jackknifing) (Hijmans, 2012; Shcheglovitova, Anderson, 2013; Radosavljevic, Anderson, 2014). Суть метода заключается в разбиении района исследований на несколько участков и использовании каждого из них в качестве тестовой выборки отдельно. При этом лучшие результаты получаются, если во время обучения модели исключать из анализа территорию, с которой берется тестовая выборка (Radosavljevic, Anderson, 2014). Такой метод реализован в среде R (Muscarella et al., 2014; Naimi, Araújo, 2016; Cobos et al., 2019) и в виде надстройки в ArcGIS Esri Inc. (Brown, 2014). По мнению Уоррена и Сиферта (Warren, Seifert, 2011), байесовы информационные критерии AIC_c и BIC лучше подходят для оценки качества моделей, чем AUC ; для их расчета создана программа ENMTools (Warren et al., 2010), а также пакеты для R: kuenm (Cobos et al., 2019) и Biomod 2 (Thuiller et al., 2019).

Анализ случайных подвыборок из исходных ТР (в MaxEnt реализованы три алгоритма: кроссвалидация, бутстрэп и сабсэмплинг) полезен не только для сравнения AUC_{training} и AUC_{test} , но и для оценки воспроизводимости модели. Отсутствие повторяемого результата говорит о нехватке или пространственном смещении данных (ТР). Исходные ТР часто бывают смещены относительно их реального положения в пространстве вследствие неверного определения их местоположения (например, определение места по коллекциям зоологических музеев или гербариев), недостаточной квалификации сборщика данных (например использование любительских баз данных), изменения местообитаний с момента сбора материала и т.п. (Fourcade et al., 2014). Поиск и отбраковка (или коррекция, если это возможно) “выскакивающих” ТР (outliers) возможны в процессе анализа подвыборок. Сравнение показателей качества моделей и содержимого файлов вид_samplePredictions.csv (содержащего предсказанные величины относительного присутствия для всех ТР; у “вылетающих” за пределы предсказанного распространения ТР эти значения минимальны)

позволит выявить список ТР, негативно влияющих на общее качество модели.

Использование анализа подвыборок для поиска оптимальной модели (а не оптимизации списка ТР), т.е. выбор итоговой модели из числа лучших результатов анализа подвыборок, едва ли оправданно, поскольку уменьшает и без того недостаточное количество ТР (Phillips et al., 2004), но не решает других проблем наборов ТР (см. раздел 5). Поиск оптимальной модели требует подбора значений параметра сложности β , а в ряде случаев и типа ФТ (Warren, Seifert, 2011; Shcheglovitova, Anderson, 2013). Для изучения оптимальной параметризации модели предложено контролировать изменения AUC вместе с изменением доли неверно предсказанных присутствий (omission rate). Предполагается, что у излишне параметризованных моделей часть ТР не будет попадать в предсказанную область, таким образом, доля ошибок будет выше нуля (Shcheglovitova, Anderson, 2013; Radosavljevic, Anderson, 2014). Долю ошибок можно оценить по предлагаемым MaxEnt величинам пороговых значений, связанных с чувствительностью (minimum training presence, 10 percentile training omission).

Как видно из текста этого раздела, методику поиска оптимальной модели нельзя считать устоявшейся. Разные предложенные критерии оценки качества имеют свои недостатки (объективные или связанные со сложностью расчетов). На настоящем этапе при сравнении набора моделей целесообразно сравнивать результаты нескольких тестов, выбирая максимальное (величины, связанные с AUC , не забывая, что у более сложных моделей AUC будет неизбежно больше) или минимальное (информационные критерии, omission rate) значение для оптимальной модели.

5. ВХОДНЫЕ ДАННЫЕ: ТОЧКИ РЕГИСТРАЦИИ ВИДОВ

Анализ присутствия в MaxEnt подразумевает случайное распределение ТР (Phillips et al., 2009; Elith et al., 2011; Guillera-Aroita et al., 2015). Чаще всего ТР расположены в пространстве не случайно: они привязаны к элементам инфраструктуры (дорогам, биостанциям, заповедникам и т.п.) и, наоборот, редки в малонаселенных районах. Таким образом, в исходных данных часто существуют агрегации или пространственные кластеры. Искусственное сгущение и разрежение ТР неизбежно приводит к смещению результатов анализа. Для того чтобы избежать этого эффекта, можно бороться с пространственной автокорреляцией путем фильтрации и пространственного разделения данных, а также менять вероятность выбора ФТ путем создания слоя коррекции (bias file) (Hijmans, 2012; Boria et al., 2014; Fourcade et al., 2014; Radosavljevic, Anderson, 2014; Fei, Yu, 2016).

