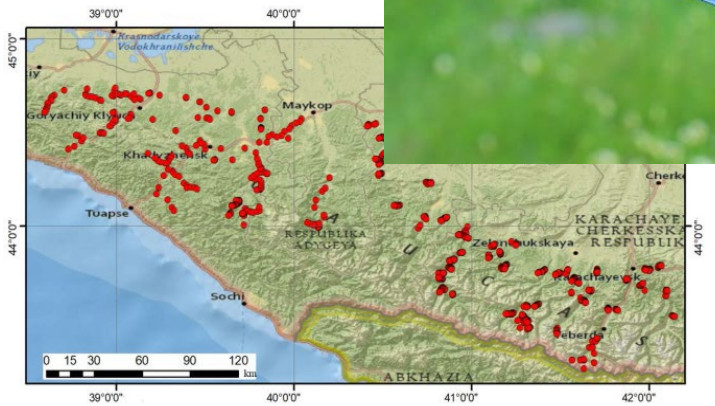
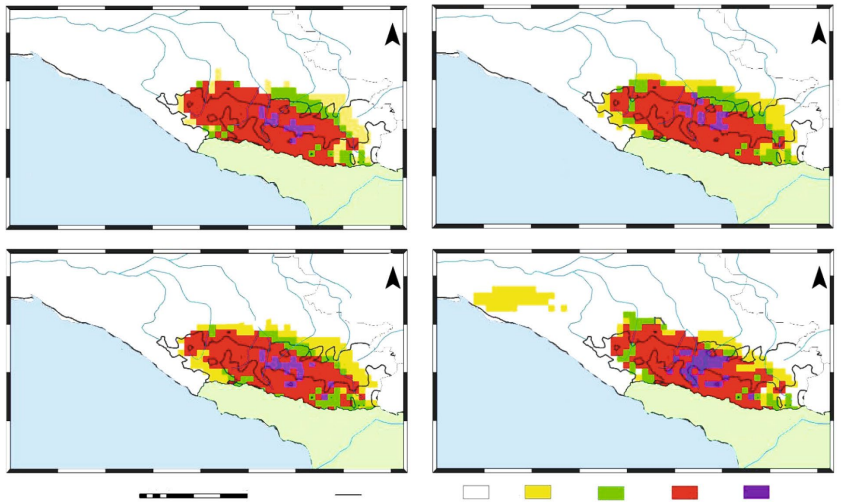
Earthworms of the North-West Caucasus: from the field to the model

Anna Geraskina
Isaev Centre for Forest Ecology and Productivity
of the Russian Academy of Sciences, Moscow





Study of earthworms in the North-West Caucasus

Kulagin, 1889;

Michaelsen, 1907,1910;

Malevich, 1947, 1959, 1966, 1967;

Perel, 1966, 1967, 1979; Vsevolodova-Perel, 1997, 2003;

Kvavadze, 1971, 1973, 1975, **1985** - monograph (Earthworms (Lumbricidae) of the Caucasus»

Ibragimov, Dzhanaev, 1972;

Prokonova, 2004, 2005 ;

Rapoport 2003-present; **2010** – dissertation «Fauna, ecology and altitudinal-belt distribution of earthworms (Oligochaeta, Lumbricidae) of the central part of the North Caucasus»

Geraskina, 2016 – present

Chorological groups of earthworms in the North-West Caucasus

- **Cosmopolitan**
- **Mediterranean**
- **Eastern-Eurasian**
- **Palaearctic**
- **Caucasian (Crimean-Caucasian endemics)**

In the North-West Caucasus, there are **24 species** of the Lumbricidae family, with **endemic and subendemic species dominating in biomass** and territory in the forest belt.

