THE SCIENCE OF WHERE

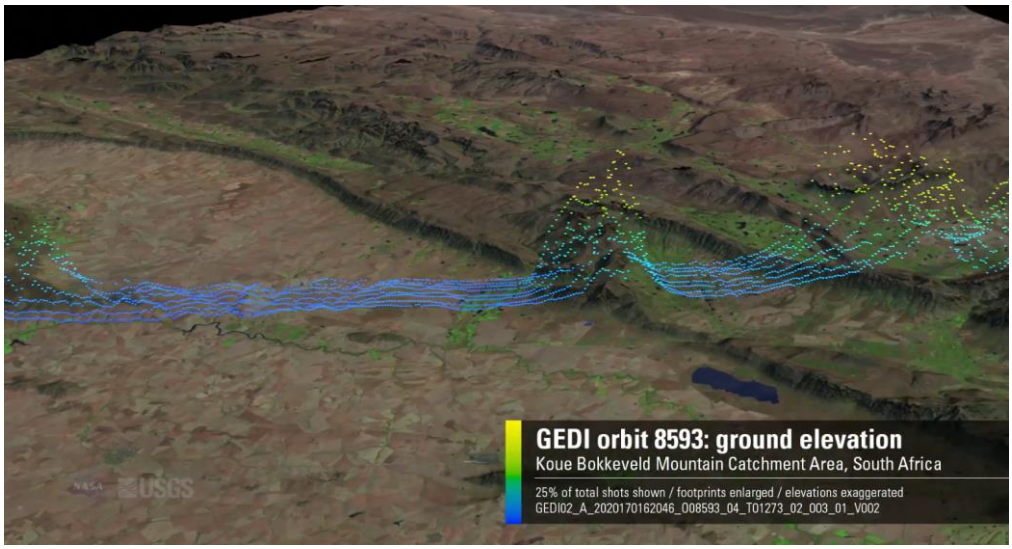
Учебное руководство к мастер-классу

Расчет биомассы растительности с помощью данных сенсора GEDI и регрессионого анализа

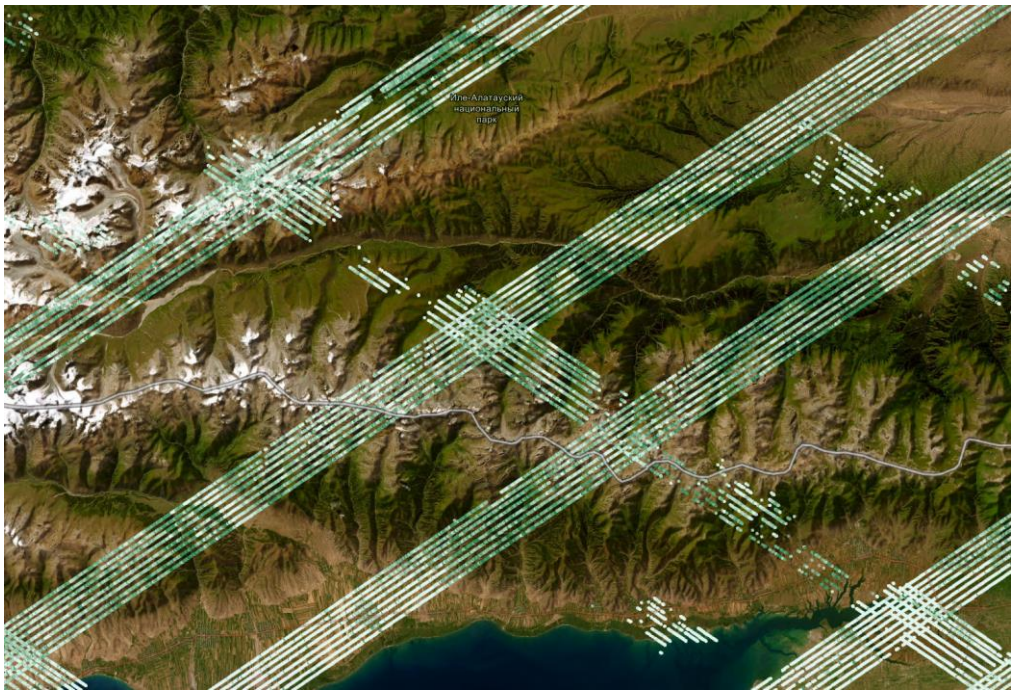
data+ esri Official Distributor

ДАННЫЕ ИНСТРУМЕНТА GEDI

GEDI - Global Ecosystem Dynamics Investigation (инструмент глобального исследования экосистем). Это лазерный сканер типа LIDAR, который размещен на борту МКС и используется для сбора данных о высоте и плотности растительности. Сканирование производится с помощью коротких импульсов, формирующих полосу шириной 25 метров вдоль наземной проекции орбиты МКС. Данные лазерного сканирования содержат информацию о высоте поверхности земли, высоте растительности, плотности растительности и других параметров.



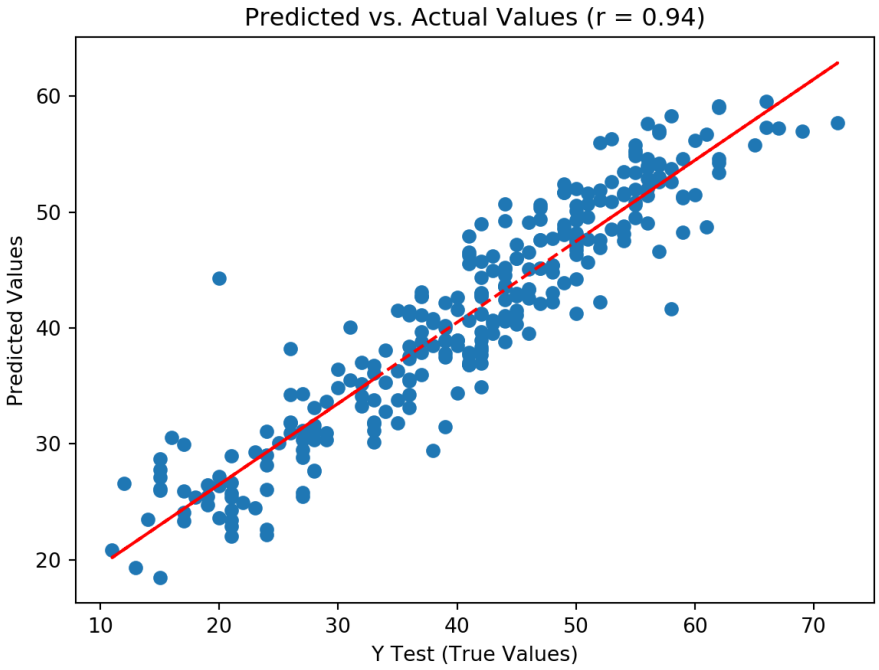
Но покрытие не сплошное, а в виде достаточно узких полос. Расстояние между импульсами – 60 метров, расстояние между полосами от 600 метров и более. Исходные данные выглядят следующим образом:



В этом упражнении мы будем решать задачу предсказания значений биомассы для участков, которые не попали в полосы измерений сенсора GEDI. Для этого мы воспользуемся алгоритмами машинного обучения.

РЕГРЕССИЯ КАК ОДИН ИЗ ИНСТУМЕНТОВ МАШИННОГО ОБУЧЕНИЯ

Задача всех алгоритмов машинного обучения – подбор оптимальных значений параметров функции, которая будет предсказывать нужные нам значения (это могут быть числа, картинки, текст и т.д.) на основе входных данных. Для алгорима машинного обучения необходима **обучающая выборка**. Обучающая выборка это набор входных данных и результатов, которые мы хотим получить на выходе. Один из самых простых случаев регрессии – линейная регрессия.



Предположим, у нас есть обучающая выборка с парами значний X и Y. Можно построить линейную регрессию, т.е. найти параметры k и b для линейной функции y = kx + b. Параметры подбираются таким образом, чтобы ошибка, т.е. сумма отклонения прямой линии от измерений была минимальной. После подбора параметров k и b, можно предсказывать значения Y по значению X. Для расчета регрессии могут использоваться гораздо сложные функции, содержащие не два, а гораздо большее количество параметров. Машинное обучение это набор алгоритмов и техник, которые подбирают параметры этих функций автоматические на основе обучающей выборки.

ПОСТАНОВКА ЗАДАЧИ

Есть нерегулярные измерения сенсора GEDI на территорию города Алматы и окрестностей. Есть сцена Landsat-9 с разрешением 30 метров и ЦМР. В отличие от измерений GEDI эти данные покрывают всю интересующую нас территорию. Мы будем предсказывать значения биомассы растительности, используя данные GEDI как обучающую выборку.

Регрессионная функция будет устанавливать связь между значениями биомассы растительности и следующими переменными:

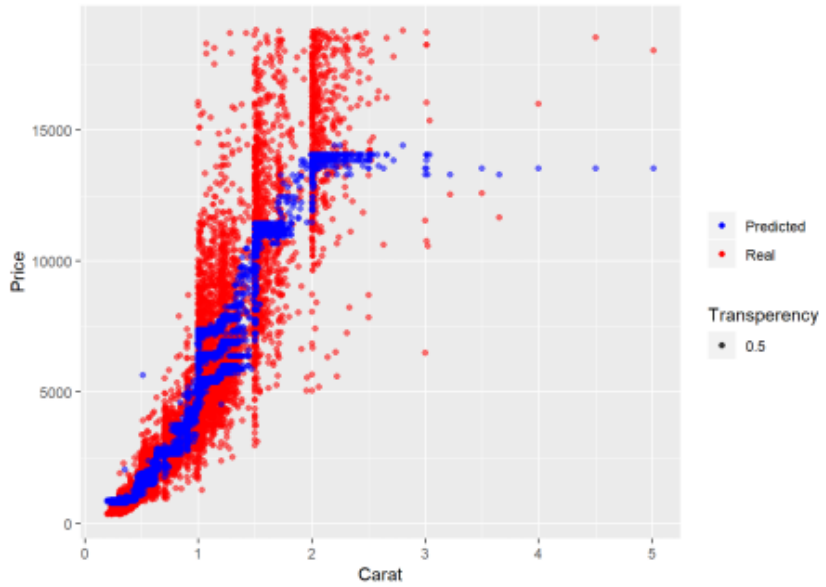
- Все каналы сцены Landsat 9
- Индекс NDVI (на базе сцены Landsat 9)
- Индекс SAVI (на базе сцены Landsat 9)
- Высота поверхности (ЦМР)
- Экспозиция склона (на базе ЦМР)
- Уклон (на базе ЦМР)

Фактически алгоритм регрессии подбирает параметры для уравнения:

Биомасса (данные GEDI) = Functon(Landsat Bands, NDVI, SAVI, DEM, Aspect, Slope)

Мы будем использовать регрессию на базе алгоритма **Random Forest**. В отличие от линейной регрессии, этот алгоритм хорошо работает с нелинейными зависимостями, достаточно устойчив к ошибкам в обучающей выборке и не требует больших вычислительных мощностей.

