Species distribution modeling
МОДЕЛИРОВАНИЕ РАСПРОСТРАНЕНИЯ ВИДОВ
Түрлердің таралуын модельдеу

SPECIES DISTRIBUTION, ABUNDANCE, AND SURVIVAL MODELING: NEW OPPORTUNITIES AND METHODS

Karyakin I.V., Knizhov K.I. (Russian Raptor Research and Conservation Network; Sibecocenter LLC, Novosibirsk, Russia)

Contact:
Igor Karyakin
ikar_research@mail.ru

Kirill Knizhov
kirillknizhov@gmail.com

Recommended citation: *Karyakin I.V., Knizhov K.I.* Species Distribution, Abundance, and Survival Modeling: New Opportunities and Methods. – Raptors Conservation. 2023. S2: 347–357. **DOI:** 10.19074/1814-8654-2023-2-347-357 **URL:** http://rrrcn.ru/en/archives/35134

Many large raptor species are currently rare and most of them are endangered, and thus details of their distribution, abundance, and survival are the most important indicators for planning conservation and restoration measures and assessing the impacts of anthropogenic transformation of the environment and/or climate change on the populations of these species.

Abundance and spatial distribution of the birds under study are determined during field surveys. At the result, we obtain the distribution density in individuals, pairs, nests per unit area (for example, pairs/100 km^2), or the distance between nearest or all neighbors (represented as mathematical values (1–5, on average 3.5±1.1 km) and/or in graphical form (ranging from simple lines connecting observation points to Delaunay triangulation and a network of polygons built from observation points). Further, to generate an estimate of abundance, one must understand the area over which these data can be extrapolated. This is often challenging for many researchers – incorrect assessment of the area of the species' habitat distorts estimated abundance and neutralizes censusing efforts. How can one correctly determine the area, over which it is possible to extrapolate censusing data? The answer to this question can be found by modeling in a GIS environment using geographic layers of environmental and spatial information, or, in current terminology, species distribution modeling (SDM).

When using SDM (also known as habitat or species range modeling), environmental data (climatic and spatial variables such as temperature, humidity, wind load, topography, land cover, soils, etc. – predictor or independent variables) are calculated for geographically referenced points of a species' presence (dependent variable) and species distribution is predicted using computer algorithms and mathematical methods.

SDM is carried out in six stages: (1) idea conceptualization, (2) data preparation (presence and absence points or background points), (3) method selection (4) model fitting, (5) model evaluation and (6) habitat or area map construction.

1. Conceptualization. At this stage, we formulate the main goal of the study and decide on the modeling process design based on our knowledge of the species and the study. Data selection about the species and the environment is an important point at the initial stage. We decide whether to use only our data, or use other available data. Doing so will require some adjustments to the sample design. Next, we need to test the basic assumptions underlying the SDM, such as whether the species is in equilibrium with available environmental variables, whether the data is biased in any way (sampling bias, spatial autocorrelation, etc.), whether there are any environmental changes relative to the time of data collection, etc. Selection of adequate environmental and spatial variables, modeling algorithm, and model complexity should be based on study goals and the hypothesis regarding the relationship between the species under study and the environment in the area selected for study.

2. Data preparation. At this stage, we collect and process factual data about the species (both points of presence and points of

absence) and the environment. When preparing data, particular attention should be paid to any inconsistencies in spatial and temporal scaling of dependent and independent variables, i.e. cases where there is a large spatial or temporal difference between species and environmental data, or between environmental data (spatial and climate variables). Also, special attention should be paid to the quality of georeferencing of points of presence and the quality of species identification, which, as a rule, suffers greatly if data is collected by amateurs. In these cases, we need to make decisions about adjusting the data or discarding it. All SDM algorithms require species absence information. If such information is not available, it is replaced by background points or "pseudo-absence" data, which naturally has a negative impact on the quality of the simulation, especially on a large scale. Consideration should be given in advance to how species data will be separated for model training and model testing if the simulation uses all data collected and there are no plans for further testing of the model in the field.

3. Method selection. At this stage, we select one or several modeling methods to combine into ensemble models.

While simple factor or cluster analyses integrated into desktop GIS were used in early stages of modeling, today the selection of algorithms has expanded significantly:

Linear regression methods:

– Generalized linear model (GLM) (Nelder, Wedderburn, 1972),

– Generalized additive model (GAM) (Hastie, Tibshirani, 1990);

Machine learning methods:

– Maximum entropy method implemented in the MaxEnt program (Soberson, Peterson, 2005; Phillips *et al.*, 2006; Phillips, Dudik, 2008),

– Random Forest (RF) is an ensemble learning method for classification and regression that works by constructing multiple decision trees during training (Breiman, 2001),

– Boosted Regression Trees (BRT),

– Convolutional Neural Networks (CNN) (LeCun *et al.*, 1989),

– Genetic algorithm for Rule Set Production (GARP) (Stockwell, 1999; Stockwell, Peters, 1999),

– Machine learning supporting vector networks (Support Vector Machines, SVM) (Cortes, Vapnik, 1995; Vapnik *et al.*, 1997),

– XGBoost (eXtreme Gradient Boosting, XGB) (Chen, Guestrin, 2016).

MaxEnt and Random Forest are integrated into ArcGIS, supported in R, and available online for Google Earth Engine (GEE) users. In recent years, GEE has become increasingly popular as a resource for SDM (Crego *et al.*, 2022).

4. Fitting the model. This stage is key in SDM. Having received preliminary modeling data, we evaluate the contribution of multicollinearity and decide how to deal with it, determine how many variables can be included in the model without retraining, evaluate spatial or temporal autocorrelation and decide how to deal with it, determine the settings of the model or several models and choose which one provides the result, best or average. At the same stage, we check the plausibility of the selected relationships between species' points of presence and environmental variables by comparing coefficients and visually inspecting the plotted curves on the graphs.