Morpho-ecological groups of earthworms in the North-West Caucasus

Epigeic

Endogeic

Epi-endogeic

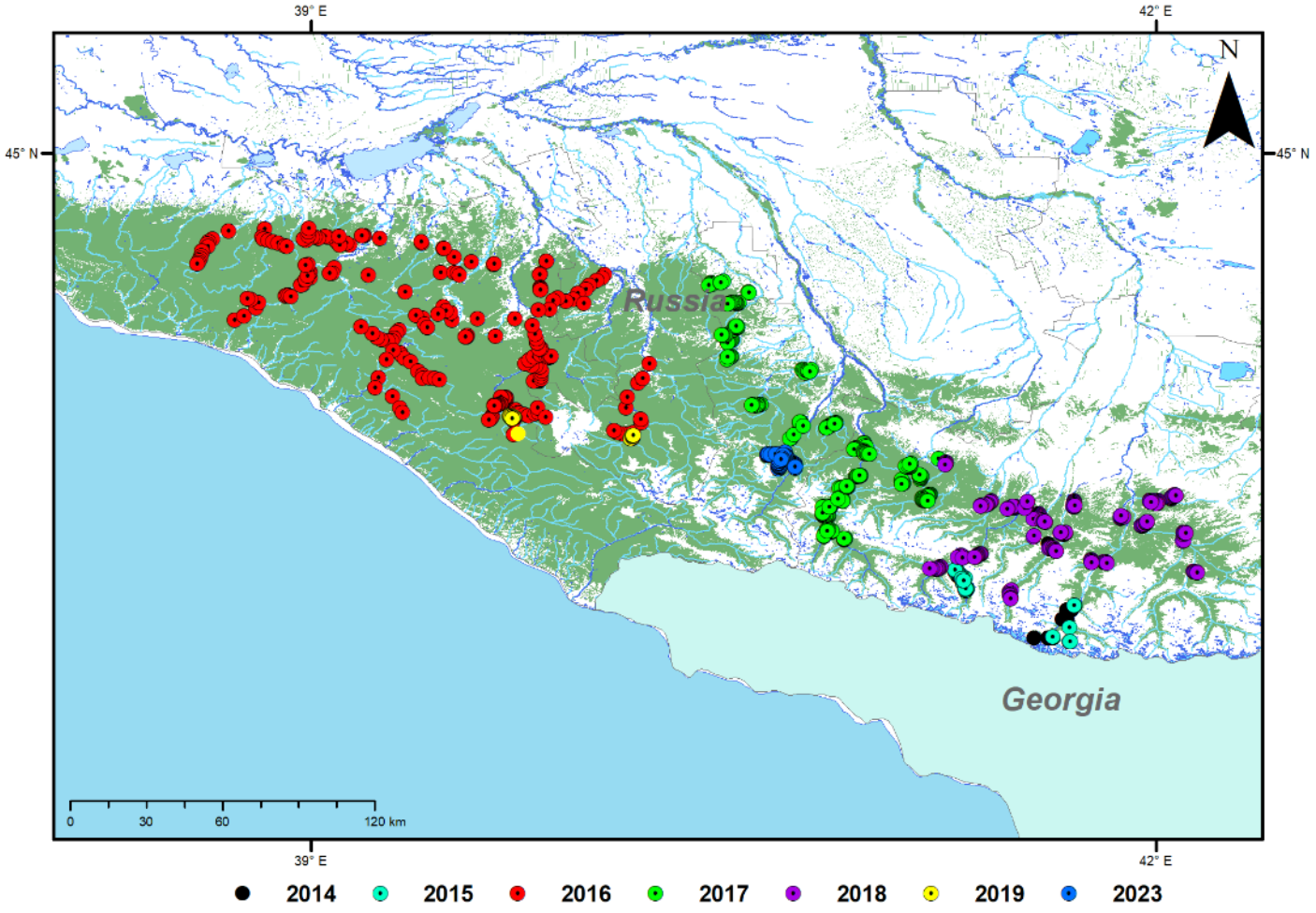
Anecic

Biology Bulletin, 2021, Vol. 48, No. 8, pp. 1176–1188

Table 1. Species composition, types of distribution ranges, and morpho-ecological groups of earthworms inhabiting the main forest types in the Bol'shaya Laba River basin (Northwestern Caucasus)

Lumbricidae species	Distribution range	Morpho-ecological group
Allolobophora chlorotica (Savigny 1826)	Palearctic	Endogeic
Dendrobaena schmidti (Michaelsen 1907)	Crimean–Caucasian	Epigeic and endogeic
Dendrobaena mariupolienis Wyssotzky 1898		Anecic
Dendrobaena attemsi Michaelsen 1902	Mediterranean	Epigeic
Dendrobaena hortensis (Michaelsen 1890)		
Dendrobaena veneta (Rosa 1886)		Epi-endogeic
Aporrectodea jassyensis (Michaelsen 1891)		Endogeic
Dendrobaena tellermanica Perel 1966	East Eurasian	
Dendrodrilus rubidus tenuis (Eisen 1874)	Cosmopolitan species	Epigeic
Eiseniella tetraedra tetraedra (Savigny 1826)		
Dendrobaena octaedra (Savigny 1826)		
Eisenia fetida (Savigny 1826)		Epi-endogeic
Lumbricus rubellus Hoffmeister 1843		
Octolasion lacteum (Örley 1885)		Endogeic
Aporrectodea rosea (Savigny 1826)		

Own research of earthworms in the North-West Caucasus



Map of points (N. Shevchenko)

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации базы данных

№ 2024623421

Видовой состав, численность и биомасса дождевых червей Северо-Западного Кавказа

Правообладатель: *Федеральное государственное бюджетное учреждение науки Центр по проблемам экологии и продуктивности лесов Российской академии наук (RU)*

Авторы: *Гераськина Анна Петровна (RU), Шевченко Николай Евгеньевич (RU)*

Заявка № 2024622949
Дата поступления 09 июля 2024 г.
Дата государственной регистрации в Реестре баз данных *06 августа 2024 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

DATABASE
«Species composition, abundance and biomass of earthworms»

The database contains information on the sampling points of soil and zoological samples (**1044** points), characteristics of vegetation, soil, species composition of earthworms, their abundance and biomass. Quantitative indicators (abundance, biomass) are given for 17 species of earthworms (one of which is represented by three forms), as well as for 4 morpho-ecological groups (epigeic, epi-endogeic, endogeic, anecic) and 5 chorological groups (cosmopolitans, Crimean-Caucasian endemics, Mediterranean, East Eurasian, Palearctic).

