# OPEN DATA SOURCES FOR SPECIES DISTRIBUTION MODELLING: BIODIVERSITY INFORMATION SYSTEMS AND SPATIAL DATASETS OF ENVIRONMENTAL CONDITIONS VARIABLES

*Shashkov M.P. (Karaganda Buketov University, Karaganda, Kazakhstan)*

***Contact:***
*Maxim Shashkov*
*max.carabus@gmail.com*

***Recommended citation:*** *Shashkov M.P.* Open Data Sources for Species Distribution Modelling: Biodiversity Information Systems and Spatial Datasets of Environmental Conditions Variables. – Raptors Conservation. 2023. S2: 358–362. **DOI:** 10.19074/1814-8654-2023-2-358-362 **URL:** http://rrrcn.ru/en/archives/35137

BIOKLIM, the first algorithm for habitat modelling (Species Distribution Modelling (SDM)), was developed in the 1980s. This area of population and ecological research began to gain prominence with the wide availability of computers, development of the Internet, and development of open resources that provide access to data on species occurrences and environmental factors.

Most algorithms for SDM (with the exception of the first "bioclimatic envelope" methods) are based on regression analysis and machine learning. The most commonly used today is the MaxEnt maximum entropy method. All methods achieve the objective of revealing quantitative relationships between occurrences of the focal species and environmental variable values where the species occurs, with subsequent extrapolation of the ensuing patterns across the entire study area. The result is an assessment of habitat suitability (probability of occurrence) for the species within the study area.

Species distribution modelling methods are implemented both as standalone software products (MaxEnt) and as modules for GIS (and for QGIS, SDMToolbox for ArcGIS, etc.) and packages for the R environment (dismo, biomod2, ENMTools, etc.).

Any species distribution modelling method requires two types of input data: (1) occurrences of the focal species, represented as a set of points with geographic coordinates; and (2) environmental variables (predictors) that may be valuable for species distribution, in the format of continuous raster layers.

Considerable advances in the digitization of scientific collections around the globe and development of other sources for species distribution data have made it possible for researchers to significantly augment their own data to develop more accurate models. Such data are available through thematic repositories, the largest of which is the Global Biodiversity Information Facility – GBIF, which currently provides over 2.5 billion occurrences, two-thirds of which relate to birds. Along with data derived from scientific collections, GBIF hosts data from multiple citizen science systems as well. The largest of these is eBird, with 1,277.5 million observations. The iNaturalist system has about 20 million bird observations. A much smaller fraction of data comes from biological collections (8.5 million) and automatic observation systems (camera traps and satellite trackers, 9.5 million). GBIF has accumulated 195,000 bird observations in Kazakhstan, in addition to the above-mentioned data, originating in the following observation systems: Raptors of the World, RU-BIRDS.RU, Hatikka.fi, and Observation.org.

The volume of available data on focal species occurrence can reach tens of thousands of records, but much less is required for modelling; for this reason, data filtering and quality control are important steps. When compiling an input dataset of occurrences, the researcher must consider the biological features of focal species. In birds, the circumstances in which a particular individual was encountered is important: on the nest, while hunting during nesting, overwintering, migrating, etc., as well as age group. It is also necessary to note which part of the range is used in the model: breeding, wintering ground, or year-round presence. The records of target species' occurrences should be more or less evenly distributed over the area of interest, should not raise questions regarding identification, and should have a geographical accuracy comparable to the resolution of the predictor layers used.

The environmental variables most in demand are bioclimatic data from the WorldClim resource. Those data reflect the distribution of precipitation and long-term average temperature. Information on soil conditions is provided by SoilGrid250. Layers for land surface classification by habitat type are also available: qualitative (Global Land

Cover 2000) and quantitative (Global 1-km Consensus Land Cover). Remote sensing imagery data from the Landsat and Sentinel satellite series are often used as predictors. Both particular image channels and layers with indexes calculated on that basis (e.g., NDVI – Normalized Difference Vegetation Index) can be included in the analysis. The SRTM (Shuttle Radar Topography Mission) digital surface model is also in wide use.