5. Model evaluation. At this stage, we evaluate the forecast performance of the final model using a set of validation or test data: AUC (ROC) (Fielding, Bell, 1997; Fawcett, 2006; Hosmer, Lemeshow, 2013), TSS (Liu *et al.*, 2005; Allouche *et al.*, 2006); R2 and Kappa (Brownlee, 2016; Zhang *et al.*, 2021). Cross-validation (spatial blocks) is commonly used for this purpose (Roberts *et al.*, 2017; Valavi *et al.*, 2019; Crego *et al.*, 2022). We also select thresholds to binarize predicted probabilities based on cross-validated predictions.

Cross-validation (spatial blocks) is commonly used for this purpose (Roberts *et al.*, 2017; Valavi *et al.*, 2019; Crego *et al.*, 2022). We also select thresholds to binarize predicted probabilities based on cross-validated predictions.

6. Constructing a map of habitats or range. This is the final stage of SDM, during which we convert our predictive model into a raster and obtain a classified image with the percentage probability of the species occurring in the study area for each pixel. We calculate a probability threshold for the species' presence on pixels that we include in the final range map, and the size of the buffer built around these pixels to determine the area of habitat. The expediency of using a buffer depends on the scale of the resulting raster; the smaller the scale, the lower the relevance of the buffer. Buffer size is usually determined by the mean nearest neighbor distance (MND) and, depending

on the modeling's goals and objectives, is half, exactly, or twice the MND.

One must always critically evaluate the underlying assumptions in SDM and be aware of the potential limitations associated with a variety of factors: the ability to detect the species, uneven sampling, limitations in the selection of environmental variables, ignorance regarding certain aspects of the species' biology to identify patterns in its biotopic and territorial preferences, etc. SDM assumes that the species is in equilibrium with its environment, that we know and have carefully selected both the species' point of presence and environmental data, and that we have included all the major factors that determine the species' range limits. It should be understood that these aspects are not stable for several reasons. First, species, especially predators, respond dynamically to changes in the environment, so they will exhibit certain spatial and temporal dynamics and need to be properly taken into account in the modeling. Important factors that determine a species' response to changes in its habitat are its physiology, demography, ability to disperse, degree of tolerance to urbanization, degree of adaptation to changes in environmental factors, and interspecific interactions. All these factors engage seemingly constantly over time, including here and now, and ignoring them can significantly distort modeling results. Therefore, the ideal option for SDM is to check results in the field and adjust them.

Unfortunately, most ornithologists have difficulty using R and desktop GIS, a fact that prevents them from processing the results of their field research in accordance with modern standards. For better implementation of modeling in practice when working with rare species, we have created a software product that allows bird specialists with minimal knowledge of GIS and programming languages, but who have a certain understanding of SDM algorithms and abundance assessment, to solve problems related to modeling distribution and abundance and survival of rare species.

This software product is designed for processing various geodata containing observations of species; obtaining data from GEE rasters; classification of biotopes; population estimates, survival rates, etc.

The main interface of the product is a web interface that allows the user to select the process of interest, enter the necessary data, and receive a link to an archive containing processing results[37].

For geodata (points, polygons, etc.), it is possible to enter csv, shp, geojson files, as well as manual input using a map. To run algorithms in which it is necessary to add data from GEE rasters, a selection field is provided from the list of available earth remote sensing (ERS) products: NASADEM (NASA JPL, 2020), MOD13A1.061 Terra Vegetation Indices 16-Day Global 500m (Didan, 2021), Geomorpho90m (Amatulli *et al*., 2020), Global Habitat Heterogeneity (Tuanmu, Jetz, 2015), Global Wind Atlas (Badger *et al*., 2021), World Clim (Fick, Hijmans, 2017), ERA5-Land Monthly Aggregated – EC-MWF Climate Reanalysis (Muñoz Sabater, 2019), ESA WorldCover 10m v100 (Zanaga *et al*., 2021), Dynamic World V1 (Brown *et al*., 2022), unclassified satellite data such as surface reflectivity (SR) collection 2 Landsat 8 atmospheric-corrected (blue, red, green, near-infrared and shortwave infrared 1 bands with 30 m spatial resolution) and ALOS-2 PALSAR L-band dual-polarization (HH and HV) SAR data, and NDVI and EVI calculation data from Landsat 8 images using the GEE (normalizedDifference) function. To run algorithms using various third-party libraries, data is entered in csv files in the formats required by the corresponding libraries. At the current stage, the product includes the following modules:

1) Obtaining data from GEE rasters for given points (result presented in a table with data selected for points from rasters included in the GEE collection);

2) Obtaining a classified raster for a given area and a set of points of presence and absence of a view (training points) using the RF and MaxEnt classifiers based on GEE (both classifiers allow, for a given area of interest, a set of training points and selected remote sensing products from GEE, to obtain a classified one with using appropriate GEE raster methods of the area of interest. It is possible to cross-validate the selected models and evaluate their predictive effectiveness);

3) Three different methods to stimulate population size:

3.1) Generation of random points in a regular network – a heuristic algorithm that, based on data on the points of presence of the species and on the studied areas, generates random points, simulating

[37] *http://www.gis.altaiproject.org*

species' distribution in the general area of interest;

3.2) Distance – a method based on the Distance Sampling model (Thomas *et al*., 2010; Buckland *et al*., 2015; Miller *et al*., 2019), that accepts input of a file with the necessary variables for points and areas and displays detailed statistics as a result;

3.3) Simple site surveys using calculation of a weighted average indicator for species distribution density (Karyakin, 2004) with an calculation of asymmetric confidence interval (Ravkin, Chelintsev, 1990);

4) Estimation of nest survival based on the RMARK library (Laake, 2013). The survival calculation module includes processing of nest survival data using the nest method of the RMARK library, which can account for various variables in remote sensing data and infers the importance of variables for nest survival.