Research on forest types and forest microsite organization

Forest canopy

Teberdinsky Reserve,
with O.V. Smirnova, 2014
(photo N. Shevchenko)

Gaps in forest canopy

Deadwood

Caucasian Nature Reserve, 2019, 2024

From the field to the model

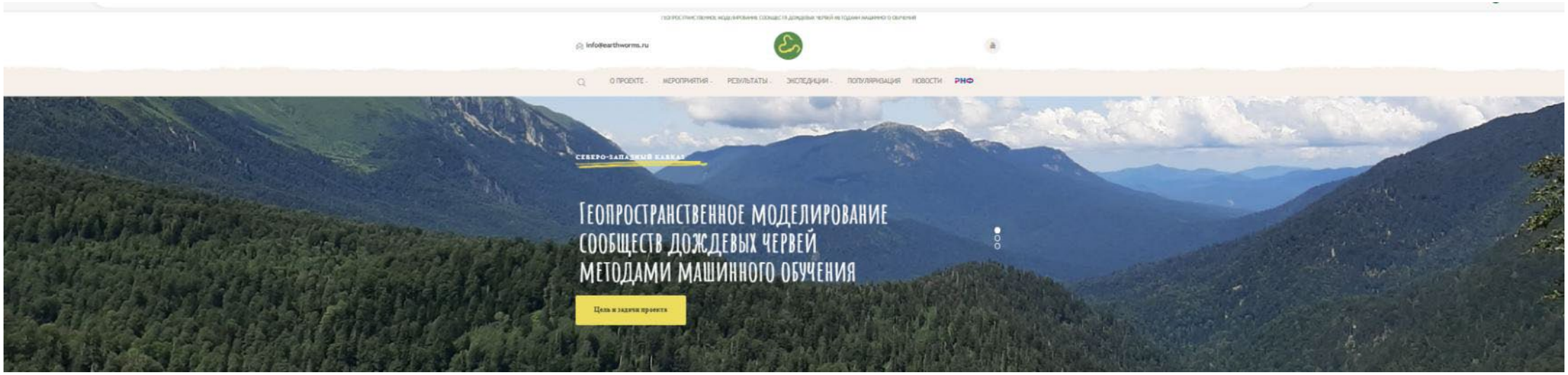
From a point to space

Machine learning modeling makes it possible to solve complex problems and make predictions based on large amounts of data. Instead of following strictly programmed instructions, algorithms learn to identify patterns, predict outcomes, and adapt to change. The goal of machine learning is to increase the accuracy of decisions and the speed at which they are made.

Моделирование методами машинного обучения даёт возможность решать сложные задачи и делать прогнозы на основе больших объёмов данных. Вместо того чтобы следовать строго запрограммированным инструкциям, алгоритмы учатся выявлять закономерности, прогнозировать результаты и адаптироваться к изменениям. Цель машинного обучения - увеличить точность решений и скорость их принятия.

GEOSPATIAL MODELING OF EARTHWORM COMMUNITIES IN THE NORTHWESTERN CAUCASUS BY MACHINE LEARNING METHODS (2024-2025) № 23-24-00543.

Geraskina A.P., Plotnikova A.S., Narykova A.N., Shevchenko N.E., Arkhiptseva E.A. (CEPF RAS)



Сайт проекта:
https://earthworms.ru/

MACHINE LEARNING METHODS

~~Naive Bayes~~

Random Forest

GradientTreeBoost

Maxent

Naive Bayes

- Naive Bayes is a machine learning classification algorithm based on **Bayes' theorem**. It assumes that all **variables** in a dataset are "naive", i.e. **uncorrelated** with each other.
- Naive Bayes works better with categorical variables, since it considers each feature value to be a separate category. There are variants of the Naive Bayes classifier that can handle continuous variables, provided that the data has a normal distribution.
- In the problem of building geospatial models of earthworm communities, the training data is not normally distributed and the predictors are continuous.
- Naive Bayes is used exclusively for solving classification problems and is not applicable for building regression models.

- Naive Bayes - это алгоритм классификации в машинном обучении, основанный на **теореме Байеса**. Он предполагает, что все переменные в наборе данных «наивные», **то есть не коррелируют** друг с другом. Теорема Байеса - одна из основных теорем элементарной теории вероятностей, которая позволяет определить вероятность события при условии, что произошло другое статистически взаимозависимое с ним событие.
- Naive Bayes лучше работает с **категориальными** переменными, поскольку считает каждое значение признака отдельной категорией. Существуют варианты наивного байесовского классификатора, которые могут обрабатывать **непрерывные переменные с условием того, что данные имеют нормальное распределение**.
- В задаче построения геопространственных моделей сообществ дождевых червей данные обучающей выборки **не подчиняются закону и предикторы являются непрерывными.** Ввиду перечисленных ограничений использование метода Naive Bayes для создания регрессионных моделей сообществ дождевых червей невозможно.
- Метод Naive Bayes используется исключительно для решения задач классификации и не применим для построения регрессионных моделей.