Способы создания слоя коррекции рассмотрены в разделе 6, методы фильтрации ТР перечислены ниже. Агрегации ТР следует разреживать с определенным шагом. Его размер определяют масштабом исследования (Fourcade et al., 2014; Velazco et al., 2017), вычисляют на основе меры пространственной автокорреляции — индекса Морана (Rangel et al., 2006; Nuñez, Medley, 2011) или подбирают экспериментально (Kramer-Schadt et al., 2013). Технически для этого можно использовать штатные средства ГИС-пакетов, например “Delete Identical” в ArcGIS 10.3, инструмент “Points Filter” в SAGA (Conrad et al., 2015), пакет “spThin” в R (Aiello-Lammens et al., 2015) или реализованный в пакете SDMtoolbox инструмент, позволяющий варьировать шаг разреживания в зависимости от гетерогенности природных условий (Brown, 2014). При значительных различиях в плотности сбора данных на большой территории рекомендуется раздельно анализировать территории с более или менее гомогенной изученностью (Fourcade et al., 2014).

Неизбежно возникает вопрос, сколько ТР требуется для построения модели, какой объем данных является оптимальным. При этом стоит учитывать, что наборы ТР обычно малы по меркам машинного обучения (Phillips et al., 2004). Как указывалось выше (раздел 1), число ТР в MaxEnt прямо связано с возможностями использования разных ФП (Elith et al., 2011), и при выборках меньше 80 ТР нельзя ожидать расчета сложной модели. Малые выборки затруднительно использовать для алгоритмов анализа случайных подвыборок ТР (subsample, crossvalidate, bootstrap) из-за ограничений в объеме данных (Bean et al., 2012). MaxEnt способен работать с выборками больше двух экземпляров, однако малое число ТР ограничивает интерпретацию модели только как территории, где условия позволяют существовать виду, и не дает оснований для описания пределов распространения (Pearson et al., 2007). В то же время показано, что при калибровке модели можно достичь удовлетворительного результата и при малых (от 20–25 ТР) выборках (Shcheglovitova, Anderson, 2013; Morales et al., 2017).

Для контроля пространственного смещения данных (см. раздел 4) полезно фиксировать точность сбора данных. В качестве меры точности можно использовать, например, радиус окружности, в который попадают все возможные варианты локализации ТР. При фиксации ТР при помощи бытового GPS-навигатора точность будет составлять около 25–50 м, при привязке этикеток музеев и гербариев она может падать до нескольких километров и более. Влияние пространственного смещения на результаты моделирования пока еще недостаточно изучено, но в целом ясно, что чем ниже точность сбора данных, тем больше вероятность того, что ТР попадут не в свою ячейку

растровых данных и “присутствие” окажется ложным. Соответственно, чем мельче ячейка растра, тем точнее должны быть собраны ТР.

6. ВЫБОРКА ФОНОВЫХ ТОЧЕК

При настройках по умолчанию MaxEnt формирует случайную выборку из 10000 (max number of background points) ФТ. Возникает вопрос о репрезентативности выборки из 10000 ФТ для описания любой территории. Например, на территории европейской части России (ЕЧР) помещается более двух миллионов ячеек размером 2×2 км. Таким образом, при запуске программы с установленным по умолчанию числом ФТ в анализе будет использовано менее 0.5% ячеек. С другой стороны, если в выборку ФТ попадает все разнообразие ландшафтов, то 10000 — более чем достаточная цифра для анализа свойств распределения природных факторов.

Вопрос о том, может ли 0.5% данных репрезентативно отражать все разнообразие сочетаний природных факторов на изучаемой территории, не имеет однозначного ответа и должен решаться в рамках каждого конкретного исследования. Некоторые исследователи предполагают (Renner, Warton, 2013), что установленное по умолчанию число фоновых точек недостаточно для репрезентативного анализа вариативности природных факторов на обширной территории. При этом увеличение этого числа приводит к заметному увеличению времени расчетов и объемов занимаемой памяти. В некоторых работах число фоновых точек составляет 75–300 тысяч (El-Gabbas, Dormann, 2018).