The software product is hosted on the servers of organizations recognized as undesirable in Russia, access to which is blocked by Roskomnadzor. The authors are considering options, including creating a clone on a Russian internet resource.

This work is carried out with financial support from the Critical Ecosystem Partnership Fund (CEPF)[38] within the framework of the project "Endangered Raptors Conservation on the Indo-Palaearctic Flyway").

МОДЕЛИРОВАНИЕ РАСПРОСТРАНЕНИЯ, ЧИСЛЕННОСТИ И ВЫЖИВАЕМОСТИ ВИДОВ: НОВЫЕ ВОЗМОЖНОСТИ И МЕТОДЫ

Карякин И.В., Книжов К.И. (Российская сеть изучения и охраны пернатых хищников; ООО «Сибэкоцентр», Новосибирск, Россия)

Контакт:
Игорь Карякин
ikar_research@mail.ru

Кирилл Книжов
kirillknizhov@gmail.com

Рекомендуемая цитата: *Карякин И.В., Книжов К.И.* Моделирование распространения, численности и выживаемости видов: новые возможности и методы. – Пернатые хищники и их охрана. 2023. Спецвып. 2. С. 347–357. **DOI:** 10.19074/1814-8654-2023-2-347-357 **URL:** http://rrrcn.ru/ru/archives/35134

Многие виды крупных хищных птиц в настоящее время являются редкими, большая часть находится под угрозой исчезновения, поэтому детали их распространения, численности и выживаемости являются важнейшими показателями для планирования мероприятий по охране и восстановлению, для оценки воздействия на популяции этих видов антропогенного преобразования среды и/или изменений климата.

Численность и распределение в пространстве изучаемых птиц определяются в ходе полевых учётов. На выходе мы получаем плотность распределения в особях, парах, гнёздах на единицу площади (например, пар/100 км^2) или дистанции между ближайшими или всеми соседями, которые можно представить как в виде математических значений (1–5, в среднем 3,5±1,1 км), так и в графическом виде (от простых линий, связывающих точки наблюдений, до триангуляции Делоне и сети полигонов, построенной по точкам наблюдений). Далее, для получения оценки численности, необходимо понимать площадь, на которую возможно экстраполировать эти данные. И с этим у многих исследователей возникают проблемы – неправильная оценка площади мест обитания учитываемого вида приводит к искажению оценки численности и нивелирует учётные усилия. Как правильно определить площадь, на которую возможно экстраполировать учётные данные? Ответ на этот вопрос может дать моделирование в среде ГИС с использованием географических слоёв экологической и пространственной информации, в современной терминологии – моделирование распространения видов (Species distribution modelling, SDM).

В ходе процесса SDM, также известного как моделирование среды обитания или ареала вида, для географически

[38] *http://www.cepf.net*

привязанных точек присутствия вида (зависимая переменная) определяются данные об окружающей среде – климатические и пространственные переменные, такие как температура, влажность, ветровая нагрузка, рельеф, растительный покров, почвы и т.п. (предикторы или независимые переменные), и посредством компьютерных алгоритмов и математических методов прогнозируется распределение вида в географическом пространстве и/или времени.

SDM проводится в 6 этапов: (1) концептуализация идеи, (2) подготовка данных (точек присутствия и отсутствия или фоновых точек), (3) выбор метода, (4) подгонка модели, (5) оценка модели и (6) построение карты местообитаний или ареала.

1. Концептуализация. На этом этапе мы формулируем основную цель исследования и принимаем решение о схеме процесса моделирования на основе наших знаний о виде и исследовании. Важным моментом на начальном этапе является выбор данных о виде и об окружающей среде. Мы принимаем решение об использовании только наших данных, или привлечении каких-то других доступных данных. Это потребует внесения корректив в дизайн выборки. Далее, нам надо проверить основные предположения, лежащие в основе SDM, например, находится ли вид в равновесии с доступными переменными окружающей среды, могут ли данные быть каким-либо образом смещены (неравномерность выборки, пространственная автокорреляция и т.п.), имеются ли изменения в окружающей среде относительно времени сбора данных и т.д. Выбор адекватных экологических и пространственных переменных, алгоритма моделирования и сложности модели должен основываться на цели исследования и гипотезе, касающейся взаимоотношения исследуемого вида и окружающей среды на выбранной для исследования территории.

2. Подготовка данных. На этом этапе мы собираем и обрабатываем фактические данные о виде (как точки присутствия, так и точки отсутствия) и окружающей среде. Особое внимание при подготовке данных следует уделить любым несоответствиям пространственного и временного масштабирования зависимых и независимых переменных, т.е. случаям, когда имеется большая пространственная или временная разница между данными о виде и окружающей среде, либо между данными об окружающей среде (пространственными и климатическими переменными). Также особое внимание надо уделить качеству географической привязки точек присутствия и качеству видовой идентификации, что как правило, сильно страдает, если данные собираются любителями. В этих случаях нам необходимо принять решения о корректировке данных или их отбраковке. Все алгоритмы SDM требуют информации об отсутствии вида. Если таковой информации нет, она заменяется фоновыми точками или так называемыми данными о псевдоотсутствии, что естественно сказывается отрицательно на качестве моделирования, особенно в крупных масштабах. Заранее следует подумать на то, как данные о виде будут разделены для обучения и проверки модели, если в моделировании используется весь объём собранных данных и не планируется дальнейшая проверка модели на местности.

3. Выбор метода. На этом этапе мы выбираем метод моделирования или несколько методов, для объединения в ансамблевые модели.