It is important to test the predictors for multicollinearity, as strongly correlated factors will introduce uncertainty in the resulting model. The test is performed over the set of values that spatially correspond to the species occurrences, rather than over the entire area of the layers. Among two correlated layers, the less environmentally dependent one is usually left, for which the working hypothesis is tested or allowing comparison of the results with other studies. It is recommended that correlation coefficient values > 0.7 be taken as critical. Predictor selection should be based on focal species biology and ecology. For some species, topography may be important, not only elevation but also, for example, slope steepness. In species associated with wetlands, it is important to include layers related to the hydrological network. The influence of factors can be both direct and indirect. For example, a particular bird species nests in an area with a certain range of mean annual temperatures, but at the local level, it chooses habitats rich in food resources, which in turn may be associated with certain soil characteristics or vegetation types. Therefore, initial model builds typically use multiple layers of environmental variables to identify significant factors and the nature of their influence on the probability of encountering the target species. Usually, no more than ten predictors remain in the final model. There must be at least ten points of occurrence of the focal species for each predictor to build a good quality model.

# ОТКРЫТЫЕ ИСТОЧНИКИ ДАННЫХ ДЛЯ МОДЕЛИРОВАНИЯ АРЕАЛОВ: ИНФОРМАЦИОННЫЕ СИСТЕМЫ О БИОРАЗНООБРАЗИИ И НАБОРЫ ПРОСТРАНСТВЕННЫХ ДАННЫХ УСЛОВИЙ СРЕДЫ

*Шашков М.П. (Карагандинский университет имени академика Е.А. Букетова, Караганда, Казахстан)*

***Контакт:***
*Максим Шашков*
*max.carabus@gmail.com*

***Рекомендуемая цитата:*** *Шашков М.П.* Открытые источники данных для моделирования ареалов: информационные системы о биоразнообразии и наборы пространственных данных условий среды. – Пернатые хищники и их охрана. 2023. Спецвып. 2. С. 358–362. **DOI:** 10.19074/1814-8654-2023-2-358-362 **URL:** http://rrrcn.ru/ru/archives/35137

Первый алгоритм для моделирования ареалов (Species Distribution Modelling – SDM), BIOCLIM, появился в 1980х. Набирать популярность данное направление популяционных и экологических исследований начало с появлением доступной компьютерной техники, развитием сети Интернет, а также с разработкой открытых ресурсов, предоставляющих доступ к данным о распространении биологических видов и условиям среды.

Большинство алгоритмов построения моделей ареалов (за исключением первых методов «биоклиматической оболочки») основаны на регрессионном анализе и машинном обучении. Наиболее используемым на сегодня является метод максимальной энтропии MaxEnt. Все методы решают задачу установления количественных взаимоотношений между точками встреч целевого вида и значениями переменных среды в них с последующей экстраполяцией установленных закономерностей на всю территорию исследования. Результатом является оценка пригодности местообитаний (вероятности встречи) для целевого вида на исследуемой территории.

Методы моделирования ареалов реализованы как виде отдельных программных продуктов (MaxEnt), так и виде модулей для ГИС (smd для QGIS, SDMToolbox для ArcGIS и др.) и пакетов для среды R (dismo, biomod2, ENMTools и пр.).

Работа любого метода моделирования ареалов основана на двух типах входных данных: (1) встречи целевого вида, представленные в виде набора точек с географическими координатами, и (2) условия среды, которые могут определять распространение изучаемого вида (предикторы), в формате непрерывных растровых слоёв.

Благодаря значительным успехам в области оцифровки мировых научных коллекций и других источников данных о распространении видов у исследователей появилась возможность существенно дополнить собственные сборы для получения более точных моделей. Такие данные доступны через тематические репозитории, крупнейшим из которых является Глобальная Информационная Система о биоразнообразии GBIF, объединяющая на сегодняшний день более 2,5 млрд. находок, две трети из которых относятся к птицам. Помимо научных коллекций в GBIF широко представлены данные из систем Citizen Science. Крупнейшей из них является eBird, включающая 1277,5 млн. наблюдений. Система iNaturalist насчитывает около 20 млн. наблюдений птиц. Гораздо меньший объём данных происходит из биологических коллекций – 8,5 млн. и систем автоматического наблюдения (фотоловушек и спутниковых трекеров) – 9,5 млн. Для Казахстана в GBIF можно найти 195 тыс. находок птиц, кроме вышеупомянутых, происходящие также из систем: Raptors of the World, RU-BIRDS.RU, Hatikka.fi и Observation.org.