Не менее важно, что анализ случайной выборки фоновых точек (см. раздел 1) относительно неслучайной выборки ТР может заметно смещать результаты анализа (Merow et al., 2013; Fourcade et al., 2014; Guillera-Aroita et al., 2015). Одним из решений обеих проблем (смещенности и репрезентативности выборки) является создание слоя коррекции (Fourcade et al., 2014). Такой растровый слой должен иметь в точности такой же географический охват и проекцию, как слой-предикторы. Отношение значений, присвоенных каждой ячейке этого растра, указывает на отношение вероятности выбора этих ячеек при формировании выборки ФТ (Merow et al., 2013). Например, если всем ячейкам присвоить значение 1, то они будут иметь равный вес при формировании выборки, которая останется случайной. Если некоторым ячейкам присвоить значение 10, то вероятность их отбора повысится в 10 раз. Таким образом, если количество ячеек с повышенными значениями вероятности отбора окажется больше порогового числа анализируемых ячеек, состав ячеек, из которых будет формироваться выборка, будет полностью контролироваться.

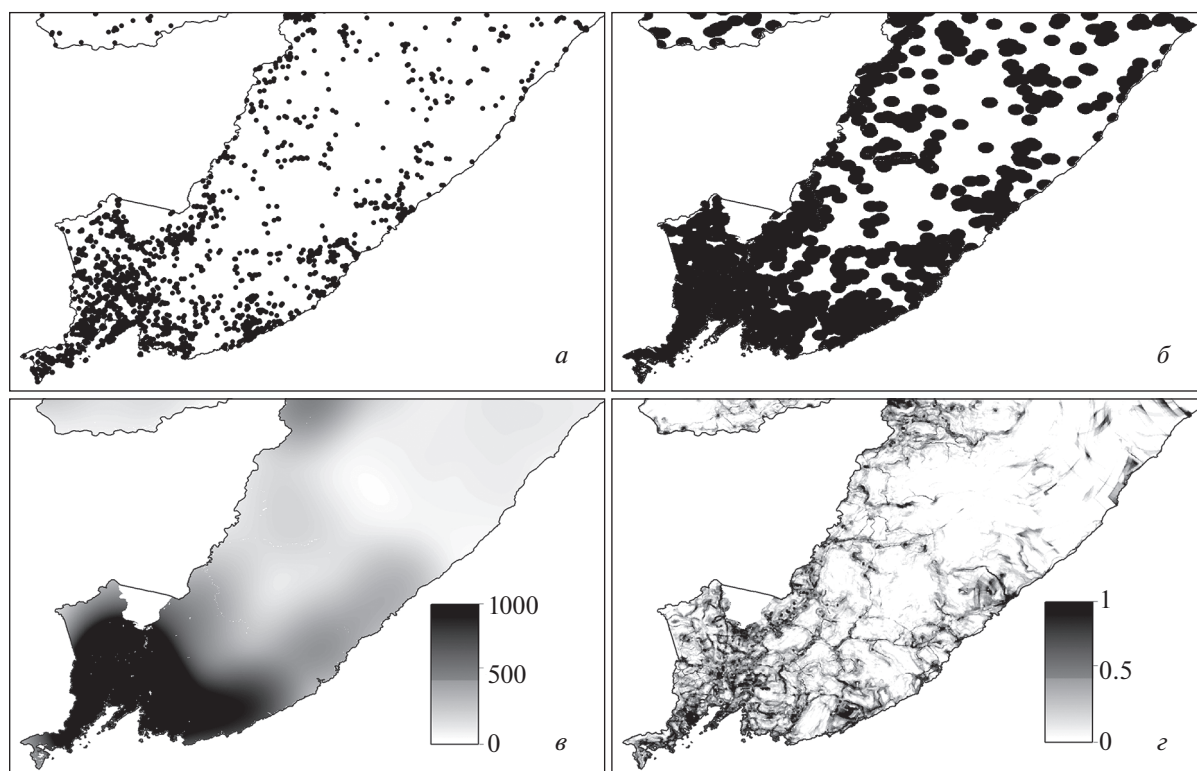


Рис. 2. Варианты создания слоя коррекции ошибки: *а* — места находок близких по выявляемости таксонов; *б* — буферные зоны, построенные вокруг точек находок; *в* — плотности коллекционных данных на основе взвешенной оценки плотности ядра Гаусса; *г* — модель “активности коллекторов” как функции генеральной совокупности точек находок (в данном случае базы данных гербарных образцов 100 видов сосудистых растений бассейна р. Амур) от удаленности от железных и автомобильных дорог, судоходных рек, населенных пунктов, заповедников.