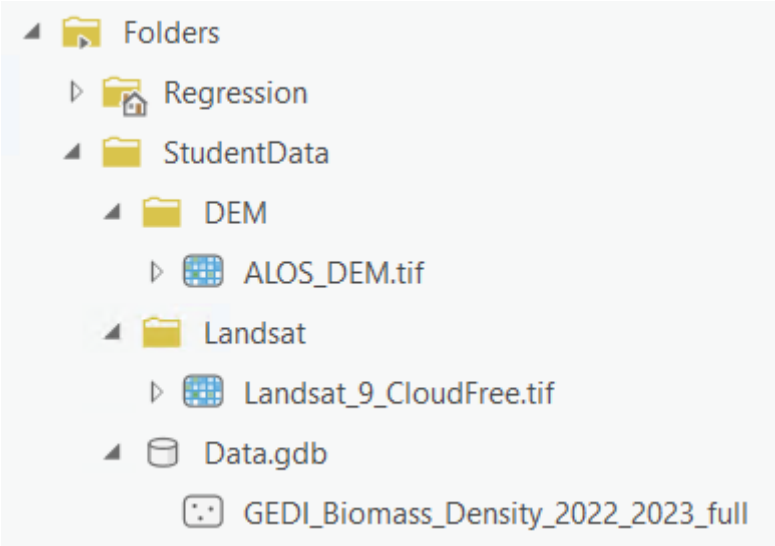
Регрессия Random Forest может выглядеть следующим образом:



Красные точки – реальные данные, синие точки – предсказанные. То есть, зависимость может быть нелинейная.

ДАННЫЕ ДЛЯ УПРАЖНЕНИЯ

В директории StudentData можно найти базу геоданных Data.gdb с точечным слоем GEDI_Biomass_Density_2022_2023_full (измерения сенсора GEDI), сцену Landsat9 и цифровую модель рельефа ALOS_DEM.tif.



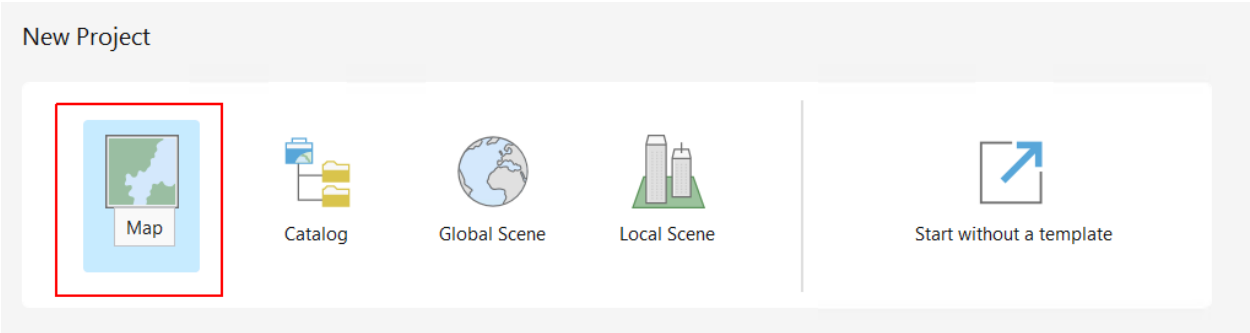
Если в вашем учебном компьютере нет этих данных, их можно скачать по ссылкам:

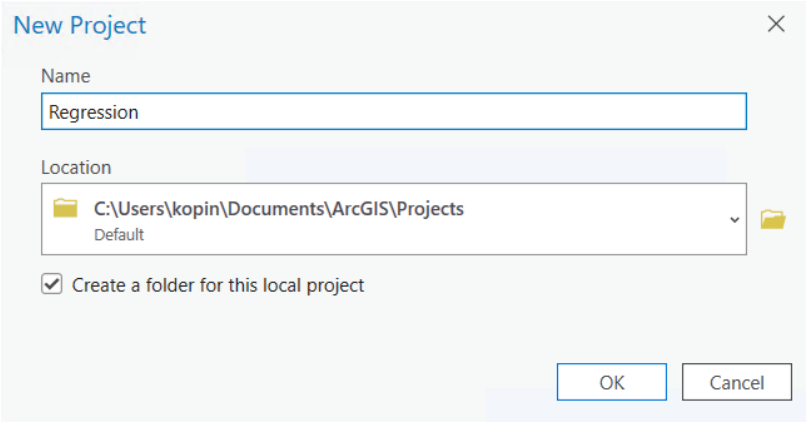
https://downloads.esri-cis.com/files/ykopin/StudentData.zip

https://t.ly/D3ixb

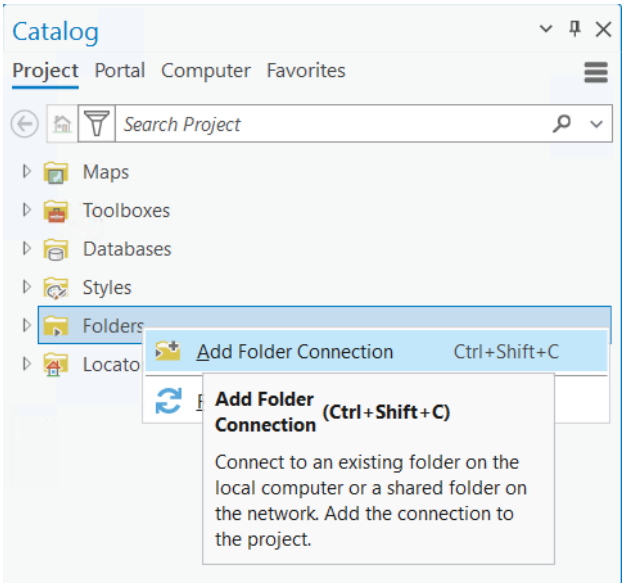
СОЗДАНИЕ ПРОЕКТА И ЗНАКОМСТВО С ИСХОДНЫМИ ДАННЫМИ

Создайте новый проект с картой и назовите его Regression.

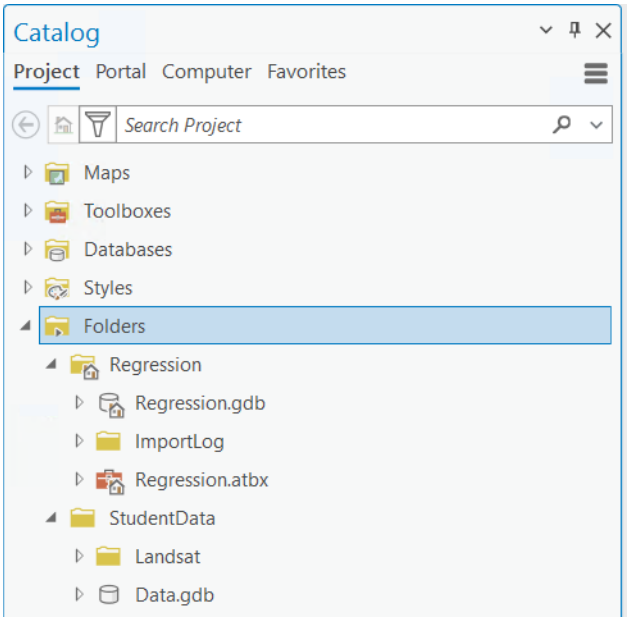




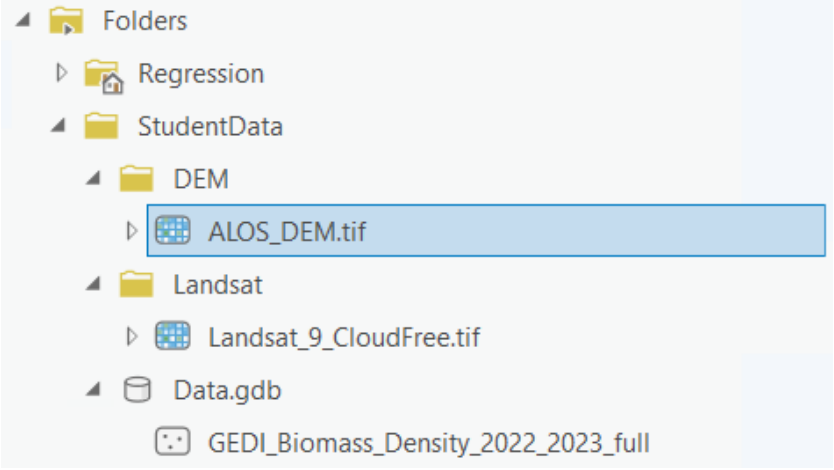
Далее, подключите к проекту директорию с учебными данными, которая называется StudentData и должна лежать в корне на диске C. Для этого в Каталоге выберите раздел Folders, нажмите правую кнопку мыши и выберите Add Folder Connection.



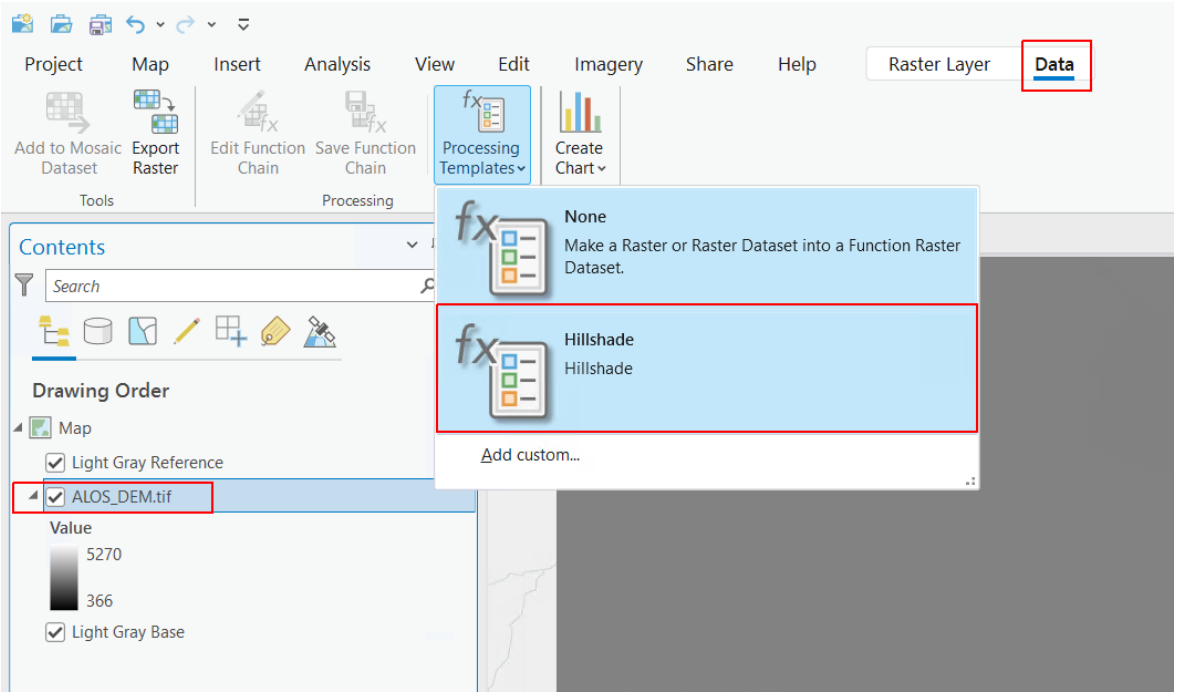
Результат должен выглядеть следующим образом:



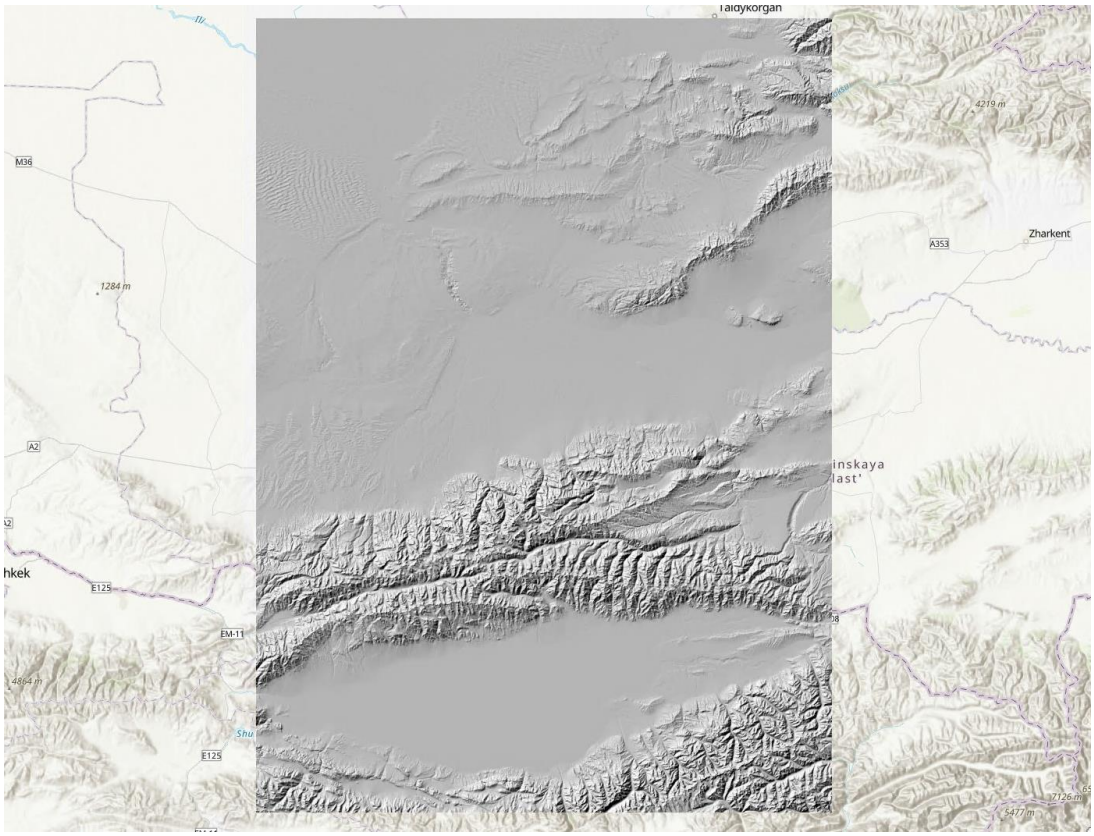
В проекте должна быть рабочая директория Regression с базой геоданных по умолчанию и отдельная директория с данными для упражнения. Выберите ЦМР ALOS_DEM.tif и перетащите мышкой на карту.



В ArcGIS есть возможность прикреплять функции обработки растров к растровым датасетам. В данном случае в ЦМР добавлена дополнительная функция отмывки рельефа и мользователь может переключаться между высотами и отмывкой. Сделать это можно следующим образом. Выберите слой DEM_ALOS в таблице содержания. В верхнем меню перейдите на закладку Data, далее нажмите на кнопку Processing Templates и выберите Hillshade.

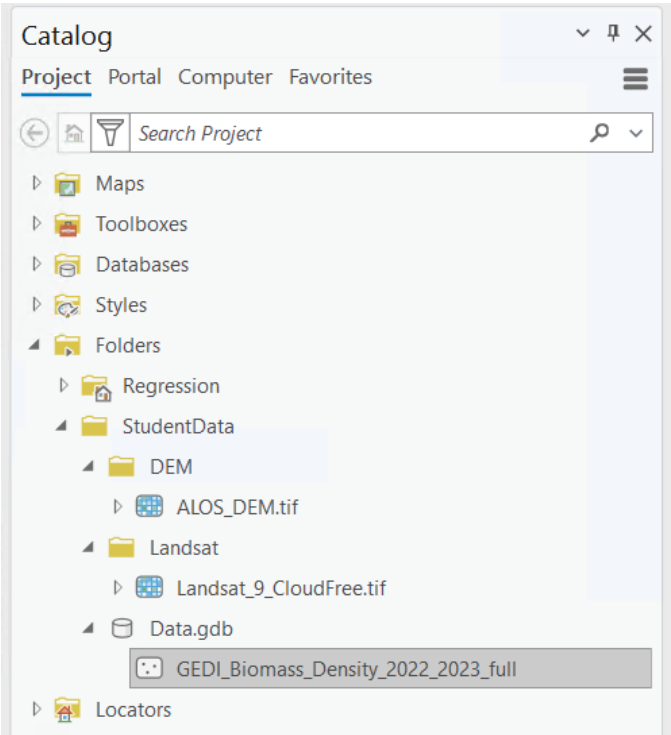


Результат должен выглядеть следующим образом.

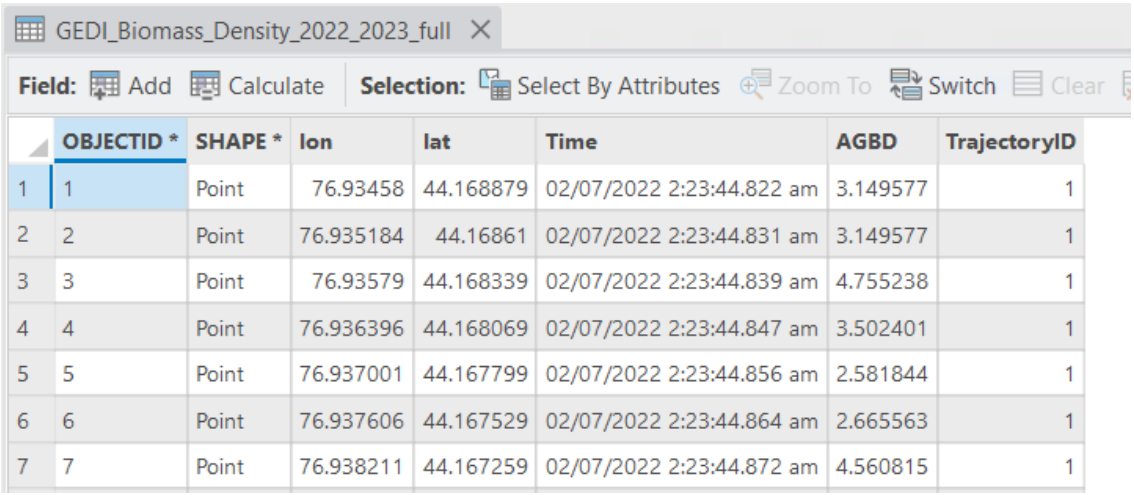


Переключитесь обратно с шаблона Hillshade на шаблон None и выключите видимость слоя DEM_ALOS в окне содержания (Content).

Далее добавим на экран точечный слой GEDI_Biomass_Density_2022_2023_Full

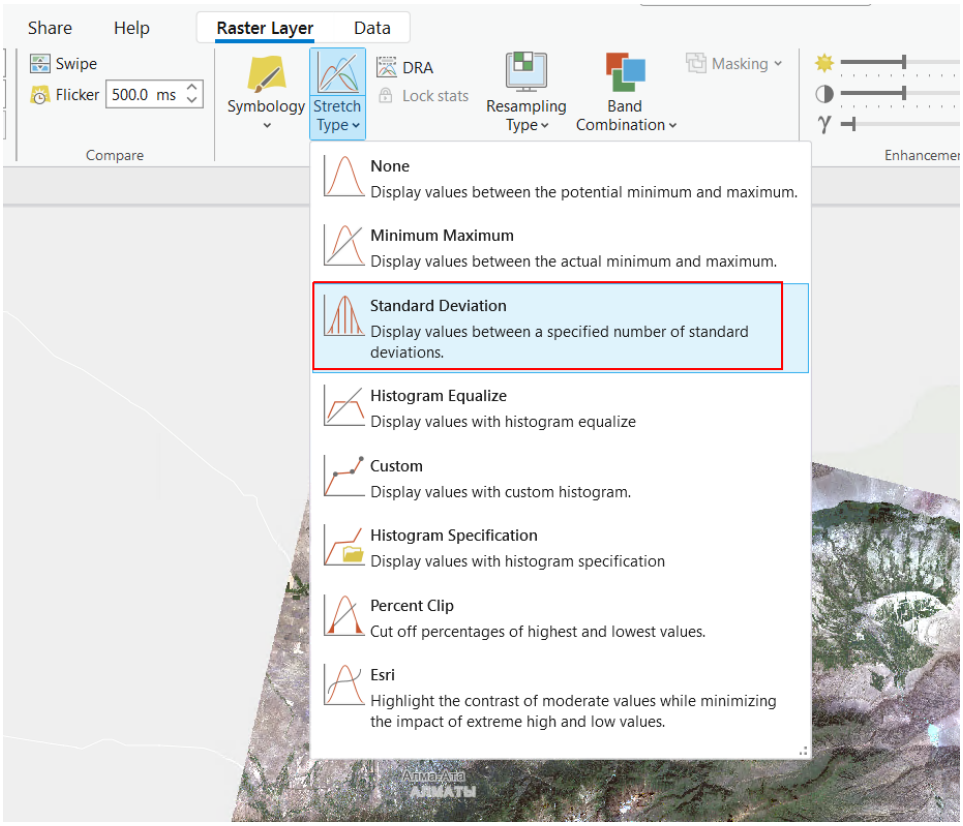


С помощью контекстного меню, откройте таблицу атрибутов этого слоя. В нем присутствуют поля с координатами точек, временем съемки, ID траектории и основное интересующее нас поле AGBD, которое расшифровывается как Aboveground Biomass Density, т.е. оценка биомассы растительности по данными измерений с помощью лазерного луча. На эту оценку влияет как высота растительности, так и ее плотность.



	OBJECTID *	SHAPE *	lon	lat	Time	AGBD	TrajectoryID
1	1	Point	76.93458	44.168879	02/07/2022 2:23:44.822 am	3.149577	1
2	2	Point	76.935184	44.16861	02/07/2022 2:23:44.831 am	3.149577	1
3	3	Point	76.93579	44.168339	02/07/2022 2:23:44.839 am	4.755238	1
4	4	Point	76.936396	44.168069	02/07/2022 2:23:44.847 am	3.502401	1
5	5	Point	76.937001	44.167799	02/07/2022 2:23:44.856 am	2.581844	1
6	6	Point	76.937606	44.167529	02/07/2022 2:23:44.864 am	2.665563	1
7	7	Point	76.938211	44.167259	02/07/2022 2:23:44.872 am	4.560815	1

Закройте таблицу атрибутов. Выключите видимость слоя GEDI_Biomass_Density_2022_2023_Full и добавьте на карту растр Landsat_9_CloudFree.tif из директории Landsat. После чего переключите метод растяжки гистограммы растра с None на Standard Deviation или Percent Clip.



Мы добавили на карту три базовых набора данных, которые будет использовать для регрессионного анализа.