Объём доступных данных о встречах целевого вида может исчисляться десятками тысяч записей, но для построения модели используется гораздо меньше, поэтому важным этапом является отбор данных и контроль их качества. При формировании входного набора данных о встречах целевых видов исследователю необходимо учитывать биологические особенности объектов. Для птиц важно, при каких обстоятельствах была встречена данная особь: на гнезде, во время охоты на гнездовом участке, зимовке, пролёте и т.д, а также к какому возрастному состоянию она относится. Необходимо также учитывать, какая часть ареала будет включена в модель: гнездования, зимовки или круглогодичного присутствия. Точки встреч целевого вида должны быть более-менее равномерно распределены по территории интереса, не вызывать сомнения в корректности определения вида и иметь точность географической привязки, сопоставимую с разрешением используемых слоёв предикторов.

Наиболее востребованные переменные среды – это биоклиматические данные ресурса WorldClim, описывающие распределение осадков и средней многолетней температуры. Сведения о почвенных условиях предоставляет ресурс SoilGrid250. Также доступны слои, классифицирующие земную поверхность по типам местообитаний: качественные (Global Land Cover 2000) и количественные (Global 1-km Consensus Land Cover). Кроме того, в качестве предикторов часто используются данные спутниковой съёмки, полученные со спутников серий Landsat и Sentinel. В анализ можно включать как отдельные каналы изображений, так и слои с характеристиками, вычисленными на их основе (например, NDVI – нормализованный относительный вегетационный индекс). Также широко используется цифровая модель поверхности SRTM (Shuttle Radar Topography Mission).

Слои предикторов важно проверять на мультиколлинеарность, так как сильно взаимосвязанные факторы будут вносить неопределенность в результат моделирования. Проверка идёт не по всей площади слоёв, а только по набору значений, пространственно соответствующих находкам вида. Из двух связанных слоёв обычно оставляют менее зависимый, либо в отношении которого проверяется рабочая гипотеза, либо позволяющий сравнить результаты с данными других исследований. Рекомендуется принимать значения коэффициента корреляции > 0,7 как критическое. Выбор предикторов должен быть обусловлен особенностями биологии и экологии целевого вида. Для каких-то видов может быть важен рельеф, причём не только высоты над уровнем моря, но и, например, крутизна склонов. Для видов, связанных с водно-болотными угодьями, важно использовать гидросеть. Воздействие факторов может быть как прямым, так и опосредованным. Например, конкретный вид птиц гнездится на территории с определённым диапазоном среднегодовых температур, но

на локальном уровне выбирает место-обитания, богатые пригодными для него пищевыми ресурсами, которые в свою очередь могут быть связаны с определёнными почвенными характеристиками или типом растительности. Поэтому при тестовых построениях моделей, как правило, используют много слоев с характеристиками среды с целью выявления значимых факторов и характера их влияния на вероятность встречи целевого вида. В конечной модели остаётся обычно не более десяти предикторов. Для построения качественной модели необходимо, чтобы на каждый предиктор было не менее десяти точек встреч целевого вида.

# ТАРАЛУ АЙМАҚТАРДЫ (АРЕАЛДАРДЫ) МОДЕЛЬДЕУ ҮШІН АШЫҚ ДЕРЕКТЕР КӨЗІ: БИОАЛУАНТҮРЛІЛІК ТУРАЛЫ АҚПАРАТТЫҚ ЖҮЙЕЛЕР ЖӘНЕ ҚОРШАҒАН ОРТА ЖАҒДАЙЛАРЫНЫҢ КЕҢІСТІКТІ ДЕРЕКТЕР ЖИНАҒЫ

*Шашков М.П. (Е.А. Букетов ат. Қарағанды мемлекеттік университеті, Қарағанды, Казақстан)*

