Санкт-Петербургский государственный университет

Кафедра информационных систем

ЖИВУЛИН Евгений Алексеевич

Магистерская диссертация

МОНИТОРИНГ ТЕРРИТОРИИ НА ОСНОВЕ СПУТНИКОВЫХ СНИМКОВ

Направление 01.04.02 «Прикладная математика и информатика»

Образовательная программа ВМ.5517

«Методы прикладной математики и информатики в задачах управления»

Профиль «Вычислительные методы и информационные технологии в современном естествознании»

Научный руководитель,
кандидат техн. наук,
доцент Гришкин В.М.

Рецензент,
Институт компьютерных технологий
Государственного университета
аэрокосмического приборостроения,
кандидат техн. наук,
профессор Космачев В.М.

Санкт-Петербург

2021

Введение

Большая проблема в ряде регионов России, которую давно пытаются решить - опустынивание земли. Например, ранее такой регион, как Калмыкия, представлял собой степь, покрытую растительностью, на которой жили скотоводы, которые вели кочевой образ жизни, что позволяло землям восстанавливаться после скота. Затем стали появляться колхозы, поголовье скота существенно выросло, что привело к большой нагрузке на почву и, как следствие, появлению антропогенной пустыни.

На данный момент в этой республике ведется активная борьба против распространения пустыни, за сохранение и увеличения площади зеленой зоны. Для эффективной борьбы необходимо знать ситуацию в текущий момент времени, а также иметь возможность отслеживать ее динамику.

Поэтому разработка методов и программных средств, позволяющей видеть и оценивать текущее состояние земли является достаточно актуальной задачей.

Наиболее оптимальным способом для этого является наблюдение за территорией с помощью спутниковых снимков. Данный метод позволяет охватить всю территорию, при этом предоставляя достаточную для таких масштабов точность. Кроме того, он не требует больших финансовых затрат и позволяет собирать и обрабатывать информацию в течение всего периода времени, когда земля республики не покрыта снегом.

Благодаря этому методу наблюдения есть возможность отслеживать динамику распространения песка, изменения площади территорий, занятых растениями.

Кроме того, при дальнейшем развитии этого метода в направлении детекции почвы и песка, как двух отдельных классов, будет информация о тех областях, на которых более выгодно сажать растения и которые следует оберегать в первую очередь, т.к. иначе наступление пустыни будет уменьшать количество почвы.

Этот способ наблюдения также может быть коммерчески выгодным. Фермеры заинтересованы в наблюдении за своими сельскохозяйственными полями. Соответственно, при анализе спутниковых снимков высокого разрешения или снимков с дронов появляется возможность наблюдать за всеми полями без необходимости их посещения.

Источником данных для наблюдения за поверхностью является набор спутниковых снимков, полученные с сервиса Sentinel Hub. Причем, используются не только снимки, демонстрирующие видимое глазу излучение, но и другие. Кроме того, используется маска, классифицирующая поверхность земли на различные типы, полученная посредством обработки данных самим сервисом Sentinel Hub.

Загрузка одного снимка доступна с помощью использования утилиты curl. Для автоматической загрузки множества снимков, покрывающих территорию республики, был написан скрипт на python, формирующий запросы на все требуемые изображения и адресующий их curl.

Для расчета площади поверхности, занятой растениями, производится подсчет количества пикселей, определенных как растительность и принадлежащих территории республики. После чего производится умножение площади одного пикселя на их количество.

Для детекции почвы и песка, как отдельных классов, применяются различные методы бинарной классификации. Перед тем, как производить тренировку классификатора на основе данных снимков, снимки подвергаются предобработке, что существенно увеличивает точность разделения класса почвы от класса песка.

В ходе данной работы были разработаны программы для загрузки спутниковых снимков, для подсчета площади поверхности Калмыкии, занятой интересующим классом, для детекции песка и почвы. Также были рассмотрены разные методы детекции земли и почвы и их точность. Результаты данной

работы могут быть использованы для дальнейшего развития работ в этой и смежных областях.