Если на ранних этапах моделирования использовались простой факторный или кластерный анализы, интегрированные в настольные ГИС, то в настоящее время набор алгоритмов существенно расширился:

Методы, основанные на линейной регрессии:

– Обобщённая линейная модель (GLM) (Nelder, Wedderburn, 1972),

– Обобщённая аддитивная модель (GAM) (Hastie, Tibshirani, 1990);

Методы машинного обучения:

– Метод максимальной энтропии, реализованный в программе MaxEnt (Soberson, Peterson, 2005; Phillips *et al*., 2006; Phillips, Dudik, 2008),

– Случайный лес (Random Forest, RF) – метод ансамблевого обучения для классификации и регрессии, который работает путём построения множества деревьев решений во время обучения (Breiman, 2001),

– Усиленные деревья регрессии (BRT),

– Свёрточные нейронные сети (CNN) (LeCun *et al*., 1989),

– Генетический алгоритм создания набора правил (GARP) (Stockwell, 1999; Stockwell, Peters, 1999),

– Машинное обучение, поддерживающее векторные сети (Support Vector

Machines, SVM) (Cortes, Vapnik, 1995; Vapnik *et al*., 1997),

– XGBoost (eXtreme Gradient Boosting, XGB) (Chen, Guestrin, 2016).

MaxEnt и Random Forest интегрированы в ArcGIS, имеют поддержку в среде R и доступны онлайн для пользователей Google Earth Engine (GEE). В последние годы GEE приобретает всё большую популярность в качестве ресурса для SDM (Crego *et al*., 2022).

4. Подгонка модели. Этот этап является ключевым в SDM. Получив данные предварительного моделирования, мы оцениваем вклад мультиколлинеарности и решаем, как с ней бороться, определяем сколько переменных необходимо включить в модель без её переобучения, оцениваем пространственную или временную автокорреляцию и решаем, как с ней бороться, определяем настройки модели или нескольких моделей и выбираем какой результат использовать, лучший или средний. На этом же этапе мы проверяем правдоподобие подобранных взаимосвязей между точками присутствия вида и переменными окружающей среды путём сравнения коэффициентов и визуального осмотра построенных кривых на графиках.

5. Оценка модели. На данном этапе мы оцениваем эффективность прогноза итоговой модели с помощью набора проверочных или тестовых данных: AUC (ROC) (Fielding, Bell, 1997; Fawcett, 2006; Hosmer, Lemeshow, 2013), TSS (Liu *et al*., 2005; Allouche *et al*., 2006); R2 и Kappa (Brownlee, 2016; Zhang *et al*., 2021). Обычно для этой цели используется перекрёстная проверка (пространственные блоки) (Roberts *et al*., 2017; Valavi *et al*., 2019; Crego *et al*., 2022). Также мы выбираем пороговые значения для бинаризации прогнозируемых вероятностей на основе перекрёстно проверенных прогнозов.

6. Построение карты местообитаний или ареала. Это заключительный этап SDM, в ходе которого мы конвертируем в растр нашу прогнозную модель и получаем классифицированное изображение с вероятностью распространения вида на исследуемой территории в процентах для каждого пикселя. Мы определяем порог вероятности присутствия вида для пикселей, которые включаем в итоговую карту ареала, и размер буфера, строящегося вокруг этих пикселей для определения площади местообитаний. Целесообразность использования буфера зависит от масштаба результирующего растра, чем меньше масштаб, тем ниже актуальность буфера. Размер буфера обычно определяется по средней дистанции между ближайшими соседями (MND) и, в зависимости от целей и задач моделирования, представляет собой половину, полную или удвоенную MND.

Всегда следует критически оценивать основные предположения в SDM и осознавать потенциальные ограничения, связанные с целым комплексом факторов: способность обнаруживать вид, неравномерность выборки, ограничения в выборе переменных окружающей среды, незнание определённых сторон биологии вида для выявления закономерностей в его биотопических и территориальных предпочтениях и пр. SDM предполагает, что виды находятся в равновесии с окружающей средой, что мы знаем и тщательно отобрали как точки присутствия вида, так и данные об окружающей среде, и что мы включили все основные факторы, определяющие пределы ареала вида. При этом надо понимать, что эти аспекты нестабильны по нескольким причинам. Во-первых, виды, особенно хищники, динамически реагируют на изменения среды, поэтому они будут демонстрировать определённую пространственную и временную динамику, и необходимо её правильно учесть в моделировании. Важными факторами, определяющими реакцию вида на изменения среды обитания, являются его физиология, демография, способность к расселению, степень толерантности к урбанизации, степень адаптации к изменению экологических факторов и межвидовые взаимодействия. Все эти факторы действуют на вид постоянно во времени, в том числе здесь и сейчас, и их игнорирование может существенно исказить результаты моделирования. Поэтому идеальным вариантом SDM является проверка результатов в поле и их корректировка.

К сожалению, большинство орнитологов испытывают сложности в работе с R и в настольных ГИС, что не позволяет им обрабатывать результаты своих полевых исследований в соответствии с современными требованиями. Для лучшего внедрения в практику моделирования в работе с редкими видами мы создали программный продукт, позволяющий специалистам по птицам с минимальными знаниями в ГИС и языках программирования, но имеющим определённое пред-

ставление об алгоритмах SDM и оценке численности, решать задачи, связанные с моделированием распространения, численности и выживаемости редких видов.

Программный продукт предназначен для обработки различных геоданных, содержащих наблюдения видов; получения данных с растров GEE; классификации биотопов; оценки популяции, выживаемости и т.д.

Основным интерфейсом продукта является веб-интерфейс[37], который позволяет выбирать интересующий процесс, вводить необходимые данные и получать ссылку на архив с результатами обработки.