Random Forest

- ✓ **Random Forest** is a machine learning algorithm based on an ensemble of decision trees. Each tree is built independently from each other on different subsamples of training data, using different combinations of features (characteristics) of objects.

Advantages:

- ✓ resistance to overfitting due to the use of subsamples and random features;
- ✓ ability to process a large number of features and work with data of different types (e.g. numeric, categorical, text);
- ✓ ability to assess the importance of each feature for the task.

- ✓ **Random Forest** - алгоритм машинного обучения, который основан на ансамбле деревьев решений. Каждое дерево строится независимо друг от друга на разных подвыборках обучающих данных, при этом используются разные комбинации признаков (характеристик) объектов.

Преимущества:

- ✓ устойчивость к переобучению (overfitting) благодаря использованию подвыборок и случайных признаков;
- ✓ способность обрабатывать большое количество признаков и работать с данными различных типов (например, числовыми, категориальными, текстовыми);
- ✓ возможность оценить важность каждого признака для задачи.

GradientTreeBoost

- A machine learning method that uses an ensemble of decision trees.
- The method differs from Random Forest in that instead of averaging the decisions of individual elements of the ensemble, additional elements are successively added to obtain the most accurate estimate of the variable being modeled.
- Compared to RF, gradient boosting is more susceptible to overfitting. The modeling performance when using gradient boosting can depend on the hyperparameter settings.

- Метод машинного обучения, использующий ансамбль деревьев решений.
- Метод отличается от Random Forest тем, что вместо усреднения решений отдельных элементов ансамбля, происходит последовательное добавление дополнительных элементов, чтобы получить наиболее точную оценку моделируемой переменной.
- По сравнению с RF градиентный бустинг более подвержен переобучению. Эффективность моделирования при использовании градиентного бустинга может зависеть от настройки гиперпараметров.

Modeling quality assessment metrics

- coefficient of determination (R^2),
- Mean Absolute Error, MAE, (средняя абсолютная ошибка)
- Root Mean Squared Error, RMSE (корень из средней квадратичной ошибки)

The coefficient of determination R^2 shows how much the dependent variable is explained or predicted by the independent variables (spatial predictors) in the model.

MAE is a metric that characterizes the average magnitude of deviations of the predicted values of the model from the actual values. MAE measures the accuracy of the model by summing and averaging the absolute differences between the predicted and actual values.

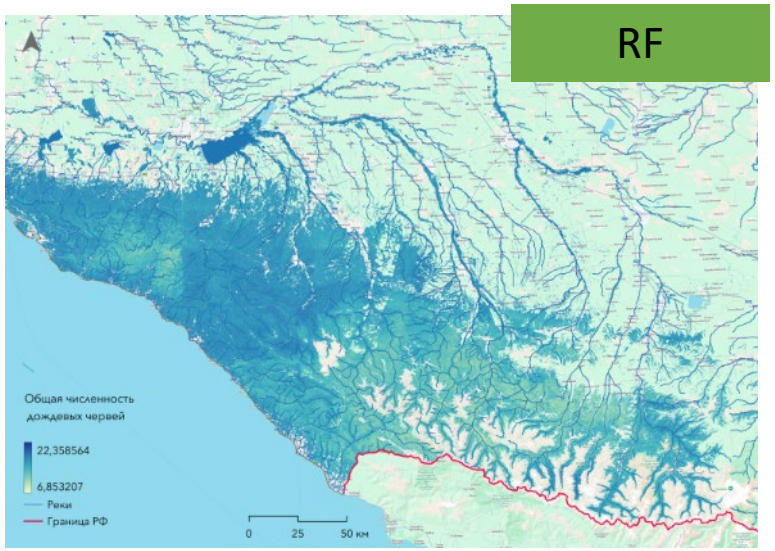
The RMSE metric is used to assess the overall accuracy of the forecast, where a lower value means a higher accuracy of the model's forecast. RMSE has the same dimension as the original data.