Постановка задачи

В данной работе поставлена задача создания инструмента для мониторинга поверхности достаточно больших регионов России на примере республики Калмыкия. Для решения этой задачи потребуется:

- провести анализ возможностей существующего интерфейса обмена с сервисом, предоставляющим доступ к спутниковым данным Sentinel Hub для загрузки необходимых данных;
- разработать и реализовать метод для автоматической загрузки набора снимков с указанной области в указанный диапазон дат;
- разработать и реализовать метод для коррекции областей облаков на изображениях поверхности;
- осуществить классификацию песка и почвы, как разных типов поверхности;
- произвести расчет территории, занятой растительностью

Обзор литературы

Для получения информации о поверхности исследуемой территории производится детекция отраженного солнечного излучения. Используются различные диапазоны электромагнитного излучения: видимый (0.38-0.72 мкм) диапазон, а также ближний (0.72-1.3 мкм), средний (1.3-3.0 мкм) и дальний (7.0-15.0 мкм) инфракрасный спектр. В связи с влиянием атмосферы, использование ультрафиолетового излучения и спектра в диапазоне 3.0-7.0 мкм для получения данных затруднено.

На данный момент анализ спутниковых снимков с целью получения информации о поверхности Земли получил широкое распространение. Например, этот способ мониторинга территории используется для получения информации об окружающей среде посредством наблюдения за состоянием растительности и водоемов. Также сейчас актуальны задачи мониторинга фермерских полей, поиск выгоревших участков.

Существует два принципиальных подхода к детекции спутниковых снимков: анализ каждого пикселя отдельно, исходя из значений цветов этого пикселя, и анализ с использованием имеющегося контекста, например информации о взаимном расположении пикселей.

Для проведения детекции на изображениях интересующих классов активно используются методы обучения с учителем и без учителя. Известным способ детекции растительного покрова является применение кластеризации методом k-средних [1] к изображению, полученному после замены значений каждого пикселя по формуле нахождения вегетационного индекса NDVI [2] . Кроме того, применяются и более сложные методы, такие как глубинные нейронные сети, сверточные нейронные сети и другие, используемые в распознавании образов.

Глава 1. Загрузка и обработка спутниковых снимков

Для проведения исследования в данной работе использовались спутниковые снимки сервиса Sentinel Hub. Данный сервис предоставляет возможность загрузить последний снимок указанной области в указанном диапазоне дат. Для просмотра всей территории республики Калмыкии требуется загрузить большое число снимков.

Так как для мониторинга поверхности необходимо производить загрузку многочисленных изображений, была разработана программа, производящая это автоматически.

На спутниковых снимков присутствует большое количество облаков и их теней, закрывающих обзор. Для просмотра участков территории, которые на одном снимке были закрыты облаками и их тенями, было решено использовать данные с других снимков, которые не имеют облаков в данных участках. Таким образом, производилось объединение информации со всех снимков, сделанных в один месяц, с целью получения одного снимка с минимально возможным количеством облаков.

1.1. Загрузка спутниковых снимков

Для исследования используются данные со спутников, предоставляемые сервисом Sentinel Hub. Снимки поверхности Земли производятся с периодичностью в 5 дней. Съемка поверхности со спутников ведется в разных диапазонах с разрешением не лучше 10м * 10м в одном пикселе изображения. Разрешения разных каналов различны, далее представлена таблица с каналами, которые загружаются с сервиса в рамках данной работы.

Канал	Длина волны	Разрешение
Blue	492.4 nm	10 m

Green	559.8 nm	10 m
Red	664.6 nm	10 m
Vegetation red edge	704.1 nm	20 m
Vegetation red edge	740.5 nm	20 m
Vegetation red edge	782.8 nm	20 m
NIR	832.8 nm	10 m
Narrow NIR	864.7 nm	20 m
Water vapour	945.1 nm	60 m
SWIR	1613.7 nm	20 m
Snow probability	-	20 m
Cloud probability	-	20 m
SCL (scene classifica data)	-	20 m

Таблица 1. Волновой диапазон и максимальное разрешение используемых каналов съемки со спутников.