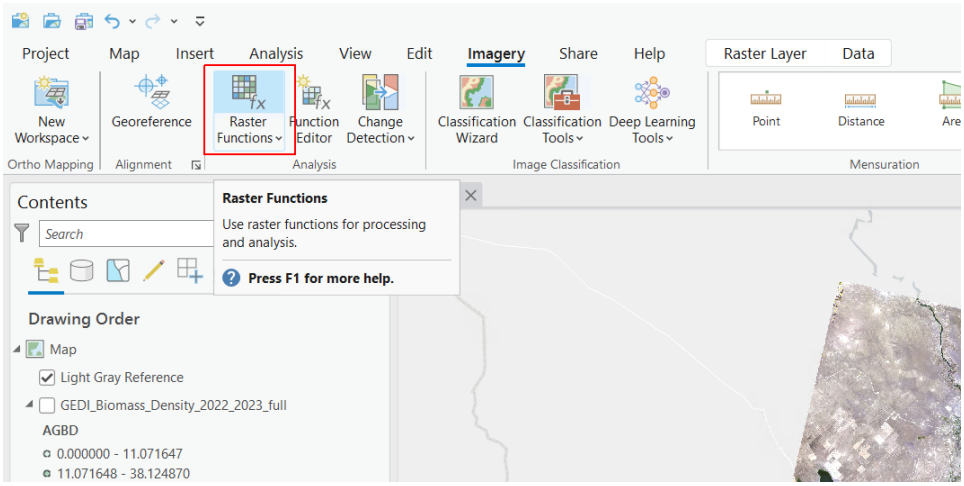
ГЕНЕРАЦИЯ ДОПОЛНИТЕЛЬНЫХ ПОВЕРХНОСТЕЙ И ИНДЕКСОВ

Какие переменные в теории могут влиять на биомассу растительности? На данный момент у нас есть информация об абсолютной высоте поверхности. От высоты зависят среднегодовая температура воздуха, соответственно этот параметр может повлиять на наличие или отсутствие растительности. Также у нас есть снимок Landsat, который содержит информацию о состоянии растительности. Мы может подсчитать дополнительные вегетационные индексы – NDVI и SAVI. Чем выше значение вегетационного индекса, тем скорее всего выше значение биомассы. Еще можно подсчитать уклон поверхности и экспозицию. Например, в горной местности деревья и кустарники лучше растут на южном склоне в определенном диапазоне углов наклона поверхности. Таким образом, для улучшение качества модели попробуем использовать дополнительные параметры, которые мы сгенерируем на основе исходных данных. Всего будет 5 исходных растров:

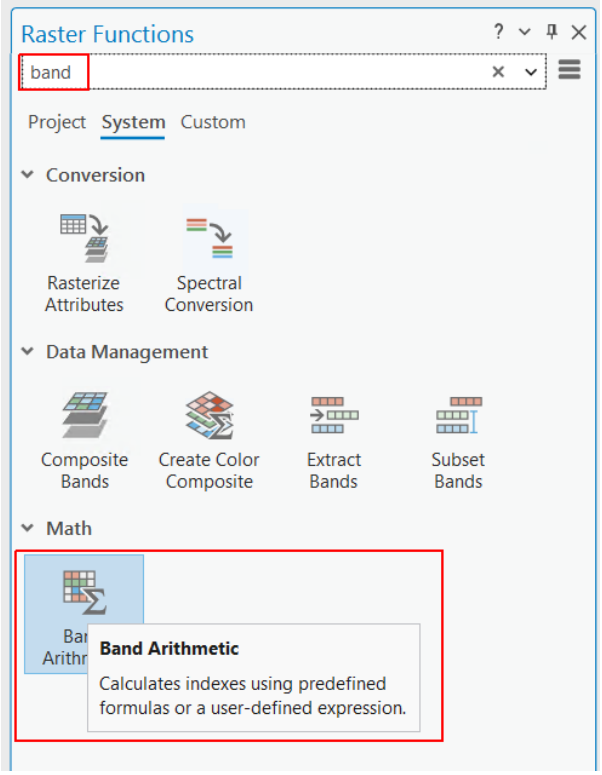
- Сцена Landsat 9
- Индекс NDVI
- Индекс SAVI
- Абсолютная высота (ЦМР)
- Экспозиция слона
- Уклон поверхности

ГЕНЕРАЦИЯ РАСТРОВ С ИНДЕКСАМИ NDVI И SAVI

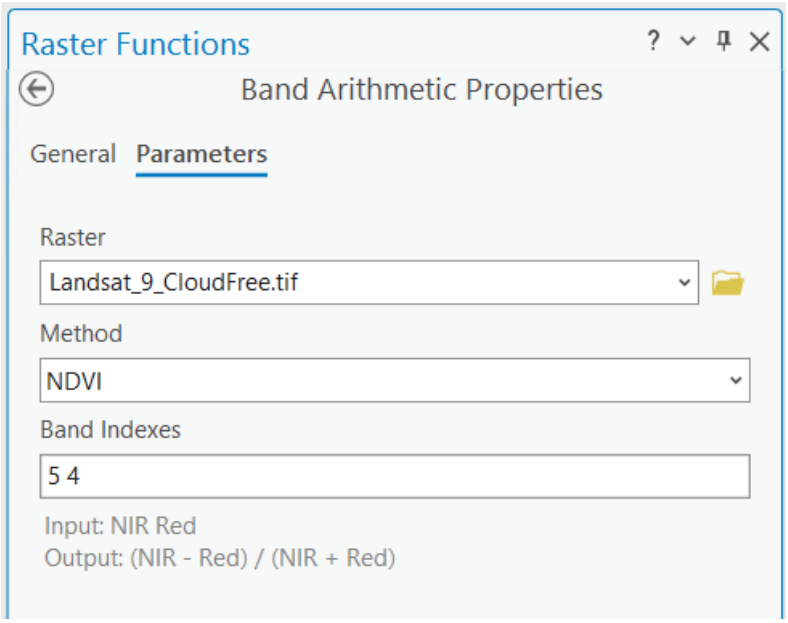
В ArcGIS есть так называемые растровые функции (Raster Functions). Это инструменты, которые позволяют выполнять обработку растров без сохранения результатов на диск. Результатом выполнения растровой функции (или цепочки функций) является временный растр, который можно использовать для расчетов и сохранить на диск, в случае необходимости. Далее мы применим растровую функцию к сцене Landsat 9. Перейдите в закладку Imagery верхней панели интерфейса ArcGIS Pro и нажмите на кнопку Raster Functions.



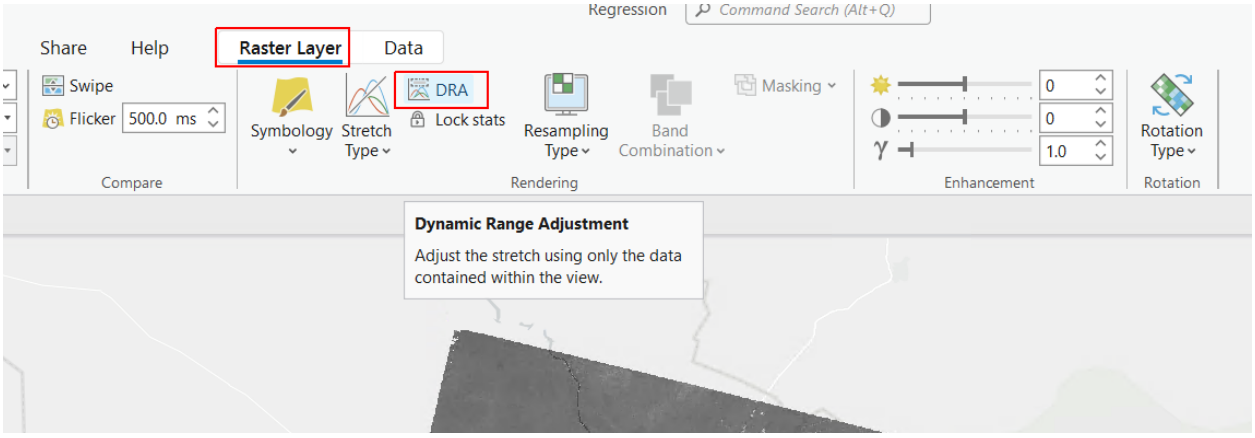
В появившемя окне растровых функций найдите функцию Band Arithmetic, сделать это можно с помощью текстового фильтра в верхней части окна (наберите в нем слово “band”).



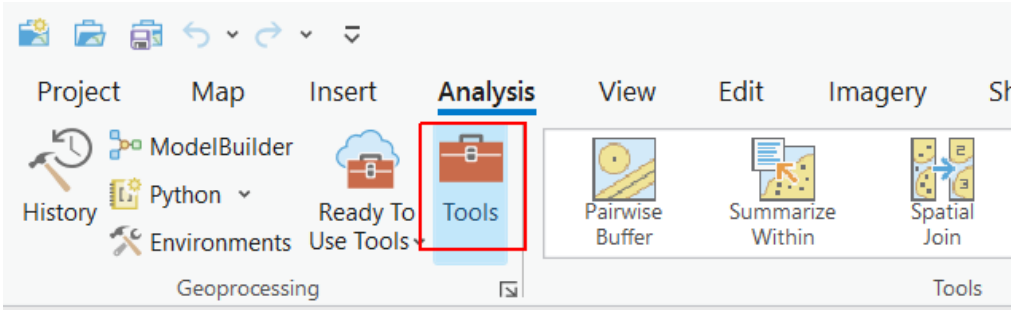
Эта функция дает возможность производить арифметические операции с каналами и содержит готовые шаблоны для расчета различных индексов. В качестве входного растра (Raster) укажите Landsat_9_CloudFree.tif, выберите метод NDVI и укажите каналы 5 и 4 (NIR и Red), далее нажмите кнопку Create New Layer в нижней части окна.



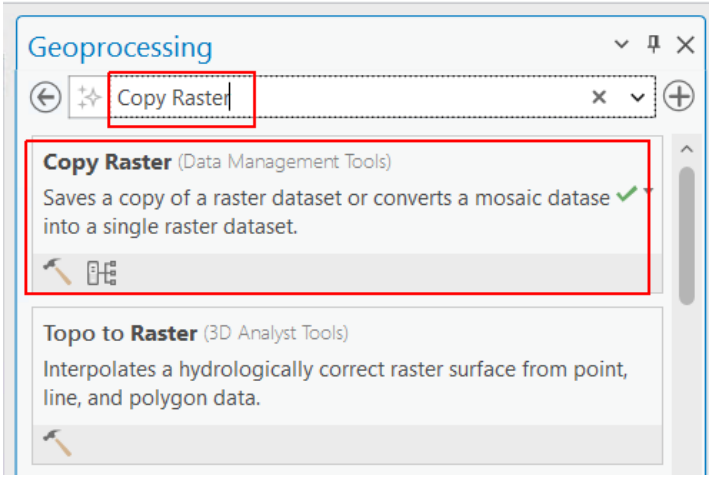
На карту будет добавлен новый виртуальный растр с рассчитанным индексом NDVI. Для того, чтобы улучшить его читаемость, включите динамическую растяжку гистограммы. В закладке Raster Layer нажмите кнопку DRA. Сейчас у растра не подсчитана статистика, включение опции DRA позволит рассчитывать ее динамически для текущего экстента.



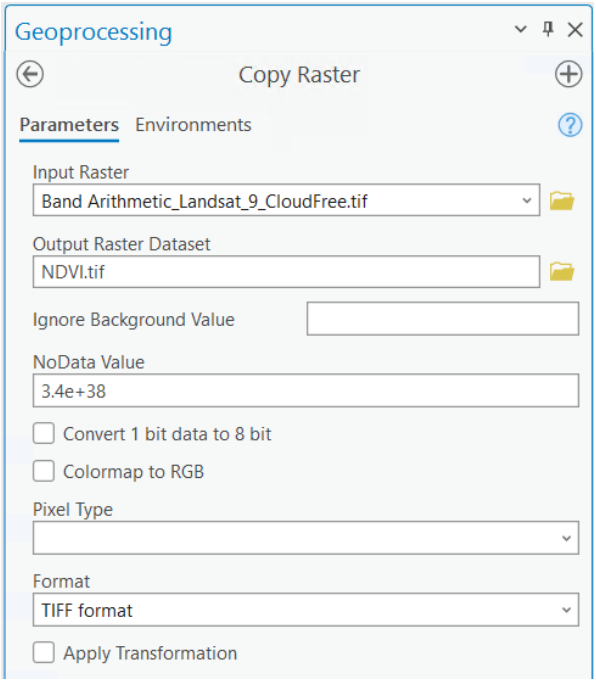
Сохранение растра на диск. В верхнем меню выберите закладку Analysis и нажмите на кнопку Tools, чтобы открыть окно ArcToolbox (инструментов геообработки).