Существуют разные подходы к созданию слоя коррекции (рис. 2): растеризация всех известных точек наблюдений всех видов некоей таксономической или иной важной с точки зрения логики исследования группы; построение буферных зон вокруг включенных в анализ ТР; вычисление относительной активности сбора коллекционных данных — функции числа наблюдений от таких переменных, как удаленность от транспортных путей, населенных пунктов, охраняемых территорий, проходимость рельефа и т.д. (Phillips et al., 2009; Kramer-Schadt et al., 2013; Syfert et al., 2013; Warren et al., 2014; El-Gabbas, Dormann, 2018). Основная идея при создании слоя коррекции сводится к описанию изученности территории: неизученные районы должны иметь меньший вес при формировании выборки ФТ по сравнению с изученными. При этом изученные места, в которых присутствовал изучаемый вид, должны иметь такой же вес, как те, где этот вид отсутствовал, но которые потенциально хорошо изучены.

Слой коррекции может учитывать также и влияние используемой картографической проекции. Так, в цилиндрической равноугольной проекции, в которой поставляются биоклиматические данные WorldClim (Hijmans et al., 2005), многократное

различие в площади ячейки раstra в зависимости от широты приводит к неравному весу точки наблюдения. Для исключения влияния ошибки проекции можно использовать готовые коэффициенты (Brown, 2014).

Фильтрация исходных данных и использование слоя коррекции сильно влияют на результирующую модель. Показано, что при использовании нескольких сценариев прореживания исходных данных и разных слоев коррекции прогнозная площадь распространения вида может различаться на 13–58% (Kramer-Schadt et al., 2013), мера AUC — на 6–25%, а коэффициент корреляции прогнозного присутствия видов с независимым набором данных может изменяться на 11–38% (Syfert et al., 2013). Некоторые авторы рекомендуют всегда использовать слой коррекции (Merow et al., 2013); мы присоединяемся к такой рекомендации.

7. ВХОДНЫЕ ДАННЫЕ: ФАКТОРЫ СРЕДЫ

Пространственные данные о параметрах окружающей среды — предикторы — часто имеют взаимную корреляцию. Это может являться причиной нестабильности модели и вносить ошибку в результаты (Dormann et al., 2013). Тем не менее

MaxEnt устойчив к влиянию взаимно коррелирующих предикторов (Elith et al., 2011) благодаря параметризации (наличие коэффициентов λ для каждого ФП; см. разделы 1 и 2). Поэтому даже если переменные сильно связаны, это не приводит к заметным искажениям предсказанного пространственного распределения. Однако есть два случая, когда взаимная корреляция ФП может привести к серьезным ошибкам: 1) если основной задачей является оценка вклада факторов среды в формирование ареала; 2) если целью работы является проекция исследуемого распространения на климаты других эпох или на иные территории, где степень или характер связи между факторами среды может быть иной.

Задача оценки вклада факторов среды в формирование ареала является традиционной для экологии и биогеографии. Тем не менее именно эта задача наталкивается на наибольшее число подводных камней в MaxEnt. Штатные возможности MaxEnt предлагают три варианта оценки связи факторов среды с результирующей моделью. Первый — “относительный вклад фактора” (percent contribution) рассчитывается на основании коэффициентов λ для каждого ФП. Здесь нужно вспомнить, что а) фактор среды может участвовать более чем в одном ФП (см. раздел 1); б) коэффициенты λ зависят от хода анализа и могут меняться от раза к разу (Phillips et al., 2019); в) коэффициенты λ распределяются между скоррелированными ФП неким образом, также зависящим от хода анализа и никак не связанными с реальным вкладом конкретных факторов среды. Таким образом, этот вариант оценки вклада фактора может использоваться лишь в частных случаях.

Второй вариант — важность при перестановках (permutation importance) — позволяет оценить вклад фактора среды, сравнивая полученную модель с аналогичной, но построенной при искаженных значениях одного из факторов (Phillips et al., 2019). Корреляция между факторами и в этом случае не позволит получить корректные значения параметра — искажая один из связанных факторов, его место при построении модели занимает второй, вне зависимости от важности.

Третий вариант — использование метода складного ножа (jackknife) — основан на сравнении моделей, построенных на каждом из факторов, с моделями, построенными без этого фактора. Очевидно, что он обладает тем же свойством, что и предыдущий вариант — если два фактора связаны между собой, то исключение одного из них приведет к меньшему искажению модели, чем если бы взаимных корреляций не было. Таким образом, используя штатные средства MaxEnt для анализа вклада факторов среды, устранение взаимных корреляций факторов обязательно.