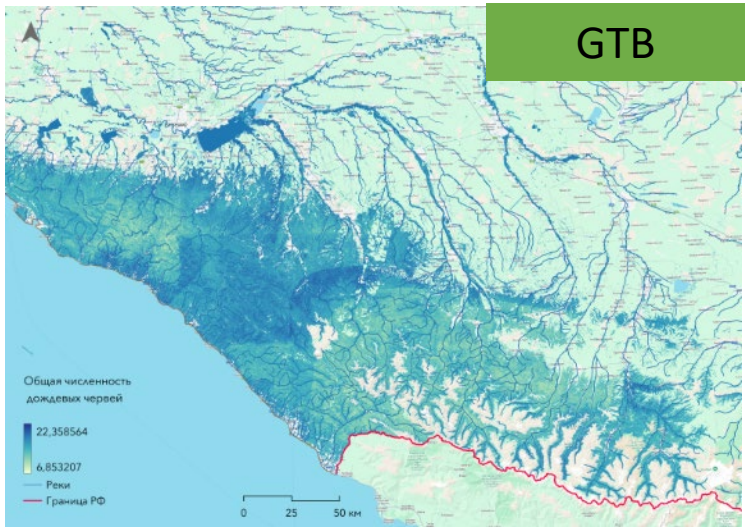
Коэффициент детерминации R^2 показывает насколько зависимая переменная объясняется или предсказывается независимыми переменными (пространственными предикторами) в модели.

MAE - это метрика, которая характеризует среднюю величину отклонений предсказанных значений модели от фактических значений. MAE измеряет точность модели, суммируя и усредняя абсолютные разности между предсказанными и реальными значениями.

Метрика RMSE используется для оценки общей точности прогноза, где меньшее значение означает более высокую точность прогноза модели. RMSE имеет одинаковую размерность с исходными данными.

Random Forest, GradientTreeBoost





In all earthworm community models, the **Random forest demonstrated higher modeling accuracy** compared to Gradient boosting, which is confirmed by the accuracy metrics.

The average R^2 value for all RF models ranges from 0.271 to 0.447, while for the BRT models it ranges from 0.172 to 0.44.

The average RMSE error in modeling the earthworm population by the RF method is in the range from 4.3 to 19.7 (ind/m^2), while for the BRT method it is from 3.6 to 23.7 (ind/m^2).

The average RMSE error in modeling the earthworm biomass by the RF method is in the range from 3.5 to 7.9 (g/m^2), while for the BRT method it is from 3.8 to 8.9 (g/m^2).More detailed hyperparameter tuning is required to improve the modeling accuracy of the GTB method.

Maxent

Maxent is a machine learning algorithm that is applicable to predicting the probability region of presence of a species. The method is based only on the points of presence of species, without taking into account the places of documented absence. The essence of the method comes down to finding patterns in the distribution of values of ecological parameters at the points where the habitat of the species is proven. Maxent chooses the probability region of presence that is most similar to the environment from which it was derived, while reducing all other assumptions (or maximizing its entropy): "*it is consistent with everything that is known, but carefully avoids assuming anything unknown*" (Jaynes, 1990).

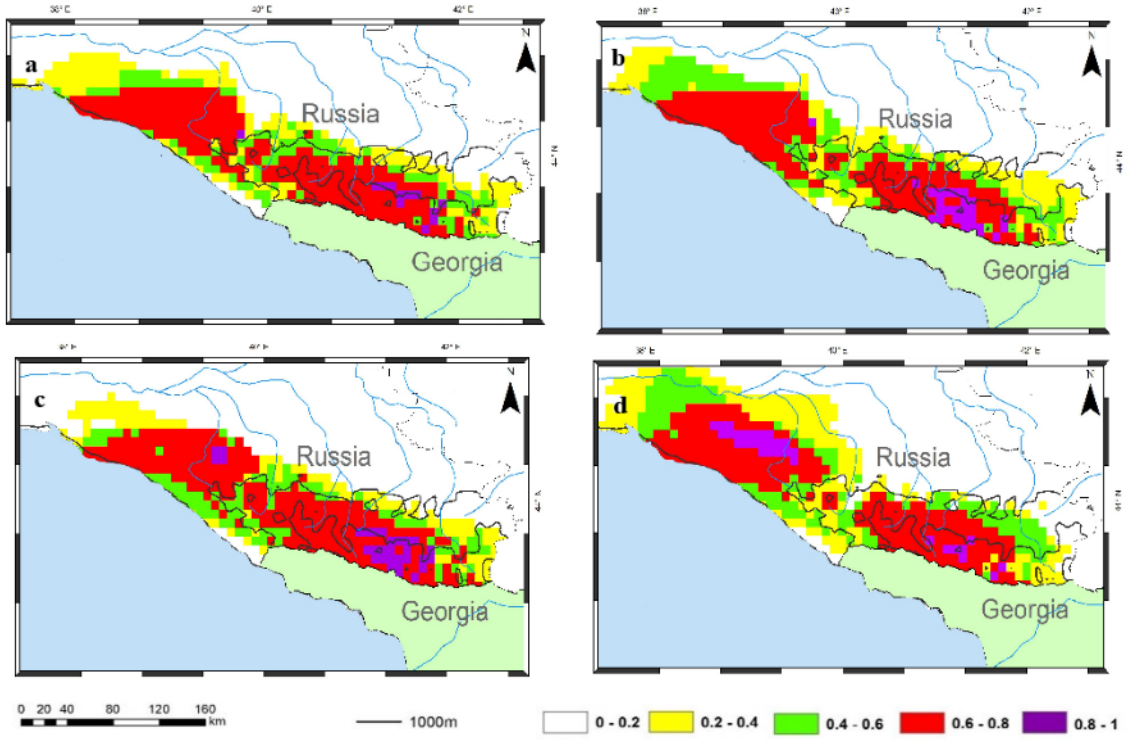
Maxent это алгоритм машинного обучения, который применим для прогнозирования области вероятности присутствия вида. Метод основывается только на точках присутствия видов (presence-only), без учета мест документированного отсутствия. Суть метода сводится к поиску закономерностей распределения значений экологических параметров в точках, где доказано обитание вида. Maxent выбирает область вероятности присутствия, которая больше всего похожа на среду, из которой она была выведена, уменьшая при этом все другие предположения (или максимизируя ее энтропию): «*она согласуется со всем, что известно, но тщательно избегает допущения чего-либо неизвестного*» (Джейнс, 1990).