Для геоданных (точек, полигонов и т.д.) предусмотрена возможность ввода файлов csv, shp, geojson, а также ручного ввода с помощью карты. Для запуска алгоритмов, в которых необходимо добавлять данные из растров GEE, предоставлено поле выбора из списка доступных продуктов дистанционного зондирования земли (ДДЗ): NASADEM (NASA JPL, 2020), MOD13A1.061 Terra Vegetation Indices 16-Day Global 500m (Didan, 2021), Geomorpho90m (Amatulli *et al*., 2020), Global Habitat Heterogeneity (Tuanmu, Jetz, 2015), Global Wind Atlas (Badger *et al*., 2021), World Clim (Fick, Hijmans, 2017), ERA5-Land Monthly Aggregated – ECMWF Climate Reanalysis (Muñoz Sabater, 2019), ESA WorldCover 10m v100 (Zanaga *et al*., 2021), Dynamic World V1 (Brown *et al*., 2022), неклассифицированные спутниковые данные, такие как коллекция 2 отражательной способности поверхности (SR) Landsat 8 с поправкой на атмосферу (синий, красный, зеленый, ближний инфракрасный и коротковолновый инфракрасный 1 диапазоны с пространственным разрешением 30 м) и наборы данных поляризации HH и HV ALOS с фазированной антенной решеткой L-диапазона с синтезированной апертурой (SAR), а также данные расчётов NDVI и EVI по изображениям Landsat 8 с использованием функции GEE (normalizedDifference). Для запуска алгоритмов, использующих различные сторонние библиотеки, вводятся данные в csv файлах, в форматах, требуемых соответствующими библиотеками.

На текущем этапе в продукт входят модули:

1) Получения данных из растров GEE по заданным точкам (результатом является таблица с выбранными для точек данными из растров, входящих в коллекцию GEE);

2) Получения классифицированного растра по заданной области и набору точек присутствия и отсутствия вида (тренировочных точек) с помощью классификаторов RF и MaxEnt на базе GEE (оба классификатора позволяют по заданной области интереса, набору тренировочных точек и выбранным продуктам ДДЗ из GEE получить классифицированный с помощью соответствующих методов GEE растр области интереса. Есть возможность провести кросс-валидацию выбранных моделей и оценку их прогностической эффективности);

3) Оценка численности популяции тремя различными методами:

3.1) генерация случайных точек в регулярной сети – эвристический алгоритм, который на основании данных о точках присутствия вида и об исследованных областях генерирует случайные точки, имитируя расселение вида по общей области интереса;

3.2) Distance – метод, основанный на модели Distance Sampling (Thomas *et al*., 2010; Buckland *et al*., 2015; Miller *et al*., 2019), который в качестве входных данных принимает файл с необходимыми переменными для точек и областей и в качестве результата выводит детальную статистику;

3.3) простые площадочные учёты с расчётом средневзвешенного показателя плотности распределения вида (Карякин, 2004) с расчётом несимметричного доверительного интервала (Равкин, Челинцев, 1990);

4) Оценка выживаемости гнёзд на основе библиотеки RMARK (Laake, 2013). В модуль расчёта выживаемости входит обработка данных о выживаемости гнёзд с помощью метода nest библиотеки RMARK, который может учитывать различные переменные из ДДЗ и выводит важность переменных для выживаемости гнезда.

В связи с тем, что программный продукт размещён на серверах организаций, признанных нежелательными в России, доступ к которым заблокирован Роскомнадзором, авторы рассматривают варианты создания клона на российском ресурсе.

Работа осуществляется при финансовой поддержке Фонда сотрудничества для сохранения экосистем, находящихся в критическом состоянии / The Critical

[37] *http://www.gis.altaiproject.org*

Ecosystem Partnership Fund (CEPF)[38] в рамках проекта «Сохранение угрожаемых видов пернатых хищников на Индо-Палеарктическом миграционном пути» (“Endangered Raptors Conservation on the Indo-Palearctic Migration Flyway”).

ТҮРЛЕРДІҢ ТАРАЛУЫН, САНЫН ЖӘНЕ ТІРШІЛІККЕ ҚАБІЛЕТТІЛІГІН МОДЕЛЬДЕУ: ЖАҢА МҮМКІНДІКТЕР МЕН ӘДІСТЕР

Карякин И.В., Книжов К.И. (Жыртқыш қанатты құстарды зерттеу және қорғау жөніндегі ресейлік желі; «Сибэкоцентр» ЖШҚ, Новосибирск, Ресей)

Контакт:
Игорь Карякин
ikar_research@mail.ru

Кирилл Книжов
kirillknizhov@gmail.com

Ұсынылатын дәйексөз: *Карякин И.В., Книжов К.И.* Түрлердің таралуын, санын және тіршілікке қабілеттілігін модельдеу: жаңа мүмкіндіктер мен әдістер. – Пернатые хищники и их охрана. 2023. Спецвып. 2. С. 347–357. **DOI:** 10.19074/1814-8654-2023-2-347-357 **URL:** http://rrrcn.ru/ru/archives/35134

Ірі жыртқыш құстардың көптеген түрлері қазіргі уақытта сирек кездеседі, олардың көпшілігі жойылып кету қаупінде, сондықтан олардың таралуы, саны және тіршілікке қабілеттілігі туралы мәліметтер сақтау және қалпына келтіру шараларын жоспарлаудың, осы түрлердің популяцияларына қоршаған ортаның антропогендік трансформациясы және/немесе климаттың өзгеруі әсерін бағалаудың маңызды көрсеткіштері болып табылады.

Зерттелетін құстардың саны мен кеңістікте таралуы далалық зерттеулер кезінде анықталады. Шығару кезінде біз жеке бастар, жұптар, ұялар бірлігіне (мысалы, жұп/100 км^2) ең жақын немесе барлық көршілер арасындағы қашықтықты аламыз немесе математикалық мәндер ретінде ұсынылуы мүмкін (1–5 орта есеппен 3,5±1,1 км), және графикалық түрде (бақылау нүктелерін Делон триангуляциясына қосатын қарапайым сызықтардан және бақылау нүктелерінен салынған көпбұрыштар желісінен). Әрі қарай, санын бағалау үшін бұл деректерді экстраполяциялауға болатын ауданды түсіну қажет. Осыған байланысты көптеген зерттеушілерде мәселе туындайды – есептелетін түрдің мекендеу орны ауданын дұрыс бағаламау санын бағалаудың бұрмалануына әкеледі және санақ жұмыстарын нивелирлейді. Есеп мәліметтерін экстраполяциялауға болатын ауданды қалай дұрыс анықтауға болады? Бұл сұрақтың жауабын қазіргі заманғы терминологияда – Түрлердің таралуын модельдеу (Species distribution modelling, SDM) экологиялық және кеңістіктік ақпараттың географиялық қабаттарын пайдалана отырып, ГИС-те модельдеу арқылы беруге болады.