Существуют два основных способа снижения влияния взаимных зависимостей факторов среды: выбраковка отдельных предикторов или трансформация данных для получения взаимно независимых (ортогональных) факторов. Выбор способа должен определяться задачами исследования. Выбраковка отдельных предикторов может производиться на основе статистических показателей, отражающих скоррелированность данных, к примеру, на основе коэффициента корреляции Пирсона или индекса VIF (Variance Inflation Factor) (Dormann et al., 2013). Эти показатели могут быть рассчитаны в широко используемых геоинформационных пакетах, например GRASS-GIS (Neteler et al., 2012). Из обнаруженных взаимно зависимых предикторов следует выбрать один. Это можно сделать, выбрав предиктор с наиболее интерпретируемым и известным по проведенным ранее исследованиям влиянием на объект моделирования (Brown, 2014). Возможным вариантом при исключении является сохранение тех предикторов, которые чаще используются в экологических исследованиях, например средней годовой температуры, годового количества осадков и т.п. Существуют подходы выбраковки взаимно зависимых предикторов на основании статистических тестов (Dormann et al., 2013; Cord et al., 2014; Warren et al., 2014; Barbosa, 2015; Cobos et al., 2019).

Уйти от коллинеарности исходных данных можно путем преобразования переменных во взаимно ортогональные, например методом главных компонент (Dormann et al., 2013). Такая трансформация данных более эффективна, поскольку в последующий анализ могут быть включены и главные компоненты с малой нагрузкой, которые в определенных случаях несут содержательную информацию (Кренке, Пузаченко, 2008). Кроме того, при таком подходе могут быть выявлены и удалены артефакты (например, технические погрешности) в самих предикторах. Однако этот метод лишает исследователя возможности прямой экологической интерпретации результата в понятных единицах измерения — связь модели с исходными предикторами должна быть исследована дополнительно.

Влияние еще одного свойства входящих факторов среды пока недостаточно изучено — это размер ячейки растровых географических данных, выбранный для исследования. Интуитивно понятно, что в зависимости от мозаичности ландшафтов ячейка одного и того же размера может включать гомогенные территории или “смесь” из разных типов ландшафтов. Кроме того, чрезмерное уменьшение размера ячейки требует повышения точности сбора данных (см. раздел 5), что не всегда возможно. Методы выбора оптимального размера ячейки растровых данных пока не разработаны.

ЗАКЛЮЧЕНИЕ

Как и большая часть научного инструментария, методы моделирования распространения видов не являются “волшебной палочкой”, позволяющей получать оптимальные результаты при минимальных усилиях. Напротив, от исследователя требуется внимание и специальные знания на разных этапах разработки модели. Экологическое моделирование является междисциплинарной технологией, в которую в равной степени вовлечены биологические навыки для грамотного сбора первичной информации и анализа результатов, но также и умения работы с ГИС для обеспечения корректного пространственного материала для анализа. Огромный материал, накопленный за последние полтора десятка лет, позволяет выработать общие рекомендации для моделирования распространения видов (Araújo et al., 2019).

Исследование начинается с формирования набора первичных данных о распространении вида. Эти данные (ТР) должны иметь известную точность географической привязки, соответствующую шагу ячейки раstra предикторов. В случае, если ТР не случайно распределены в пространстве, нужно устранять агрегации ТР. Необходимо изучить закономерности исследованности территории. Должен быть разработан файл коррекции, отражающий исследовательскую активность по регистрации объекта в пространстве. Набор данных о факторах среды должен быть адекватным исследуемому объекту. В случае, если планируется изучать влияние факторов среды на формирование ареала, требуется минимизировать корреляции предикторов.

Необходимо исследовать влияние типа ФТ и параметра сложности на результирующую модель и выбрать оптимальные параметры для моделирования. Использование анализа случайных подвыборок позволит оценить стабильность модели и выявить ТР, “выскакивающие” за пределы модели; им нужно уделить особое внимание. Обучение и тестирование модели должно проводиться на независимых наборах данных. Первичное тестирование модели проводится экспертным способом, но для окончательной проверки экстраполяции ареала необходимо проводить тестирующие полевые работы.

ГЛОССАРИЙ

Бинарная пороговая функция (threshold) — тип функции предикторов; содержит значение 0, если значения предиктора ниже определенного порога, и 1, если выше.