***Контакт:***
*Максим Шашков*
*max.carabus@gmail.com*

***Ұсынылатын дәйексөз:*** *Шашков М.П.* Таралу аймақтарды (ареалдарды) модельдеу үшін ашық деректер көзі: биоалуантүрлілік туралы ақпараттық жүйелер және қоршаған орта жағдайларының кеңістікті деректер жинағы. – Пернатые хищники и их охрана. 2023. Спецвып. 2. С. 358–362. **DOI:** 10.19074/1814-8654-2023-2-358-362 **URL:** http://rrrcn.ru/ru/archives/35137

Түрлердің таралуын модельдеу (Species Distribution Modelling – SDM) бірінші BIOCLIM алгоритмі 1980 жылдары пайда болды. Популяция мен экологиялық зерттеулердің бұл саласы қолжетімді компьютерлік технологиялардың пайда болуымен, интернеттің дамуымен, сондай-ақ биологиялық түрлердің таралуы мен қоршаған орта жағдайлары туралы деректерге қол жеткізуді қамтамасыз ететін ашық ресурстардың дамуымен танымал бола бастады.

Тіршілік ету ортасының модельдерін құруға арналған алгоритмдердің көпшілігі (бірінші «биоклиматтық қабық» әдістерін қоспағанда) регрессиялық талдауға және машиналық оқытуға негізделген. Бүгінгі танда ең көп қолданылатын әдіс MaxEnt максималды энтропия әдісі болып табылады. Барлық әдістер мақсатты түрлердің кездесу нүктелері мен олардағы қоршаған орта айнымалыларының мәндері арасындағы сандық байланыстарды орнату мәселесін шешеді, содан кейін белгіленген заңдылықтарды бүкіл зерттеу аймағына экстраполяциялайды. Нәтиже – зерттелетін аумақтағы мақсатты түр үшін тіршілік ету ортасының жарамдылығын (кездесу ықтималдығын) бағалау.

Тіршілік ету ортасын модельдеу әдістері жеке бағдарламалық өнімдер (MaxEnt) түрінде де, ГИС модульдері (QGIS үшін smd, ArcGIS үшін SDMToolbox және т.б.) және R ортасына арналған пакеттер (dismo, biomod2, ENMTools т.б.) түрінде де жүзеге асырылады.

Кез келген таралу аймағын модельдеу әдісінің жұмысы кіріс деректердің екі түріне негізделеді: (1) географиялық координаттары бар нүктелер жиынтығы ретінде үсынылған мақсатты түрлердің пайда болуы және (2) зерттелетін түрдің таралуын анықтай алатын қоршаған орта жағдайлары. (болжамдаушылар), үздіксіз растрлық қабаттар форматында.

Дүние жүзіндегі ғылыми жинақтарды және түрлердің таралу деректерінің басқа көздерін цифрландырудағы елеулі жетістіктермен зерттеушілер дәлірек үлгілерді шығару үшін өз коллекцияларын айтарлықтай толықтыруға мүмкіндік алды. Мұндай деректер тақырыптық репозиторийлер арқылы қол жетімді, олардың ең үлкені GBIF жаһандық биоалуантүрлілік ақпараттық жүйесі болып

табылады, ол қазіргі уақытта 2,5 миллиардтан астам жазбаны қамтиды, онын үштен екісі құстарға қатысты.

GBIF-тегі ғылыми жинақтардан басқа Citizen Science мәліметтер жүйелерінен мол деректер бар. Олардын ен үлкені – eBird, ол 1277,5 миллион бақылауды қамтиды. iNaturalist жүйесі 20 миллионға жуық құстарды бақылауды қамтиды. Деректердін анағұрлым аз көлемі биологиялық жинақтардан – 8,5 миллион және автоматты бақылау жүйелерінен (фототүзақтар мен спутниктік трекерлер) – 9,5 миллионнан келеді. Қазақстан үшін Raptors of the World, RU-BIRDS. RU, Hatikka.fi және Observation.org. жүйелерінен орын алған GBIF-те жоғарыда айтылғандардан басқа, 195 мын құс табылғанын табуға болады.