В окне поиска инструментов по названию наберите Copy Raster и выберите инструмент Copy Raster из списка предложенного списка.



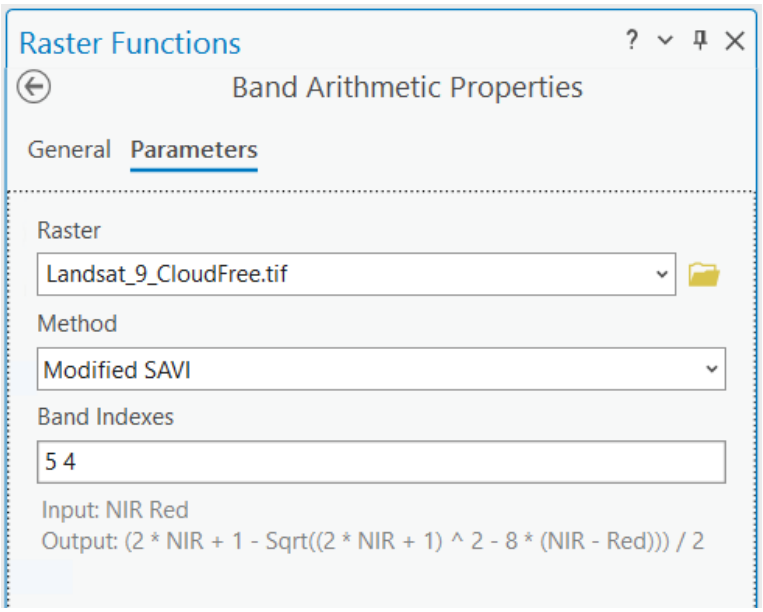
Скопируйте растр NDVI на диск в директорию Landsat под названием NDVI.tif



Запустите инструмент с помощью кнопки Run.

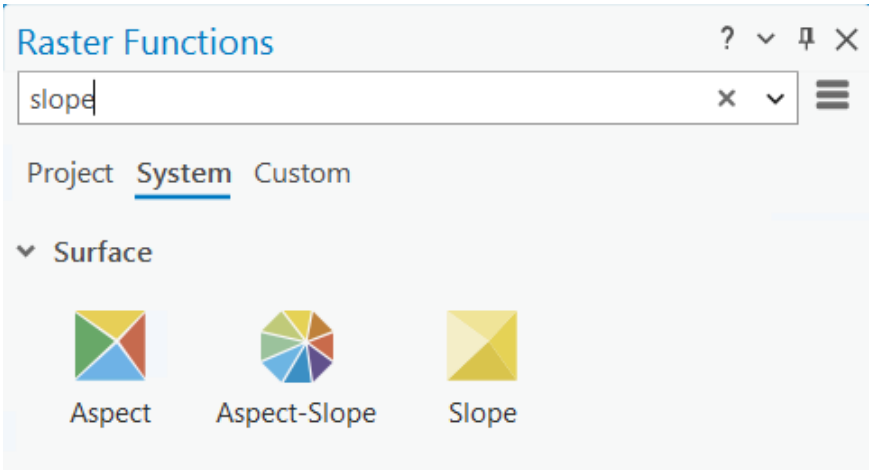
Появится новый растр NDVI.tif в таблице содержания карты. Уберите виртуальный растр Band Arithmetic_Landsat_9_CloudFree.tif из таблицы содержания (правая кнопка мыши, контекстное меню, Remove). Опционально можете включить для растра NDVI динамическую растяжку гистограммы (см. предыдущую страницу).

Теперь проделаем те же операции для создания растра SAVI. Применим растровую функцию Band Arithmetic к сцене Landsat 9. Далее сохраним виртуальный растр на диск с названием SAVI.tif в директорию Landsat.

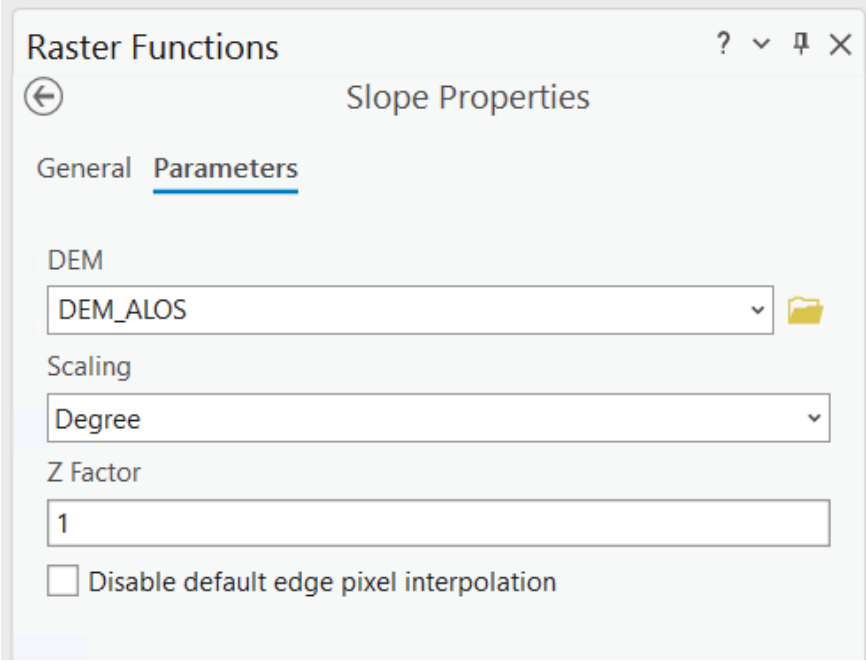


СОЗДАНИЕ РАСТРОВ УКЛОНА И ЭКСПОЗИЦИИ СКЛОНА

Для создания растра уклона поверхности воспользуемся растровой функцией Slope.



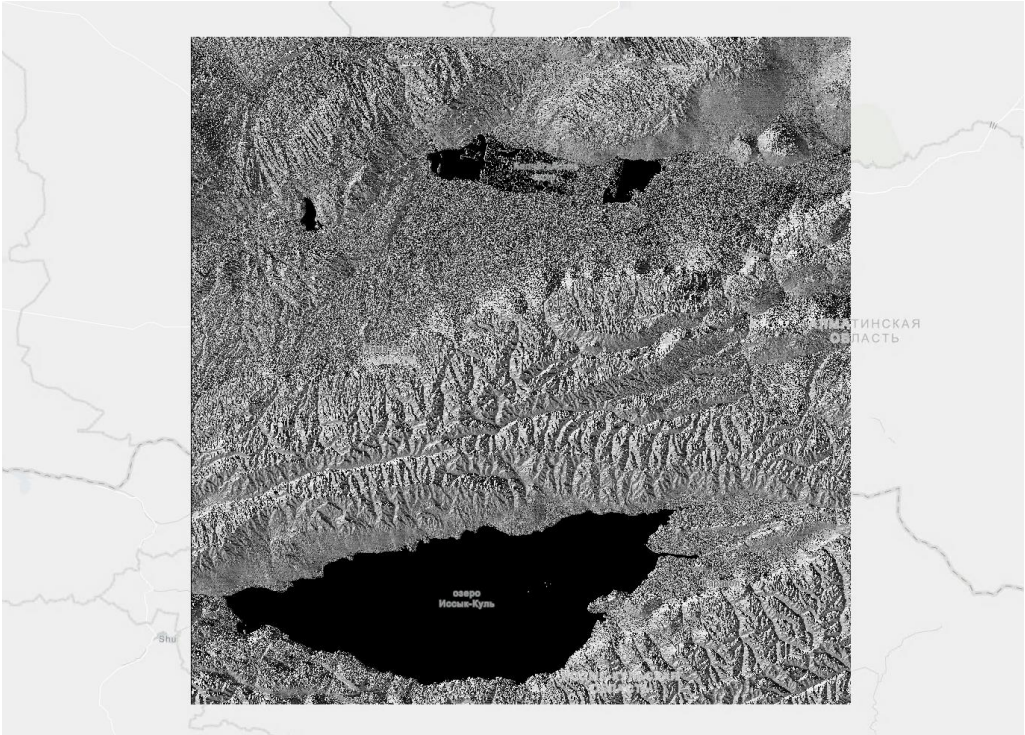
На вход функции нужно подать цифровую модель рельефа ALOS_DEM.tif. Единицей измерения угла наклона поверхности будут градусы. Значение параметра Z Factor оставим равным 1.



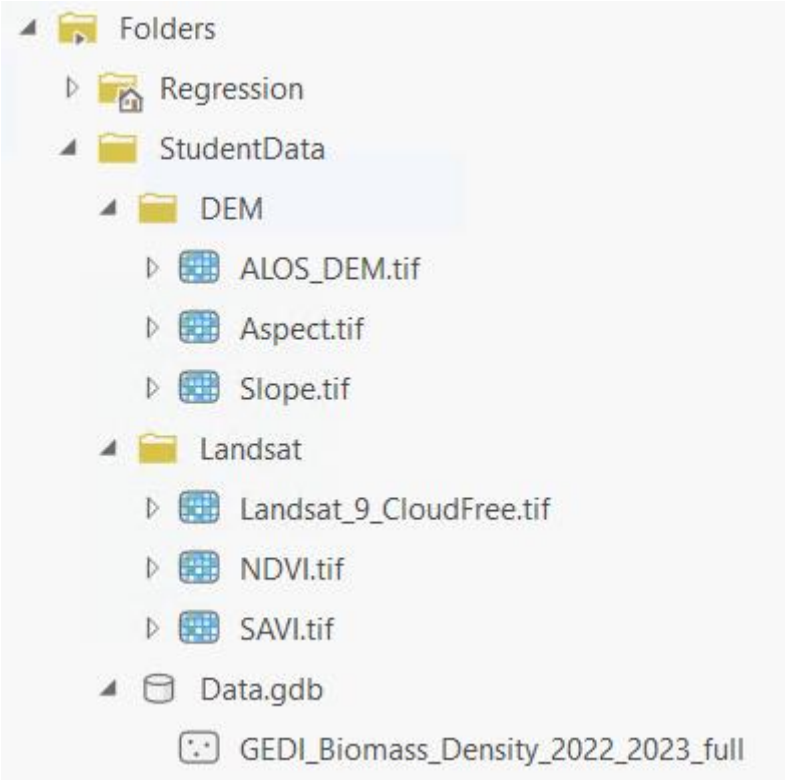
Далее полученный растр нужно сохранить в директории DEM под названием Slope.tif.

И последний растр, который нужно подготовить, это растр экспозиции склонов. Для этого в окне растровых функций выберите функцию Aspect и примените ее к цифровой модели рельефа ALOS_DEM.tif. У функции всего один параметр – входной растр. Сохраните растр на диск в директорию DEM под названием Aspect.tif.

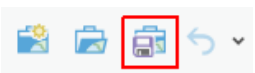
Должен получиться растр экспозиции склона.



В итоге должно получиться 6 растров на базе которых будет построена регрессионная модель.