Вероятность присутствия вида — теоретическая величина, отражающая вероятность присутствия представителей вида на конкретной территории; не зависит от методов выявления.

Встречаемость вида — практически определяемая величина, показывающая, с какой вероятностью вид может быть зарегистрирован на конкретной территории; зависит от использованных методов поиска.

Выявляемость вида — коэффициент, показывающий различия между реальным присутствием вида и собранной информацией о его присутствии — встречаемостью.

Доля ошибок (omission rate) — доля неверно предсказанных присутствий.

Квадратичная функция (quadratic) — тип функции предикторов; содержит квадрат значений факторов среды.

Линейная пороговая функция (hinge) — тип функции предикторов; содержит непреобразованные факторы среды, если значения предиктора выше определенного порога, и константу, если ниже.

Линейная функция (linear) — тип функции предикторов; содержит непреобразованные данные о факторах среды.

Множественная функция (product) — тип функции предикторов; содержит произведение двух факторов среды.

Оптимальная модель — выбранная на основе определенных условий модель из набора построенных по одним и тем же данным, но с разными параметрами.

Относительное присутствие вида — величина, линейно связанная с вероятностью присутствия вида, может быть пересчитана в эту вероятность, если рассчитать линейные коэффициенты в результате дополнительной калибровки модели. Относительное присутствие вида также линейно связано с пригодностью местообитаний.

Параметр сложности (regularization multiplier) — коэффициент, регулирующий сложность модели.

Предикторы или пространственные данные о факторах среды (environmental variables) — растровые географические данные, описывающие изменчивость факторов среды в пространстве.

Распространенность (default prevalence) — доля территории, на которой априорно ожидается наличие вида.

Специфичность (specificity) — доля верно предсказанных отсутствий.

Точки регистрации вида (occurrence data) — документированные факты находок или встреч изучаемых биологических видов.

Файл коррекции (bias file) — растровый слой, отношение значений в ячейках которого показывает, насколько вероятность отбора одной ячейки будет больше, чем другой. Необходим для коррекции отклонений распределения точек регистрации вида от случайного.

Фоновые точки (background points) — случайная выборка данных о факторах среды.

Функции предикторов (features) — преобразования исходных значений предикторов (факторов среды), используемые для вычислений в MaxEnt.

Чувствительность (sensitivity) — доля верно предсказанных присутствий.

Мы благодарны двум анонимным рецензентам, детально проработавшим текст рукописи и сделавшим ценные замечания. Работа выполнена при финансовой поддержке РНФ № 18-14-00093.