Мақсатты түрлердін кездесуі туралы қолда бар деректердін көлемі ондаған мын жазбаларды құрауы мүмкін, бірақ модельді құру үшін әлдеқайда аз пайдаланылады, сондықтан маңызды қадам деректерді таңдау және сапаны бақылау болып табылады. Мақсатты түрлердін кездесуі туралы кіріс деректер жинағын құру кезінде зерттеуші нысандардын биологиялық сипаттамаларын ескеруі керек. Құстар үшін бұл түрдін қандай жағдайда кездестіргені маңызды: ұяда, ұя салатын жерде ан аулау кезінде, қыстауда, қоныс аударуда және т.б., сондай-ақ онын қай жаста екендігі. Сондай-ақ модельге таралу аймағынын қандай бөлігі кіретінін ескеру қажет: ұя салу, қыстау немесе жыл бойы болу. Мақсатты түрлердін кездесу нүктелері қызығушылық танытатын аумақтын барлық аумағында азды-көпті біркелкі таратылуы керек, түрді сәйкестендірудін дұрыстығына күмән тудырмауы керек және пайдаланылатын болжамды қабаттардын рұқсатымен салыстырылатын геосілтеме дәлдігі болуы керек.

Ен талап етілетін ауыспалы қоршаған орта – жауын-шашыннын таралуын және орташа ұзақ мерзімді температураны сипаттайтын WorldClim ресурсынын биоклиматтық деректері. Топырақ жағдайы туралы ақпаратты SoilGrid250 ұсынады. Сонымен бірге, жер бетін тіршілік ету ортасынын түріне қарай жіктейтін қабаттар да бар: сапалық (Global Land Cover 2000) және сандық (Global 1-km Consensus Land Cover). Сонымен қатар, Landsat және Sentinel сериялы жерсеріктерінен алынған спутниктік түсірілім деректері жиі болжау ретінде пайдаланылады. Талдау жеке кескін арналарын да, олардын негізінде есептелген сипаттамалары бар қабаттарды да қамтуы мүмкін (мысалы, NDVI – өсімдік жамылғысынын нормаланған салыстырмалы индексі). SRTM (Shuttle Radar Topography Mission) сандық беттік моделі де кенінен қолданылады.

Болжамдық денгейлерді мультиколлинеарлық үшін тексеру маңызды, өйткені өзара жоғары байланысты факторлар модельдеу нәтижесіне белгісіздік экеледі. Тексеру қабаттардын бүкіл аймағында емес, тек кеністікте түрдегі табылғандарға сәйкес келетін мәндер жиынтығында жүргізіледі. Қосылған екі қабаттын эдетте жүмыс гипотезасы тексерілетін немесе нәтижелерді басқа зерттеулердін деректерімен салыстыруға мүмкіндік беретін тәуелділігі азырақ қалады. Корреляция коэффициентінін мәндерін критикалық ретінде > 0,7 қабылдау ұсынылады.

Болжамдарды таңдау мақсатты түрдін биологиясы мен экологиясынын ерекшеліктерімен анықталуы керек. Кейбір түрлер үшін рельеф теніз денгейінен биіктікте ғана емес, сонымен қатар, мысалы, беткейлердін тіктігі де маңызды болуы мүмкін. Сулы-батпақты жерлермен байланысты түрлер үшін гидроторды пайдалану маңызды.

Сулы-батпақты жерлермен байланысты түрлер үшін гидроторды пайдалану маңызды. Факторлардын әсері тікелей және жанама болуы мүмкін. Мысалы, құстардын белгілі бір түрі орташа жылдық температураның белгілі бір диапазоны бар аймақта ұя салады, бірақ жергілікті денгейде қолайлы қоректік ресурстарға бай мекендеу ортасын таңдайды, бұл өз кезегінде белгілі бір топырақ сипаттамаларымен немесе өсімдік түрімен байланысты болуы мүмкін. Сондықтан модельдерді сынау кезінде, эдетте, маңызды факторларды және олардын мақсатты түрмен кездесу ықтималдығына әсер ету сипатын анықтау үшін қоршаған ортанын сипаттамалары бар көптеген қабаттар қолданылады. Эдетте сонғы үлгіде оннан артық предиктор қалмайды. Сапалы модельді құру үшін әрбір предиктор үшін мақсатты түрдін кемінде он кездесу нүктесі болуы керек.