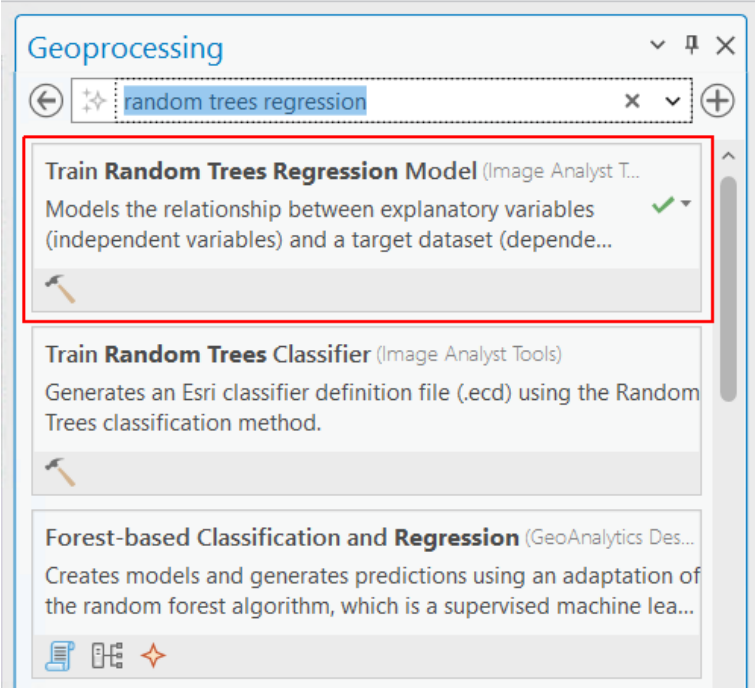
Сохраните проект нажав соответствующую кнопку в левом верхнем углу.



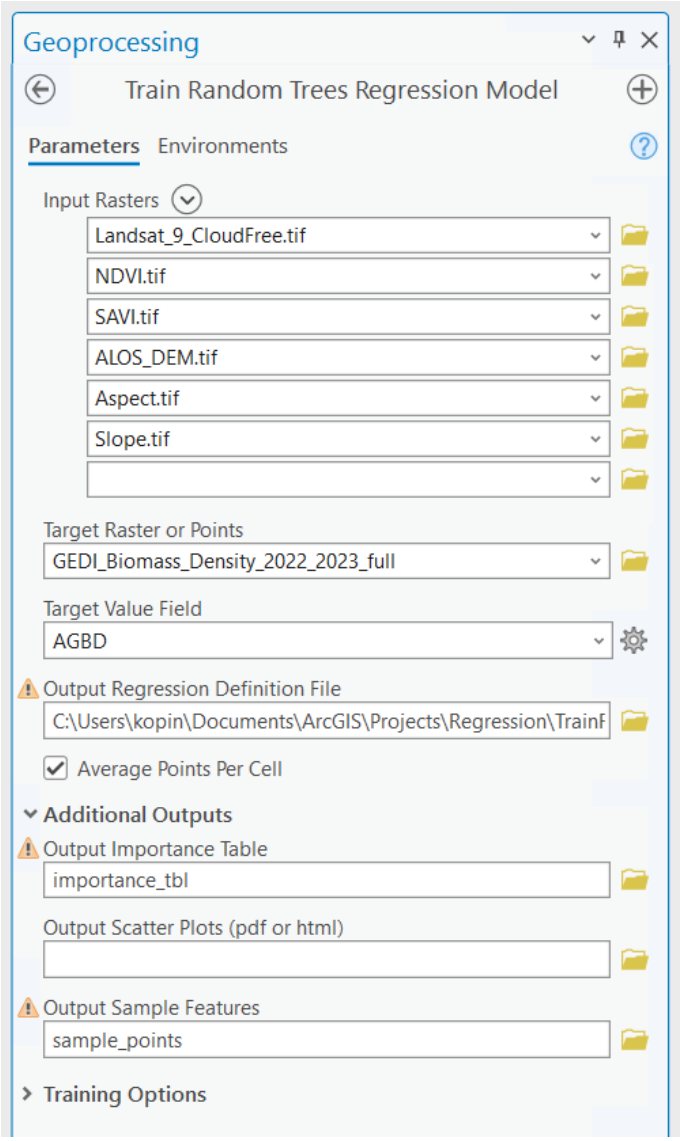
ОБУЧЕНИЕ РЕГРЕССИОННОЙ МОДЕЛИ НА БАЗЕ АЛГОРИТМА RANDOM FOREST

У нас есть опорные данные сенсора GEDI с измерениями биомассы растительности, которые безусловно содержат погрешности, но поскольку других источников у нас нет, мы будем считать эти данные основой для построения предсказательной (регрессионной) модели. Модель должна будет обучиться предсказывать значения биомассы в точках, в которых есть измерения GEDI, минимизировать ошибку предсказания и потом эту модель можно будет применить для предсказания значения биомассы на участках, для которых нет измерений сенсора GEDI. Обучающая выборка будет сформирована на базе 5 растров и данных сенсора GEDI.

В ArcGIS есть несколько инструментов построения регрессионных моделей, мы воспользуемся одним из них. Инструмент называется Train Random Trees Regression Model, его можно найти в ArcToolbox в наборе инструментов модуля Image Analyst. Откройте окно ArcToolbox и наберите в строке поиска **random trees regression**. Выберите нужный инструмент.



Этот инструмент оптимизирован для работы с растровыми данными. В его параметрах нужно будет указать входные растры, которые будут использоваться для обучения регрессионной модели, а также растровый или точечный слой, который содержит опорные данные, которые мы считаем эталоном.



Input Rasters – входные растры, на базе которых будет обучаться регрессионная модель.

Target Raster or Points – опорные данные, в данном случае это точечный слой с измерениями сенсора GEDI.

Target Value Field – поле точечного слоя, в котором содержится информация о значениях объема биомассы.

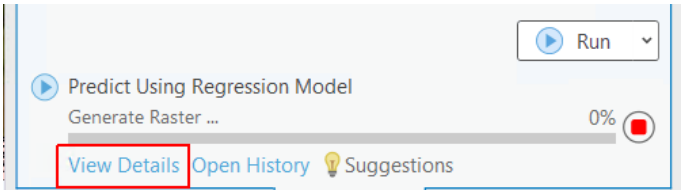
Output Regression Definition File – выходной файл с параметрами обученной регрессионной модели, сохраните в директорию по умолчанию или в любое удобное место на диске.

Average Points Per Cell – данный параметр определяет, что делать если на две и более точек с опорными данными попадают в один пиксель входных растров. Если установить флажок, то значения в точках опорных данных будет усредняться внутри одного пикселя.

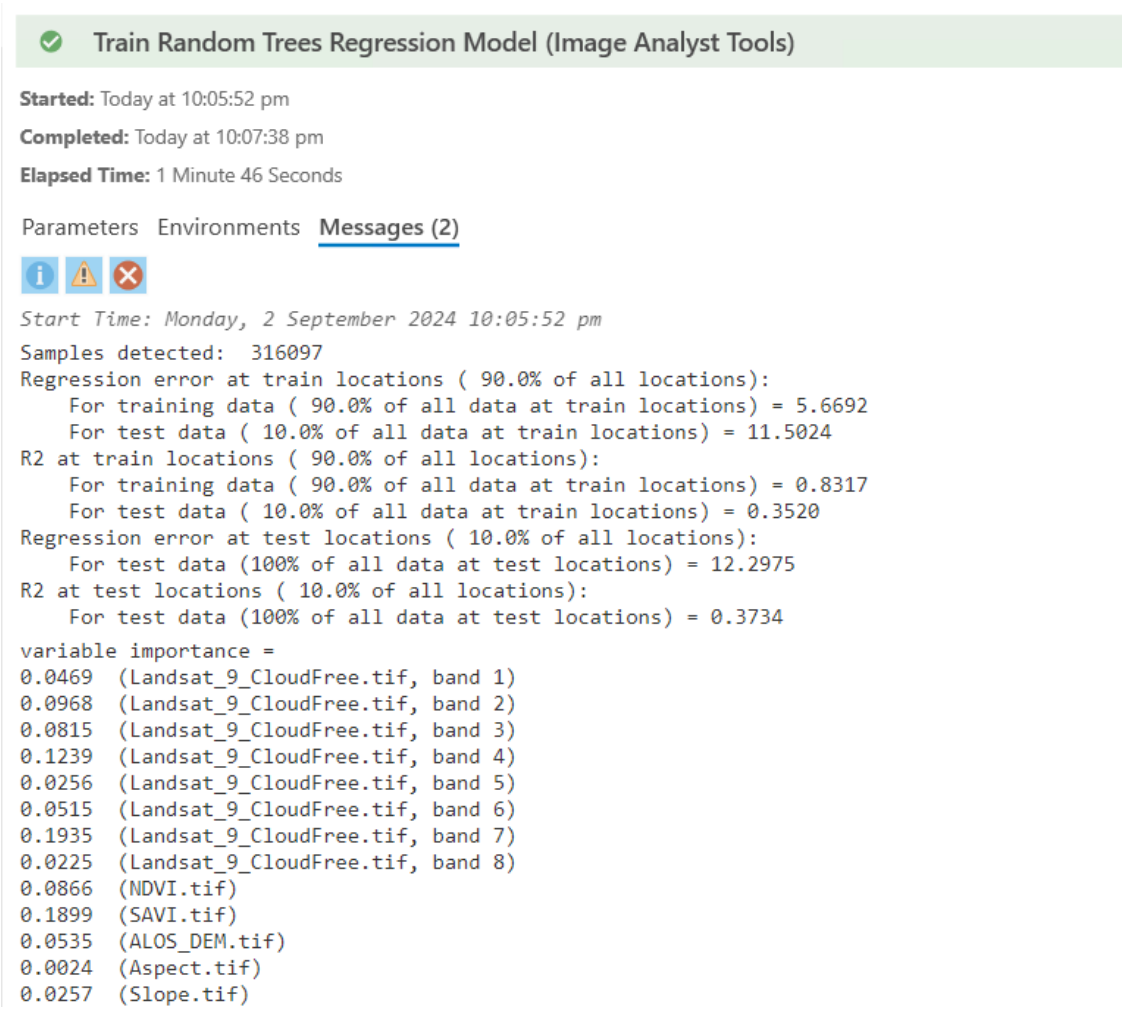
Output Importance Table – выходная таблица с коэффициентами значимости каждого входного параметра (растра) с точки зрения вклада в конечный результат.

Output Sample Features – выходной точечный слой, который будет содержать два поля: опорные значения и предсказанные.

Запустите инструмент и нажмите на кнопку View Details в нижней части окна.



После того как инструмент закончит выполнение задачи, перейдите в закладку Messages. В окне Details показаны сообщения инструмента, в которых можно прочитать различную полезную информацию. В данном случае нас интересуют значения variable importance. Значения у вас могут немного отличаться от тех, которые показаны на скриншоте, потому что процесс обучения модели содержит в себе элемент случайности. Наибольшее влияние на конечный результат оказал четвертый канал сцены Landsat (красный), седьмой канал сцены Landsat (SWIR) и индекс SAVI. Входные параметры (растры) могут сильно коррелировать друг с другом и показывать практически одно и то же, для алгоритма Random Forest это не является проблемой (в отличие от алгоритмов линейной регрессии).