Basic conditions for successful modeling in Maxent

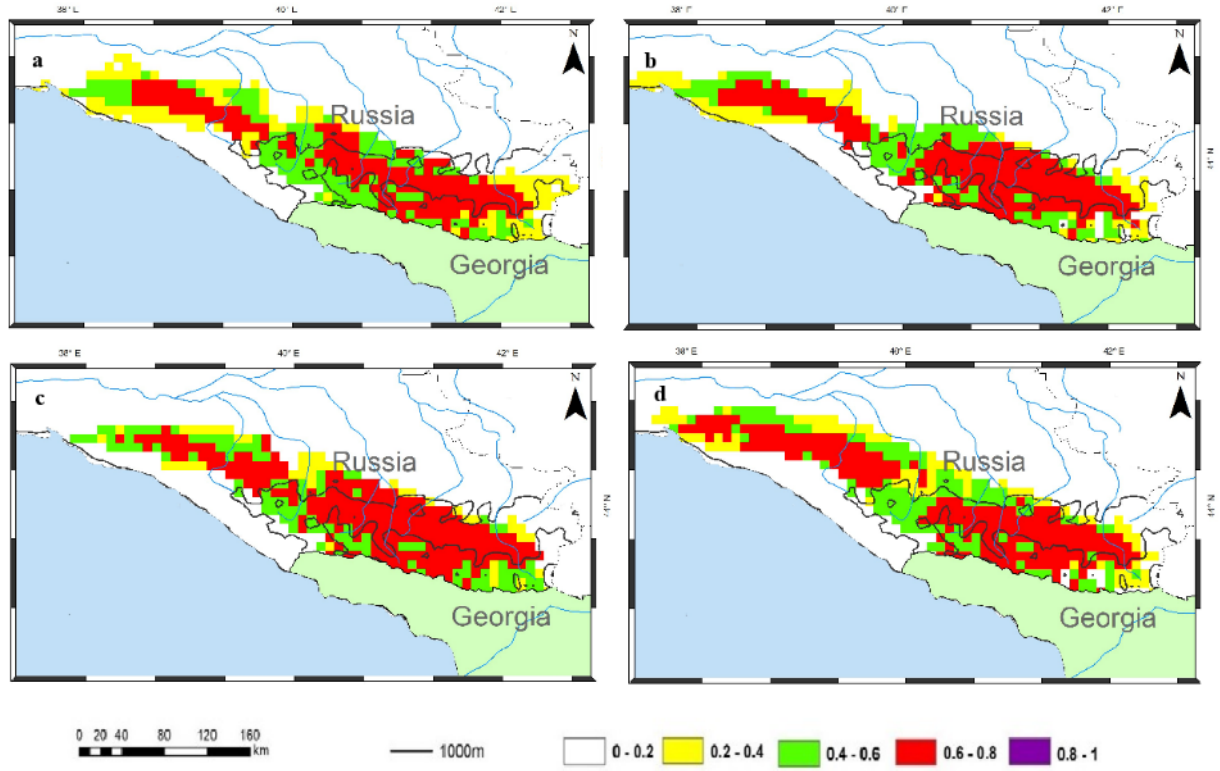
- Reliability of primary information in the field (correct species identification, clear geographic reference).
- Proper formation of a set of primary data. If the points of presence are distributed in an aggregated manner, this can have a significant impact on the forecasting results.
- Most authors suggest eliminating aggregations of species findings (an exception may be for rare species with narrow ranges).
- Knowledge and skills in working with GIS to form correct spatial material for analysis.
- Model training and testing should be carried out on independent data sets.
- To analyze the influence of environmental factors on the formation of a range, it is necessary to minimize predictor correlations.