СПИСОК ЛИТЕРАТУРЫ

- Дубинин М.Ю., Костикова А.А., 2008. Введение в геоинформационные системы. Векторные, растровые данные. <http://gis-lab.info/docs/giscourse/11-vector-raster.html>.
- Кренке А.Н., Пузаченко Ю.Г., 2008. Построение карты ландшафтного покрова на основе дистанционной информации // Экол. планирование и управление. Т. 2. № 7. С. 10–25.
- Лисовский А.А., Дудов С.В., Оболенская Е.В., 2020. Преимущества и ограничения использования методов экологического моделирования ареалов. 1. Общие подходы // Журн. общ. биологии. Т. 81. № 2. С. 123–134.
- Aiello-Lammens M.E., Boria R.A., Radosavljevic A., Vilela B., Anderson R.P., 2015. spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models // *Ecography*. V. 38. № 5. P. 541–545.
- Anderson R.P., Gonzalez I., 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent // *Ecol. Model.* V. 222. № 15. P. 2796–2811.
- Araújo M.B., Pearson R.G., Thuiller W., Erhard M., 2005. Validation of species-climate impact models under climate change // *Glob. Change Biol.* V. 11. № 9. P. 1504–1513.
- Araújo M.B., Anderson R.P., Barbosa A.M., Beale C.M., Dormann C.F. et al., 2019. Standards for distribution models in biodiversity assessments // *Sci. Adv.* V. 5. № 1. P. eaat4858.
- Austin M.P., 2002. Spatial prediction of species distribution: An interface between ecological theory and statistical modelling // *Ecol. Model.* V. 157. № 2. P. 101–118.
- Barbosa A.M., 2015. fuzzySim: Applying fuzzy logic to binary similarity indices in ecology // *Methods Ecol. Evol.* V. 6. № 7. P. 853–858.
- Bean W.T., Stafford R., Brashares J.S., 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models // *Ecography*. V. 35. № 3. P. 250–258.
- Boria R.A., Olson L.E., Goodman S.M., Anderson R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models // *Ecol. Model.* V. 275. P. 73–77.
- Braunisch V., Suchant R., 2010. Predicting species distributions based on incomplete survey data: The trade-off between precision and scale // *Ecography*. V. 33. № 5. P. 826–840.
- Brown J.L., 2014. SDMtoolbox: A python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses // *Methods Ecol. Evol.* V. 5. № 7. P. 694–700.
- Cobos M.E., Peterson A.T., Barve N., Osorio-Olvera L., 2019. kuenm: An R package for detailed development of ecological niche models using Maxent // *PeerJ*. V. 7. P. e6281.
- Conrad O., Bechtel B., Bock M., Dietrich H., Fischer E. et al., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4 // *Geosci. Model Dev.* V. 8. № 7. P. 1991–2007.
- Cord A.F., Klein D., Gernandt D.S., Rosa J.A.P., de la, Dech S., 2014. Remote sensing data can improve predictions of species richness by stacked species distribution models: A case study for Mexican pines // *J. Biogeogr.* V. 41. № 4. P. 736–748.
- Dormann C.F., Elith J., Bacher S., Buchmann C., Carl G. et al., 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance // *Ecography*. V. 36. № 1. P. 27–46.
- El-Gabbas A., Dormann C.F., 2018. Improved species-occurrence predictions in data-poor regions: Using large-scale data and bias correction with down-weighted Poisson regression and Maxent // *Ecography*. V. 41. № 7. P. 1161–1172.
- Elith J., Kearney M., Phillips S., 2010. The art of modelling range-shifting species // *Methods Ecol. Evol.* V. 1. № 4. P. 330–342.
- Elith J., Phillips S.J., Hastie T., Dudík M., Chee Y.E., Yates C.J., 2011. A statistical explanation of MaxEnt for ecologists // *Divers. Distrib.* V. 17. № 1. P. 43–57.
- Fei S., Yu F., 2016. Quality of presence data determines species distribution model performance: A novel index to evaluate data quality // *Landsc. Ecol.* V. 31. № 1. P. 31–42.
- Fourcade Y., Besnard A.G., Secondi J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics // *Glob. Ecol. Biogeogr.* V. 27. № 2. P. 245–256.
- Fourcade Y., Engler J.O., Rödder D., Secondi J., 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias // *PLoS One*. V. 9. № 5. P. e97122.
- Guillera-Aroita G., Lahoz-Monfort J.J., Elith J., Gordon A., Kujala H. et al., 2015. Is my species distribution model fit for purpose? Matching data and models to applications: Matching distribution models to applications // *Glob. Ecol. Biogeogr.* V. 24. № 3. P. 276–292.
- Hijmans R.J., 2012. Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model // *Ecology*. V. 93. № 3. P. 679–688.
- Hijmans R.J., Cameron S.E., Parra J.L., Jones P.G., Jarvis A., 2005. Very high resolution interpolated climate surfaces for global land areas // *Int. J. Climatol.* V. 25. № 15. P. 1965–1978.
- Kramer-Schadt S., Niedballa J., Pilgrim J.D., Schröder B., Lindenborn J. et al., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models // *Divers. Distrib.* V. 19. № 11. P. 1366–1379.
- Liu C., White M., Newell G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data // *J. Biogeogr.* V. 40. № 4. P. 778–789.
- Merow C., Smith M.J., Silander J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter // *Ecography*. V. 36. № 10. P. 1058–1069.
- Morales N.S., Fernández I.C., Baca-González V., 2017. MaxEnt's parameter configuration and small samples: Are we paying