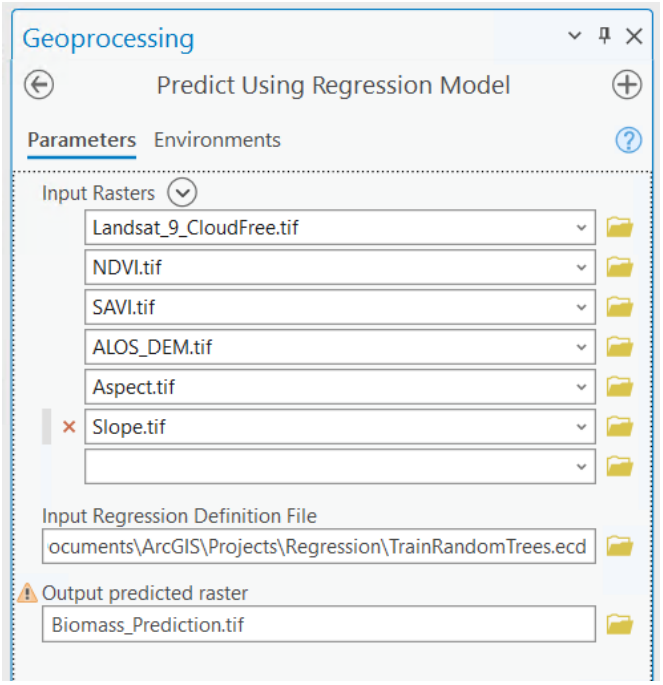


Также в этом отчете можно посмотреть на метрики точности регресионной модели. Для каждой точки опорных данных есть значение, которое мы считаем правильным, оно сохранено в поле AGBD, и есть значение, которое предсказывает регрессионная модель. Разница между этими значениями – ошибка прогноза. Также в отчете есть две категории - train locations и test locations. Train locations используются для обучения регрессионной модели, а test locations исползуются только для оценки точности. В данном случае 90% точек с опорными данными использовались как train locations, 10% как test locations. На что нужно обратить внимание? Прежде всего на коэффициент R2. Он у нас равен 0.83 для обучающей выборки и 0.35 для тестовой. В идеале эти значения должны примерно

совпадать, что будет говорить о том, что модель достаточно универсальна, не страдает от “переобучения” и хорошо работает с любыми данными. Но идеал достигается редко. В данном случае скорее всего тестовая выборка слишком мала и не очень репрезентативна. Не исключено, что можно улучшить результаты либо за счет увеличения тестовой выборки, либо за счет увеличения всего набора обучающих данных. Но мы этого делать не будем, остановимся на текущем варианте. Значение 0.83 говорит о том, что точность предсказания примерно соответствует 83%. Понятно, что у сенсора есть определенная погрешность, у модели есть определенная погрешность, все это накладывается и поэтому результат 0.83 достаточно неплохой. Далее мы сможем его визуализировать на диаграмме.

ПРИМЕНЕНИЕ РЕГРЕССИОННОЙ МОДЕЛИ

Найдите другой инструмент, который называется **Predict Using Regression Model**. На вход этого инструмента нужно подать те же самые растры, которые использовались для обучения регрессионной модели. Порядок растров должен быть такой же, как при обучении модели. В качестве выходного растра укажите Biomass_Prediction.tif.



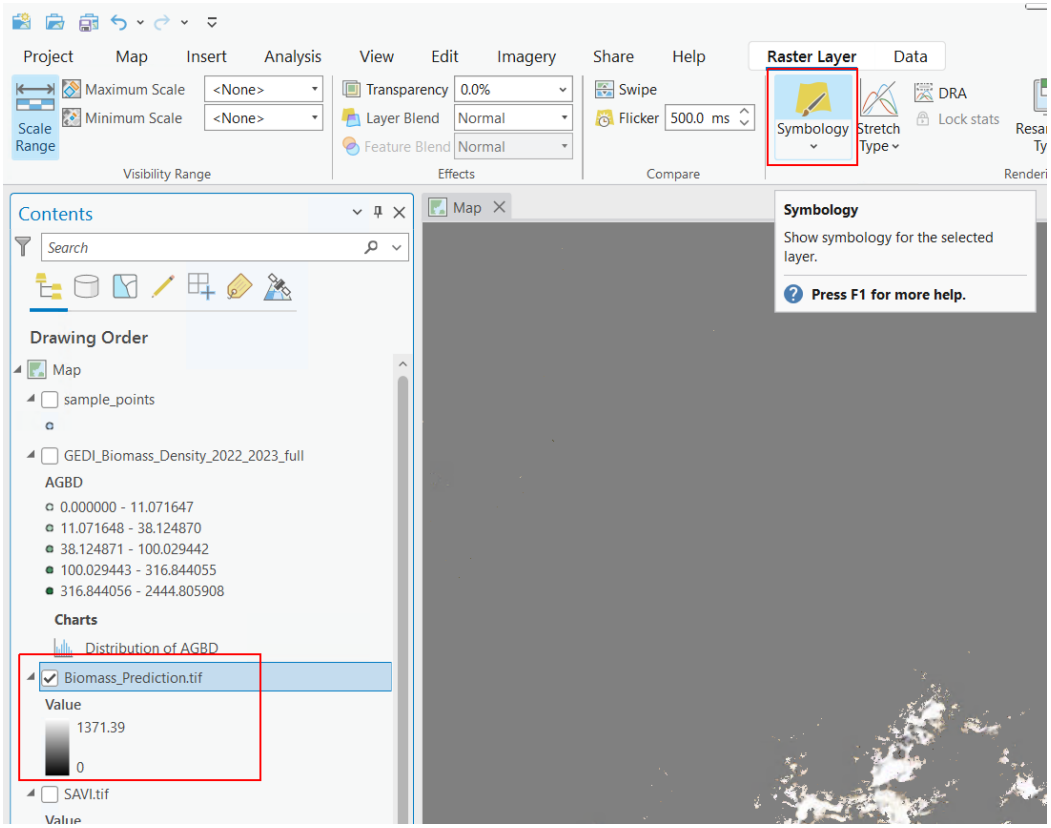
Input Rasters – входные растры

Input Regression Definition Flie – файл с расширением ecd, который мы сгенерировали на предыдущем этапе (обучение модели)

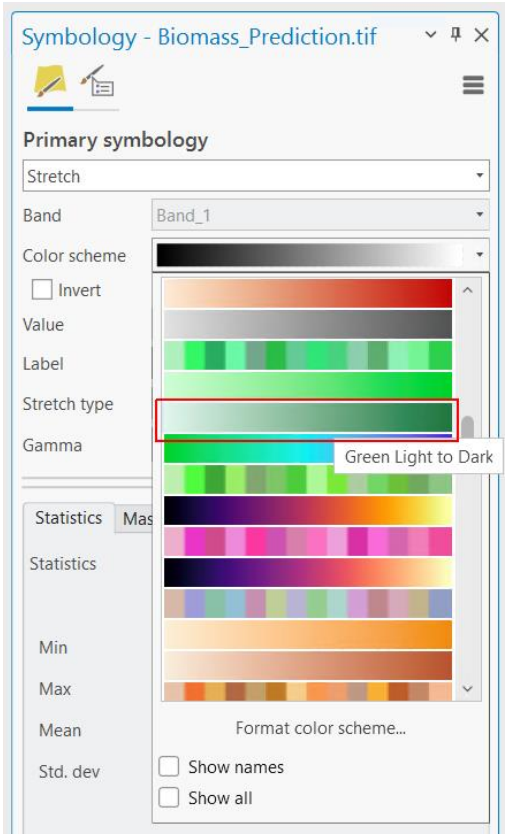
Output predicted raster – выходной растр с значениям биомассы, предсказанными с помощью регрессионной модели

Запустите инструмент и дождитесь появления результата на карте.

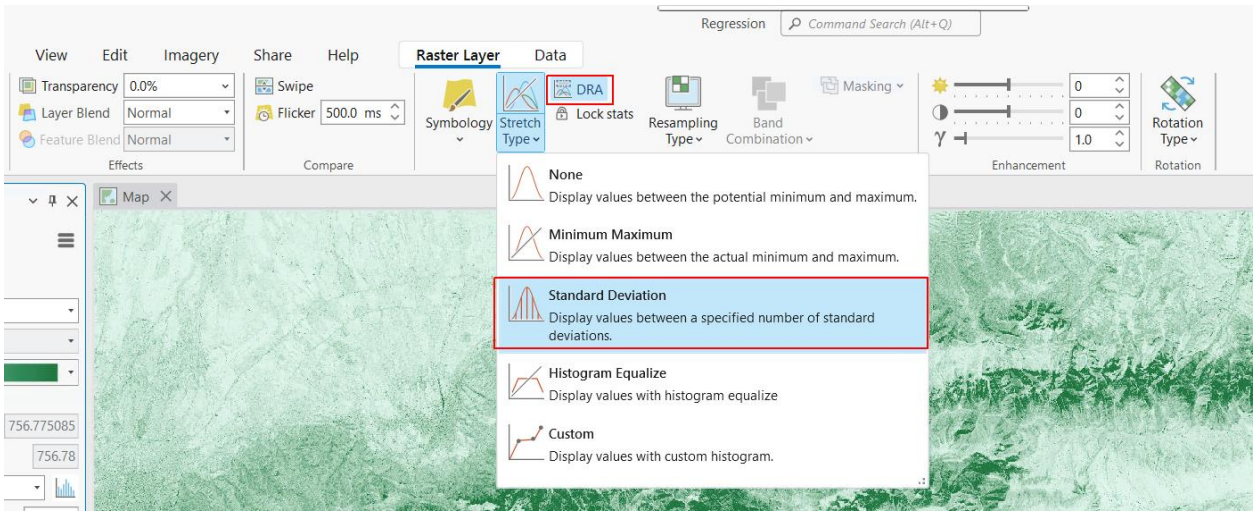
Выберите растр Biomass_Prediction в таблице содержания и нажмите на кнопку Symbology в закладке Raster.



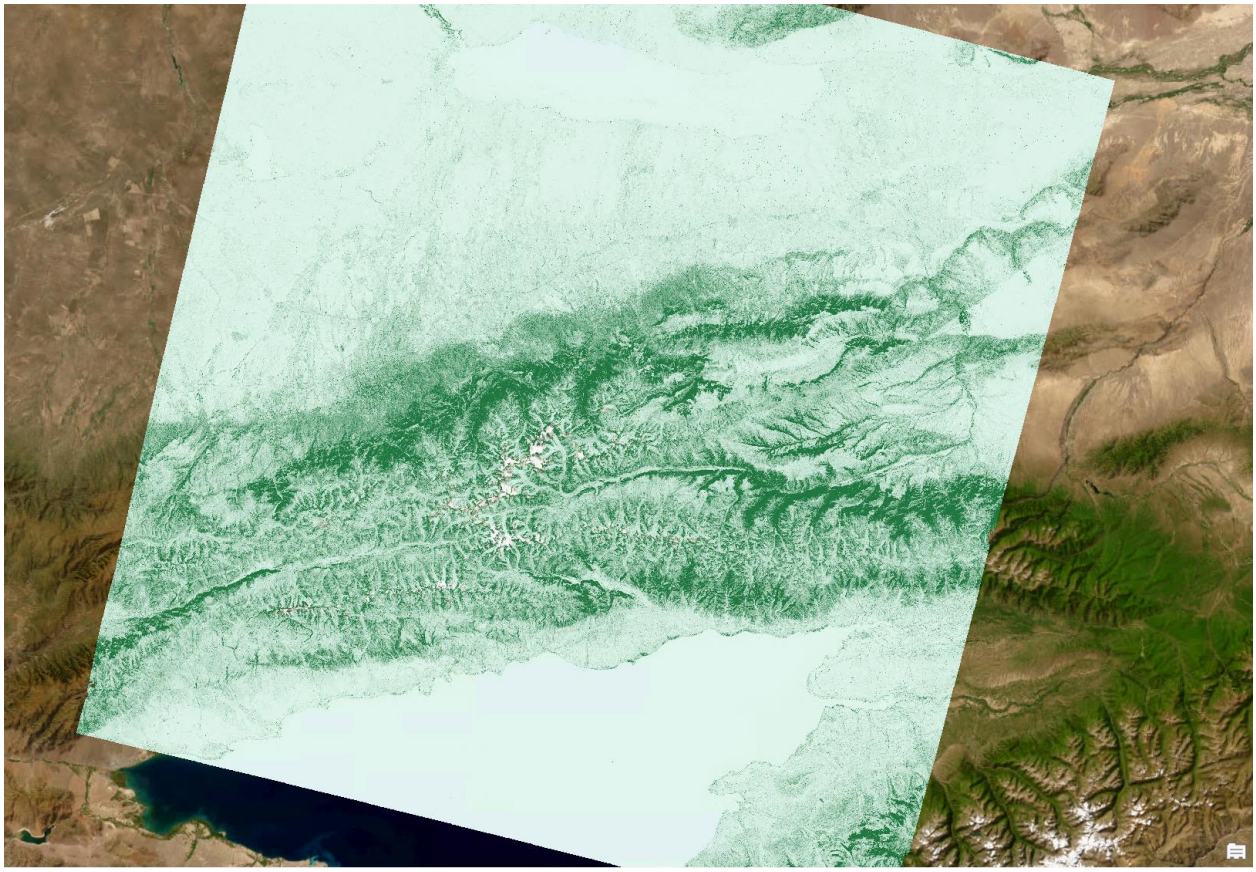
Откроется окно настройки отображения растра. Выберите темно-зеленую шкалу.