- Достоверность первичной информации, полученной в полевых условиях (верная идентификация видов, четкая географическая привязка).
- Грамотное формирование набора первичных данных. Если точки присутствия распределены агрегировано, это может оказать существенное влияние на результаты прогнозирования. Большинство авторов предлагает устранять агрегации находок видов (исключение может быть для редких видов с узкими ареалами).
- Знания и умения работы с ГИС для формирования корректного пространственного материала для анализа.
- Обучение и тестирование модели должно проводиться на независимых наборах данных.
- Для анализа влияния факторов среды на формирование ареала, необходимо минимизировать корреляции предикторов.

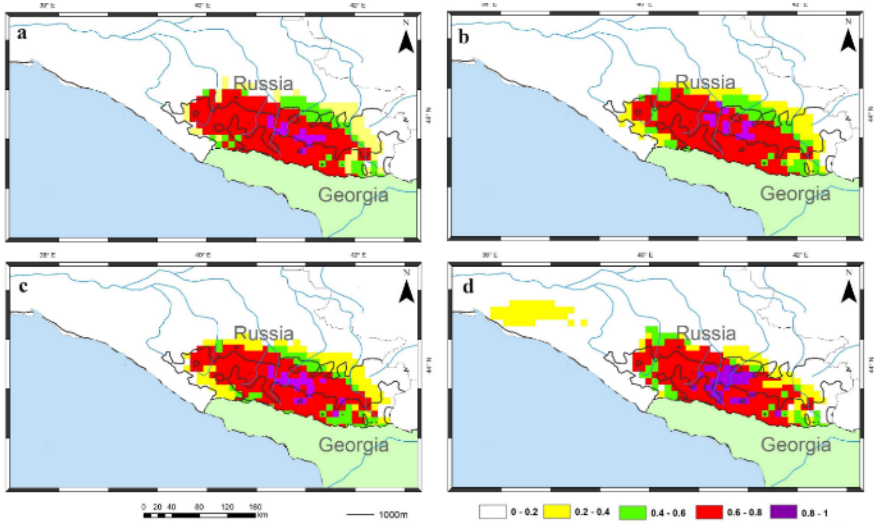
Range models for earthworm species were constructed using the maximum entropy method (Maxent) for 4 probable climate scenarios (rcp): 2.6, 4.5, 6.0 and 8.5 up to 2070.



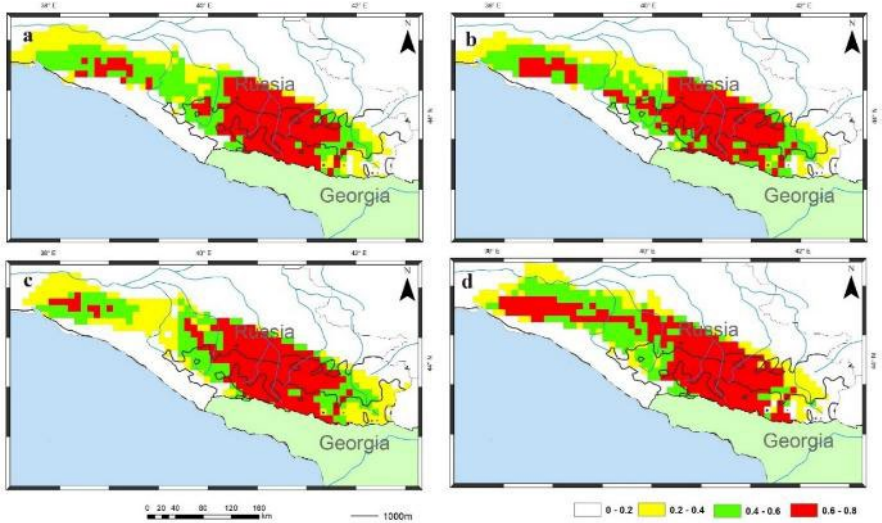
Dendrobaena attemsi (epigenic, Mediterranean)



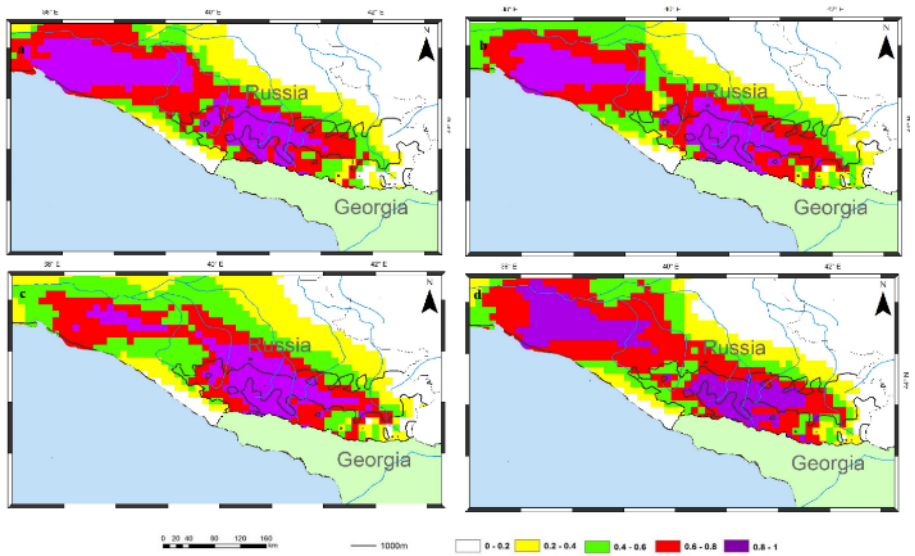
Dendrobaena octaedra (epigenic, cosmopolitan)