Первые три канала представляют собой разложение отраженного света, который способен воспринять наш глаз, на три составляющие. Следующие каналы представляют собой излучение, относящееся к инфракрасному диапазону. Затем еще два канала в ближнем инфракрасном диапазоне, испарения воды и коротковолновое инфракрасное излучение. Три последних канала - результаты обработки данных самим сервисом Sentinel Hub, а именно: вероятность того, что тот или иной пиксель является частью заснеженной

области или облака, а также маска изображения, которая каждому пикселю присваивает значение от 0 до 11. Среди классов, к которым принадлежит конкретный пиксель в маске есть следующие: растительность, голая почва, вода, облака, тени облаков, снег.

Данные, предоставляемые с каналов, являющиеся отражением электромагнитных волн находятся в диапазоне от 0 до 0.4, каналы вероятности снега и облаков от 0 до 100.

Для загрузки снимков можно воспользоваться утилитой curl, позволяющей формировать запросы на указанный утилите сайт через команду в командной строке. Сам запрос на сервис Sentinel Hub представляет собой код на javascript, в котором указывается диапазон дат(сервис предоставляет последний по времени снимок из указанного диапазона), координаты интересуемой области(координаты левого нижнего и правого верхнего углов интересуемой прямоугольной области), высота и ширина получаемого изображения в пикселях, список каналов, с которых запрашиваются данные и вид, в котором они будут загружены.

Координаты запрашиваемой области необходимо указывать в UTM. Это система координат, которая делит северное и южное полушария на 60 зон. Точность данной системы выше, чем обычных ширины и долготы. Перед тем, как загружать данные, есть возможность их предварительно обработать, произведя над ними арифметические операции, после чего загрузить их в PNG изображение, определяя, какие данные каким цветом(красным, зеленым или синим) будут обозначены.

Загрузка производится с 13 разных каналов, соответственно, учитывая то, что в одном PNG изображении может содержаться только 3 цвета, необходимо загружать по 5 изображений на один участок местности за указанный период времени.

Размер запрашиваемой области, умещающийся на изображении, выбран 10км * 10км исходя из того, что разрешение на большинстве каналов 20м, следовательно 20м * 512 пикселей ~ 10км.

Получение данных, необходимых для создания большой выборки и исследования территорий, занимающих огромные площади, потребует и большое количество загрузок. Поэтому для скачивания данных использовался скрипт на Python, который формировал запрос, загружающий требуемые данные из указанного места в указанный период дат. После формирования запроса, скрипт вводит его в командную строку операционной системы, благодаря чему происходит загрузка.

Для автоматизации загрузки, в скрипт передается массив с координатами, сохраненными в json файле, указывающими области на карте, с которых требуется загрузить снимки. Данные координаты расчитаны скриптом на Python, разбивающим указанную область на набор квадратов, и сохраняющих их в json файле. Разбиение интересующей области на участки было произведено с предварительным переводом граничных координат всей области в систему UTM. Это позволило увеличить точность разбиения. В противном случае, при переводе в UTM каждого участка отдельно, возникали некоторые погрешности у соседних участков и точки, которые должны совпадать, немного отличались, что заметно на итоговых изображениях.

Кроме того, в скрипте на Python указаны также запрашиваемые каналы и вид, в котором они будут загружены для каждого из 5 изображений на один участок загружаемой территории. В итоге для каждого цвета при сохранении данных в PNG изображении значение должно быть в диапазоне от 0 до 1, после чего оно будет переведено в используемый для RGB диапазон от 0 до 255. Таким образом, производится загрузка по всем каналам, по всей указанной области автоматически.

Кроме того, необходимо предоставлять сервису ключ, который обновляется каждый час и может быть получен в результате другого curl запроса на сервис,

поэтому в программе предусмотрена проверка размера загружаемых изображений, и если он слишком мал, это свидетельствует о том, что сервис не предоставил данные, следовательно, необходимо запросить новый ключ, что и производится в скрипте.