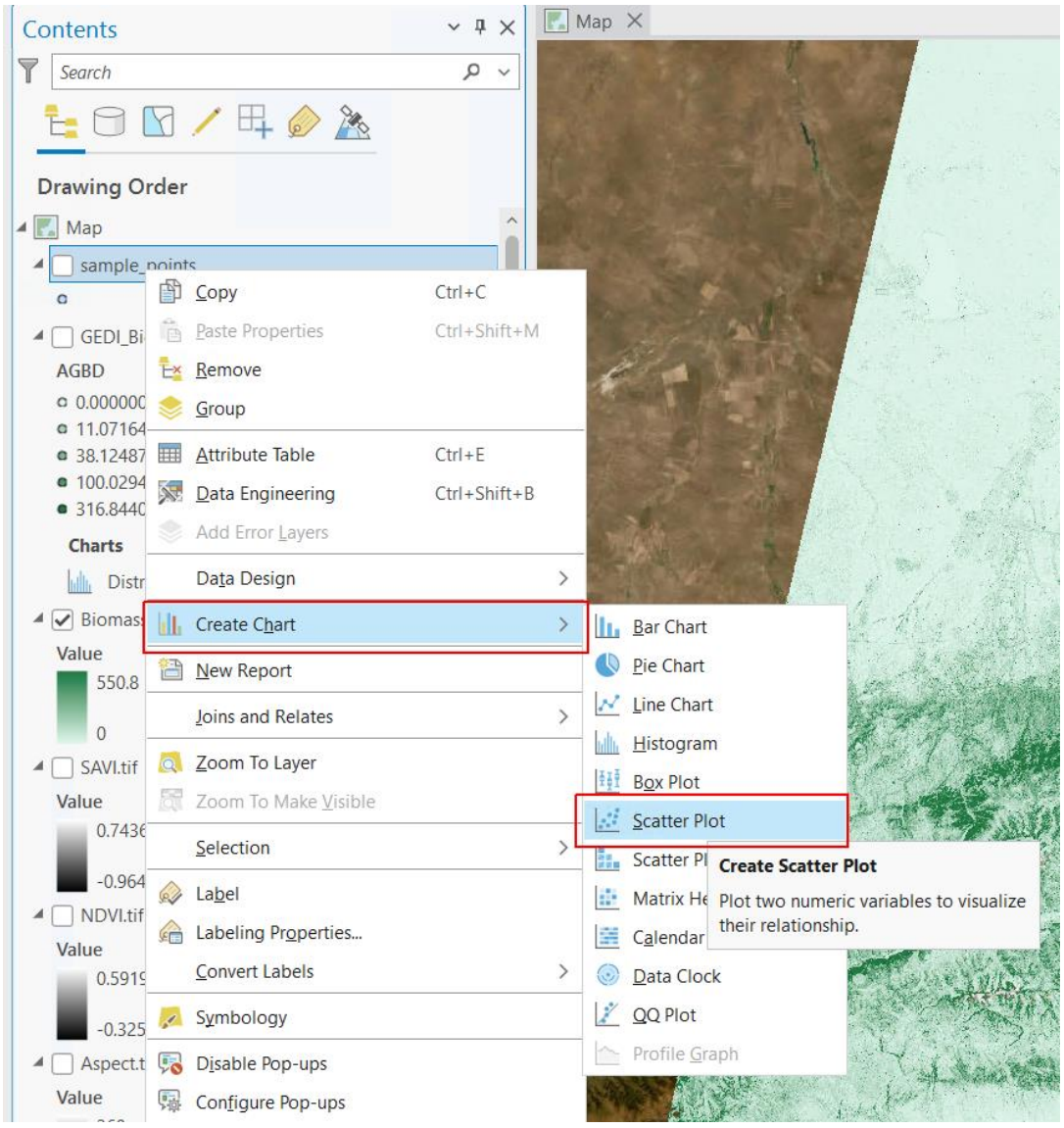
Далее нажмите кнопку DRA и выберите Standard Deviations в качестве способа растяжки гистограммы.

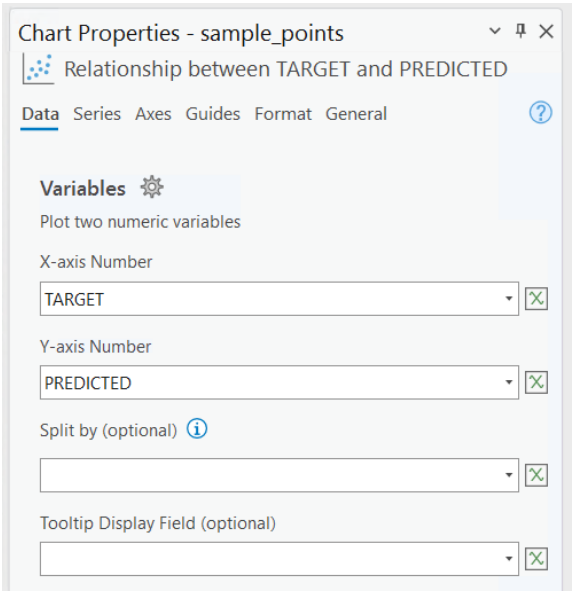


Теперь можно посмотреть на результат прогнозирования биомассы растительности с помощью регрессионной модели. Для обучения использовались отдельные точки с измерениями сенсора GEDI, регрессионая модель позволила предсказывать значения биомассы растительности на базе набора растров.



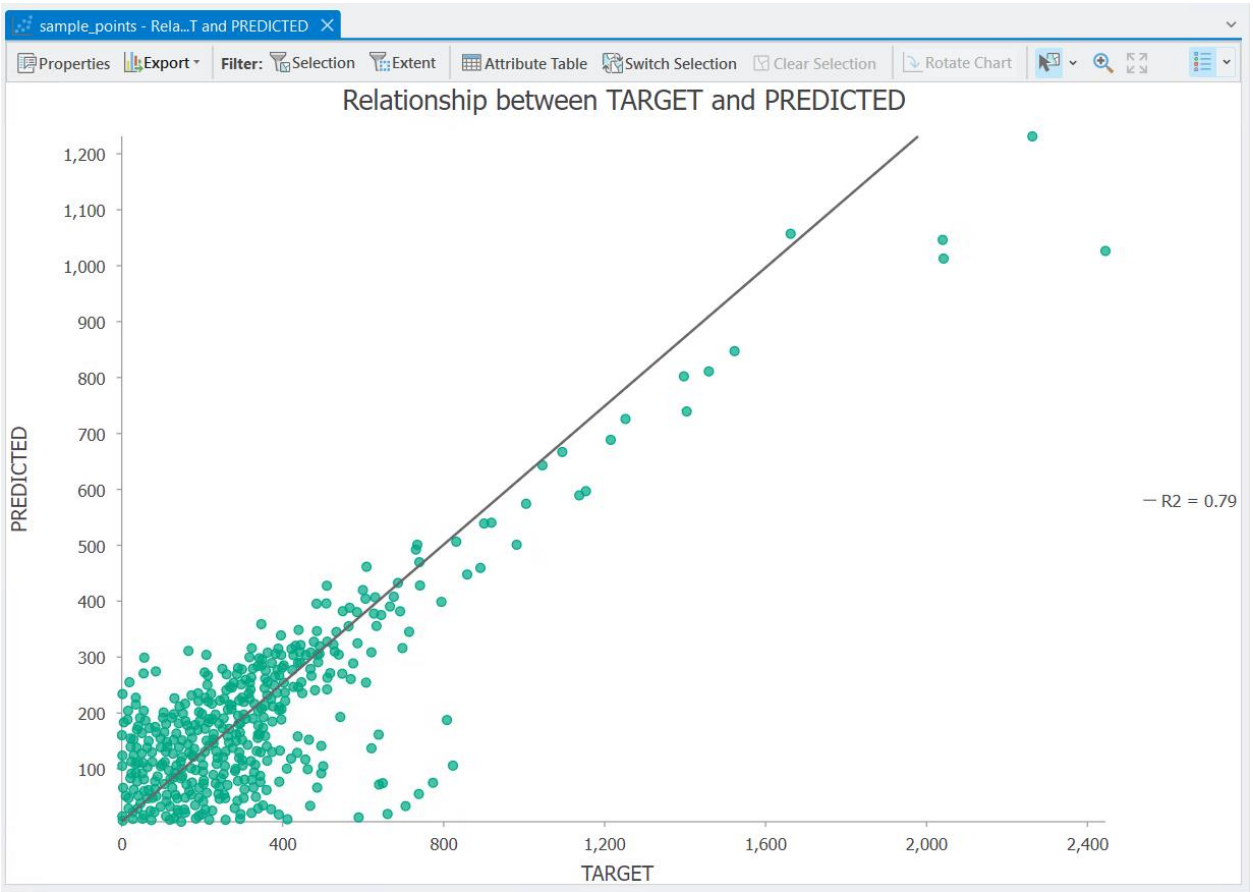
Завершающий шаг, визуализация диаграммы, отображающей соотношение между опорными значениями и прогнозом. Воспользуйте слоем sample_points, который вы сгенерировали на этапе обучения модели. В нем содержатся пары значений – опорное и предсказанное моделью.





В окне настройки диаграммы укажите в качестве оси X поле TARGET (опорные данные), в качестве значений оси Y поле PREDICTED (прогноз).

Посмотрим на диаграмму. В принципе, видна линейная зависимость. Понятно, что регрессионая модель не прогнозирует идеально. Для низких значений биомассы процент ошибок выше. Есть выбросы, которые модель прогнозирует с ошибками. Но в среднем она все же предсказывает с коэффициентом R2 в районе 0.8, т.е. с ошибкой примерно 20%.



ИТОГИ УПРАЖНЕНИЯ

Вы познакомились с инструментами построения регрессионной модели в приложении ArcGIS Pro. Такие модели можно использовать практически в в любых географических задачах. Главное, чтобы в реальности существовала какая-то статистическая зависимость между опорными данными и входными данными (в данном случае растрами), с помощью которых вы будете делать прогнозную модель. Это может быть прогноз концентрации химических элементов в почве, прогноз стоимости недвижимости и так далее.