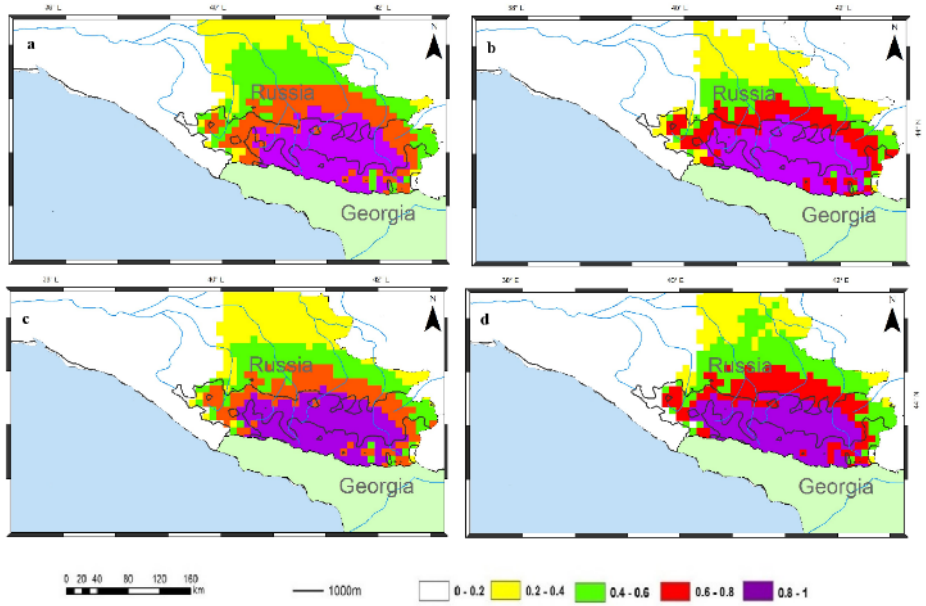
Dendrobaena schmidti (epi-endogenic, (sub)endemic)



Dendrobaena schmidti (endogenic, (sub)endemic)



Eisenia fetida (epi-endogeic, cosmopolitan)



Dendrobaena veneta (epi-endogeic, средиземноморский)

Maxent

The **most significant bioclimatic predictors** for most species are:

- ✓ **Precipitation of the driest month – bio 14 (correlates with bio 17 - amount of precipitation in the driest quarter)**
- ✓ **Seasonality of precipitation (variation coefficient) – bio 15**
- ✓ **Maximum temperature of the warmest month – bio 5**

Earthworms are sensitive to an increase in temperature combined with a decrease in precipitation, as well as to uneven precipitation, which is reflected in a reduction in the area of potential habitats in the most suitable habitats for these species of earthworms under different climate scenarios, especially the most "severe" ones – RCP 6.0 and RCP 8.5. **The most sensitive to climate change are endemic/subendemic Crimean-Caucasian species**, as well as **Mediterranean species**, for which the area of distribution in mid- and high-mountain forests is reduced. The least sensitive are cosmopolitans.

CONCLUSION

"Species distribution modeling methods are not a 'magic wand' that allows one to obtain optimal results with minimal effort. On the contrary, the researcher is required to pay attention and have special knowledge at different stages of model development."

(Lisovsky, Dudov, 2020).

«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».

(Лисовский, Дудов, 2020).

The creative team of the project

к.б.н., зав. лабораторией ЦЭПЛ РАН

АННА ГЕРАСЬКИНА

Полевые исследования. Количественный учет дождевых червей. Идентификация видов дождевых червей. Экспертная оценка результатов моделирования

к.т.н., с.н.с. ЦЭПЛ РАН

АЛЕКСАНДРА ПЛОТНИКОВА

Подготовка предикторов геопространственного моделирования. Техническое сопровождение моделирования. Разработка сайта проекта

аспирант ЦЭПЛ РАН

АННА НАРЫКОВА

Геопространственное моделирование. Выявление предикторов, определяющих состав сообществ дождевых червей. Создание WEB-ГИС проекта

к.б.н., ученый секретарь ЦЭПЛ РАН

НИКОЛАЙ ШЕВЧЕНКО

Полевые исследования. Геоботанические описания. Максент-моделирование. Экспертная оценка результатов моделирования

инженер ЦЭПЛ РАН

ЕЛЕНА АРХИПЦЕВА

Создание WEB-ГИС и сайта проекта

https://earthworms.ru/