Тіршілік ету ортасын немесе түрдің мекен ету орындарын модельдеу деп те аталатын SDM процесінде қоршаған орта деректері – температура, ылғалдылық, жел жүктемесі, жер бедері, өсімдіктер жамылғысы, топырақ және т.б. сияқты климаттық және кеңістіктік ауыспалылар – түрдің географиялық сілтеме нүктелері үшін анықталады (предикторлар немесе тәуелсіз ауыспалылар) және компьютерлік алгоритмдер мен математикалық әдістер арқылы түрдің географиялық кеңістікте және/немесе уақытта таралуы болжанады.

SDM 6 кезеңде жүзеге асырылады: (1) идеяларды түжырымдамалау, (2) деректерді дайындау (бар болу және жоқ нүктелері немесе фондық нүктелер), (3) әдісті тандау (4) модельді сәйкестендіру, (5) модельді бағалау және (6) тіршілік ету ортасының немесе таралу аймағының картасын құру.

1. Түжырымдамалау. Бұл кезеңде біз зерттеудің негізгі мақсатын түжырымдаймыз және түр және зерттеу туралы білімімізге сүйене отырып, модельдеу процесінің үлгісін туралы шешім қабылдаймыз. Бастапқы кезеңдегі маңызды сәт – түр мен қоршаған орта туралы мәліметтерді тандау. Біз тек өз деректерімізді пайдалануды немесе басқа қолжетімді деректерді пайдалану туралы

[38] *http://www.cepf.net*

шешім қабылдаймыз. Бұл үлгі дизайнына кейбір түзетулерді қажет етеді. Әрі қарай, біз SDM негізінде жатқан негізгі болжамдарды мысалы, түрдің қол жетімді қоршаған орта ауыспалыларымен тепе-тендікте болу-болмауы, деректер қандай да бір жолмен бұрмалануы мүмкін бе (іріктеменің ауытқуы, кеністіктік автокорреляция және т.б.) мәліметтерді жинау уақытына қатысты орта және т.б. тексеруіміз керек. Адекватты экологиялық және кеністіктік ауыспалыларды тандау, модельдеу алгоритмі және модель күрделілігі зерттеу мақсатына және зерттелетін түрлер мен зерттеу үшін тандалған аумақтағы қоршаған орта арасындағы қарым-қатынасқа қатысты гипотезаға негізделуі керек.

2. Мәліметтерді дайындау. Бұл кезенде біз түр (болу нүктелерін де, жоқ болу нүктелерін де) және қоршаған орта туралы нақты деректерді жинаймыз және өңдейміз. Тәуелді және тәуелсіз ауыспалылардың кеністіктік және уақытша масштабтауындағы кез келген сәйкессіздіктерге деректерді дайындау кезінде ерекше назар аудару керек, яғни, түрлер мен қоршаған орта деректері немесе қоршаған орта деректері (кеністіктік және климаттық айнымалылар) арасында үлкен кеністіктік немесе уақыттық айырмашылық бар жағдайлар. Сондай-ақ, бар болу нүктелерінің географиялық байланысу сапасына және түрлерді сәйкестендіру сапасына ерекше назар аудару керек, әдетте, деректерді әуесқойлар жинаған жағдайда үлкен зардап шегеді. Мұндай жағдайларда біз деректерді түзету немесе оларды жою туралы шешім қабылдауымыз керек. Барлық SDM алгоритмдері түрлердің жоқтығы туралы ақпаратты талап етеді. Мұндай жағдайларда біз деректерді түзету немесе оларды жою туралы шешім қабылдауымыз керек. Барлық SDM алгоритмдері түрлердің жоқтығы туралы ақпаратты талап етеді. Егер мұндай ақпарат жоқ болса, ол фондық нүктелермен немесе псевдо-болмау деп аталатын деректермен ауыстырылады, бұл эрине модельдеу сапасына, әсіресе кен ауқымда теріс әсер етеді. Егер модельдеу жиналған деректердің толық көлемін пайдаланса және үлгіні өрісте одан әрі сынау жоспарлары болмаса, модельдерді оқыту және модельді сынау үшін түр деректерінің қалай бөлінетінін алдын ала қарастыру керек.

3. Әдіс тандау. Бұл кезенде біз модельдеу әдісін немесе ансамбльдік модельдерге біріктіру үшін бірнеше әдістерді тандаймыз.

Егер модельдеудің бастапқы кезендерінде жұмыс үстеліндегі ГИС-ке біріктірілген қарапайым фактор немесе кластерлік талдаулар қолданылса, қазір алгоритмдер жиынтығы айтарлықтай кенейді:

Сызықтық регрессияға негізделген әдістер:

– Жалпылама сызықтық модель (GLM) (Nelder, Wedderburn, 1972),

– Жалпыланған аддитивті модель (ГАМ) (Хасти, Тибширани, 1990);

Машиналық оқыту әдістері:

– MaxEnt (Soberson, Peterson, 2005; Phillips *et al*., 2006; Phillips, Dudik, 2008),бағдарламасында енгізілген максималды энтропия әдісі,

– Кездейсоқ орман (Random Forest) – оқыту кезінде бірнеше шешім ағаштарын құру арқылы жұмыс істейтін жіктеу мен регрессияға арналған ансамбльдік оқыту әдісі (Breiman, 2001),

– күшейтілген регрессия ағаштары (BRT),

– өте дәл нейрондық желілер (CNN) (LeCun *et al*., 1989),

– Ережелер жинағын құрудың генетикалық алгоритмі (GARP) (Stockwell, 1999; Stockwell, Peters, 1999),

– Векторлық желілерді қолдайтын машиналық оқыту (Support Vector Machines, SVM) (Cortes, Vapnik, 1995; Vapnik *et al*., 1997),

– XGBoost (eXtreme Gradient Boosting, XGB) (Chen, Guestrin, 2016).