- attention to recommendations? A systematic review // *PeerJ*. V. 5. P. e3093.
- Muscarella R., Galante P.J., Soley-Guardia M., Boria R.A., Kass J.M. et al., 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models // *Methods Ecol. Evol.* V. 5. № 11. P. 1198–1205.
- Naimi B., Araújo M.B., 2016. sdm: A reproducible and extensible R platform for species distribution modelling // *Ecography*. V. 39. № 4. P. 368–375.
- Neteler M., Bowman M.H., Landa M., Metz M., 2012. GRASS GIS: A multi-purpose open source GIS // *Environ. Model. Softw.* V. 31. P. 124–130.
- Núñez M.A., Medley K.A., 2011. Pine invasions: Climate predicts invasion success; something else predicts failure // *Divers. Distrib.* V. 17. № 4. P. 703–713.
- Pearson R.G., Raxworthy C.J., Nakamura M., Townsend Peterson A., 2007. Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar // *J. Biogeogr.* V. 34. № 1. P. 102–117.
- Phillips S.J., Dudík M., 2008. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation // *Ecography*. V. 31. № 2. P. 161–175.
- Phillips S.J., Dudík M., Schapire R.E., 2004. A maximum entropy approach to species distribution modeling // *Proceedings of the 21st int. conf. on Machine learning*. Banff, Alberta, Canada, 4–8 July 2004. P. 655–662.
- Phillips S.J., Anderson R.P., Schapire R.E., 2006. Maximum entropy modeling of species geographic distributions // *Ecol. Model.* V. 190. № 3–4. P. 231–259.
- Phillips S.J., Dudík M., Schapire R.E., 2019. Maxent software for modeling species niches and distributions (Version 3.4.1). http://biodiversityinformatics.amnh.org/open_source/maxent.
- Phillips S.J., Anderson R.P., Dudík M., Schapire R.E., Blair M.E., 2017. Opening the black box: An open-source release of Maxent // *Ecography*. V. 40. № 7. P. 887–893.
- Phillips S.J., Dudík M., Elith J., Graham C.H., Lehmann A. et al., 2009. Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data // *Ecol. Appl.* V. 19. № 1. P. 181–197.
- Radosavljevic A., Anderson R.P., 2014. Making better Maxent models of species distributions: Complexity, overfitting and evaluation // *J. Biogeogr.* V. 41. № 4. P. 629–643.
- Rangel T.F.L.V.B., Diniz-Filho J.A.F., Bini L.M., 2006. Towards an integrated computational tool for spatial analysis in macroecology and biogeography // *Glob. Ecol. Biogeogr.* V. 15. № 4. P. 321–327.
- Renner I.W., Warton D.I., 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology // *Biometrics*. V. 69. № 1. P. 274–281.
- Shcheglovitova M., Anderson R.P., 2013. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes // *Ecol. Model.* V. 269. P. 9–17.
- Syfert M.M., Smith M.J., Coomes D.A., 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models // *PLoS One*. V. 8. № 2. P. e55158.
- Thuiller W., Georges D., Engler R., Breiner F., 2019. biomod2: Ensemble platform for species distribution modelling. R package version 3.3–7.1. <https://cran.r-project.org/web/packages/biomod2>.
- Velazco S.J.E., Galvão F., Villalobos F., Marco P., de, 2017. Using worldwide edaphic data to model plant species niches: An assessment at a continental extent // *PLoS One*. V. 12. № 10. P. 1–24.
- Ward G., Hastie T., Barry S., Elith J., Leathwick J.R., 2009. Presence-only data and the EM algorithm // *Biometrics*. V. 65. № 2. P. 554–563.
- Warren D.L., Seifert S.N., 2011. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria // *Ecol. Appl.* V. 21. № 2. P. 335–342.
- Warren D.L., Glor R.E., Turelli M., 2010. ENMTools: A toolbox for comparative studies of environmental niche models // *Ecography*. V. 33. № 3. P. 607–611.
- Warren D.L., Wright A.N., Seifert S.N., Shaffer H.B., 2014. Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern // *Divers. Distrib.* V. 20. № 3. P. 334–343.

Advantages and limitations of application of the species distribution modeling methods. 2. MaxEnt

A. A. Lisovsky^{a, *}, S. V. Dudov^{b, **}

^aLomonosov Moscow State University, Biological Faculty, Zoological Museum
Bolshaya Nikitskaya, 2, Moscow, 125009 Russia

^bLomonosov Moscow State University, Biological Faculty, Department of Geobotany
Leninskiye Gory, 1, Bld. 12, Moscow, 119234 Russia

*e-mail: andlis@zmmu.msu.ru

**e-mail: serg.dudov@gmail.com

The MaxEnt software package is one of the most popular species distribution modeling tools. Despite its popularity, researchers are usually underestimating the influence of initial parameters and spatial bias in occurrence data. The choice of feature type and regularization multiplier notably affects the spatial “compactness” or “smoothness” of the model. Nonrandom distribution of occurrences in geographical space requires obligatory data correction. In the case of searching for ecological factors that influence species distribution, it is necessary to minimize multicollinearity in predictors dataset.