MaxEnt және Random Forest ArcGIS жүйесіне біріктірілген, R ортасында қолдау көрсетіледі және Google Earth Engine (GEE) пайдаланушылары үшін онлайн қолжетімді. Соңғы жылдары GEE SDM үшін ресурс ретінде кенінен танымал бола бастауда (Crego *et al*., 2022).

4. Модельді сәйкестендіру. Бұл кезен SDM үшін маңызды болып табылады. Алдын ала модельдеу деректерін ала отырып, біз мультиколлинеарлық үлесті бағалаймыз және онымен қалай күресуге болатынын шешеміз, оны қайта оқытпай модельге қанша ауыспалыны енгізу керектігін анықтаймыз, кеністіктік немесе уақытша автокорреляцияны бағалаймыз және онымен қалай күресуге болатынын шешеміз, үлгінін немесе бірнеше модельдің параметрлерін тандап, қайсысы жақсы немесе орташа нәтижені пайдаланатынын тандаймыз. Дәл осы кезенде біз түрлердің

болу нүктелері мен қоршаған ортаның ауыспалылары арасындағы тандалған қатынастардың орындылығын коэффициенттерді салыстыру және графиктердегі сызылған қисықтарды визуалды тексеру арқылы тексереміз.

5. Үлгілік бағалау. Бұл кезеңде біз тексеру немесе сынақ деректерінің жиынтығын пайдаланып соңғы үлгінің болжамдық тиімділігін бағалаймыз: AUC (ROC) (Fielding, Bell, 1997; Fawcett, 2006; Hosmer, Lemeshow, 2013), TSS (Liu *et al*., 2005; Allouche *et al*., 2006); R2 и Kappa (Brownlee, 2016; Zhang *et al*., 2021). Осы мақсат үшін әдетте қиылыстырып тексеру (кеністіктік блоктар) қолданылады (Roberts *et al*., 2017; Valavi *et al*., 2019; Crego *et al*., 2022) Біз сондай-ақ қиылыса тексерілген болжамдар негізінде болжанған бинаризация үшін шекті мәндерді тандаймыз.

6. Тіршілік ету ортасының немесе таралу аймағының картасын құру. Бұл SDM-нің соңғы кезеңі, онын барысында біз болжамдық модельді растрға түрлендіреміз және әрбір пиксель үшін зерттеу аймағында кездесетін түрлердің пайыздық ықтималдығы бар жіктелген кескінді аламыз. Біз соңғы диапазон картасына енгізетін пикселдер үшін түрлердің болу ықтималдығынын шегін және тіршілік ету ортасынын ауданын анықтау үшін осы пикселдердің айналасында салынған буфердің өлшемін анықтаймыз. Біз соңғы диапазон картасына енгізетін пикселдер үшін түрлердің болу ықтималдығынын шегін және тіршілік ету ортасынын ауданын анықтау үшін осы пикселдердің айналасында салынған буфердің өлшемін анықтаймыз. Буферді қолданудын орындылығы алынған растрдын масштабына байланысты, масштаб неғұрлым кіші болса, буфердің өзектілігі соғұрлым төмен болады. Буфер өлшемі әдетте орташа ен жақын көршілес қашықтықпен (MND) анықталады және модельдеу мақсаттары мен міндеттеріне байланысты MND-нің жартысы, толық немесе екі еселенген мөлшері болып табылады. Эрқашан SDM-дегі негізгі болжамдарды сыни түрғыдан бағалау керек және эртүрлі факторлармен байланысты ықтимал шектеулерді білу керек: түрді анықтау мүмкіндігі, біркелкі емес іріктеу, қоршаған ортанын ауыспалыларын тандаудағы шектеулер, түрдің биологиясы онын биотопиялық және аумақтық қалауларындағы зандылықтарды анықтау үшін зерттеудің кейбір аспектілерін білмеу және т.б. SDM түрдің қоршаған ортамен тепе-тендікте екенін, біз түрдің болу нүктесін де, қоршаған орта деректерін де білеміз және мүқият тандадық және түрдің таралу шегін анықтайтын барлық негізгі факторларды енгіздік деп болжайды. Бұл аспектілер бірнеше себептерге байланысты түрақты емес екенін түсіну керек. Біріншіден, түрлер, әсіресе жыртқыштар, қоршаған ортанын өзгеруіне динамикалық жауап береді, сондықтан олар белгілі бір кеністіктік және уақыттық динамикаларды көрсетеді және модельдеу кезінде дұрыс ескерілуі керек. Түрдің тіршілік ету ортасынын өзгеруіне реакциясын анықтайтын манызды факторларға онын физиологиясы, демографиясы, таралу қабілеті, урбанизацияға төзімділік дәрежесі, қоршаған орта факторларынын өзгеруіне бейімделу дәрежесі және түр аралық өзара әрекеттесу жатады. Осы факторлардын барлығы уақыт өте келе, сонын ішінде осы жерде және қазір әрекет етеді және оларды елемеу модельдеу нәтижелерін айтарлықтай бұрмалауы мүмкін. Сондықтан, SDM үшін тамаша нұсқа өрістегі нәтижелерді тексеру және оларды реттеу болып табылады.

Өкінішке орай, орнитологтардын көпшілігі R және жұмыс үстелі ГИС-ті пайдалануда қиындықтарға тап болады, бұл олардын далалық зерттеулерінін нәтижелерін заманауи талаптарға сәйкес өндеуге кедергі жасайды. Сирек кездесетін түрлермен жұмыс істеу кезінде модельдеуді тәжірибеде жақсырақ енгізу үшін біз ГИС және бағдарламалау тілдерін аз білетін, бірақ SDM алгоритмдері және санын бағалау, сирек түрлердің таралуы мен көптігі мен тіршілігін модельдеу туралы белгілі бір түсінігі бар құс мамандарына мәселелерді шешуге мүмкіндік беретін бағдарламалық өнімді жасадық.

Бағдарламалық өнім түрлерді бақылауды қамтитын эртүрлі геодеректерді өндеуге арналған; GEE растрларынан мәліметтер алуға; биотоптардын классификациялауға; популяциясын, тіршілікке қабілеттілігін т.б. бағалауға.

Өнімнін негізгі интерфейсі веб-интерфейс[37] болып табылады, ол қызықтыратын процесті тандауға, қажетті деректерді енгізуге және өндеу нәти-

[37] *http://www.gis.altaiproject.org*

желерімен мvрағатқа сілтеме алуға мvмкіндік береді.

Геодеректер vшін (нvктелер, полигондар және т.б.) csv, shp, geojson файлдарын енгізуге, сонымен қатар картаны пайдаланып қолмен енгізуге болады. GEE растрларынан деректерді қосуды қажет ететін алгоритмдерді іске қосу vшін жерді қашықтықтан зондтау (ЖҚЗ) қолжетімді өнімдерінің тізімінен таңдау өрісі беріледі: NASADEM (NASA JPL, 2020), MOD13A1.061 Terra Vegetation Indices 16-Day Global 500m (Didan, 2021), Geomorpho90m (Amatulli *et al*., 2020), Global Habitat Heterogeneity (Tuanmu, Jetz, 2015), Global Wind Atlas (Badger *et al*., 2021), World Clim (Fick, Hijmans, 2017), ERA5-Land Monthly Aggregated – ECMWF Climate Reanalysis (Muñoz Sabater, 2019), ESA WorldCover 10m v100 (Zanaga *et al*., 2021), Dynamic World V1 (Brown *et al*., 2022),

(SR) Landsat 8 жіктелмеген спутниктік деректер, мысалы, (SR) Landsat 8 атмосфералық тvзетілген беттік 2 шағылысу жинағы (көк, қызыл, жасыл, жақын инфрақызыл және қысқа толқынды инфрақызыл 1 жолағы 30 м кеністіктік рvқсат) және L-диапазонды синтетикалық апертура массиві (SAR) антенналары бар HH және HV ALOS поляризация деректер жинағы, сондай-ақ GEE (normalizedDifference) функциясын пайдаланып Landsat 8 кескіндерінен NDVI және EVI есептеулері. Эртvрлі vшінші тарап кітапханаларының көмегімен алгоритмдерді іске қосу vшін деректер csv файлдарына сәйкес кітапханалар талап ететін пішімдерде енгізіледі.

Қазіргі кезеңде өнімге келесі модульдер кіреді:

1) Берілген нvктелер vшін GEE растрларынан мәліметтер алу (нәтиже – GEE жиынына енгізілген растрлардан нvктелер vшін таңдалған деректері бар кесте);

2) GEE негізіндегі RF және MaxEnt жіктеуіштерін пайдалана отырып, берілген аумақ vшін жіктелген растрды және көріністің бар және жоқ нvктелерінің жиынын (жаттығу нvктелерін) алу (екеуі де белгілі бір қызығушылық аймағы vшін қызығушылық аймағының сәйкес GEE растрлық эдістерін қолдана отырып, классификацияланғанын алу vшін оқу нvктелерінің және GEE-ден таңдалған қашықтан зондтау өнімдерінің жиынтығы. Таңдалған vлгілерді кросс-валидациялауға және олардың болжамдық тиімділігін бағалауға болады);

3) Популяция санын vш тvрлі эдіспен бағалау:

3.1) vдайы желіде кездейсоқ нvктелерді генерациялау – тvрдің болу нvктелері және зерттелген аумақтар бойынша деректер негізінде жалпы қызығушылық аймағы бойынша тvрдің таралуын модельдейтін кездейсоқ нvктелерді тvрлендіретін эвристикалық алгоритм;

3.2) Distance – эдісі, Distance Sampling (Thomas *et al*., 2010; Buckland *et al*., 2015; Miller *et al*., 2019) vлгісіне негізделген нvктелер мен аймақтар vшін қажетті ауыспалылары бар файлды кіріс ретінде қабылдайтын және нәтижесінде егжей-тегжейлі статистиканы шығаратын модельге негізделген эдіс;

3.3) тvрдің таралу тығыздығының орташа өлшенген көрсеткішін есептеумен (Карякин, 2004) асимметриялық сенімділік интервалын есептеумен (Равкин, Челинцев, 1990) қарапайым аудандық есептеулер;

4) RMARK (Laake, 2013) кітапханасы негізінде vялардын өміршендігін бағалау. Тіршілікке қабілеттілігін есептеу модулі қашықтан зондтау деректерінен эртvрлі айнымалы мәндерді есепке алатын және vянын аман қалуы vшін айнымалы мәндердің маныздылығын шығаратын RMARK кітапханасының nest эдісін пайдаланып vянын аман қалу деректерін өндеуді қамтиды.

Бағдарламалық өнім Ресейде қолаусыз деп танылған vйымдардың серверлерінде орналастырылып, оған кіруге Роскомнадзор тыйым салғандықтан, авторлар ресейлік ресурста клон жасау нvсқаларын қарастыруда.

Бvл жvмыс The Critical Ecosystem Partnership Fund (CEPF)[38] қаржылық қолдауымен «Үнді-Палеарктикалық көші-қон vшатын жолында жойылып бара жатқан қауырсынды жыртқыштарды сақтау» (“Endangered Raptors Conservation on the Indo-Palearctic Migration Flyway”) жобасы аясында жvзеге асырылады.

[38] *http://www.cepf.net*