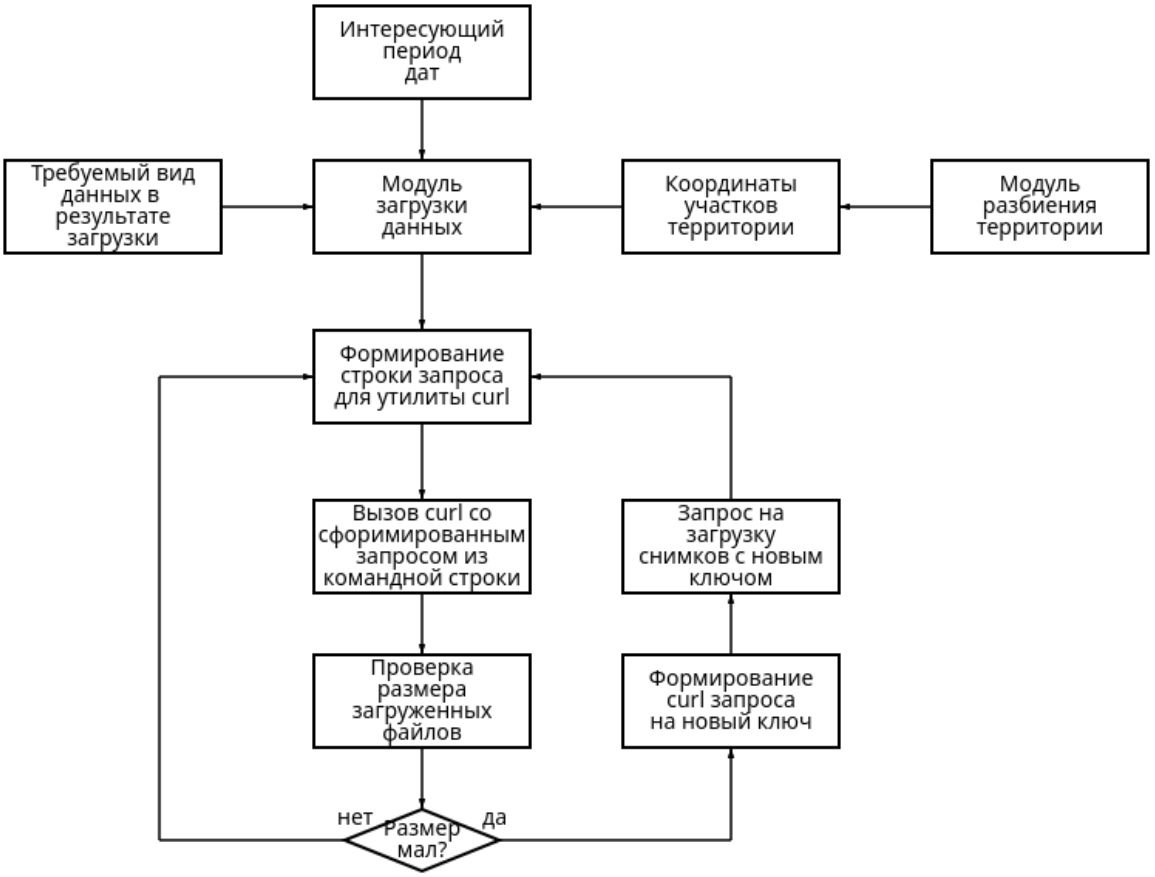

Рис. 1. Схема алгоритма загрузки данных с сервиса Sentinel Hub.

На тот случай, если соседние снимки будут немного смещены друг относительно друга, нужно производить загрузку с небольшим нахлестом соседних снимков друг на друга. Это обеспечит запас на тот случай, если снимки будут смещены в разные стороны друг от друга, чтобы не потерять часть информации. Кроме того, такой нахлест позволяет реализовать метод опорных точек, который заключается в совмещение набора точек на обоих изображениях так, чтобы они совпали друг с другом.

В рамках данной работы метод опорных точек был реализован следующим образом: края двух соседних изображений “накладывались” друг на друга с некоторым сдвигом, после чего рассчитывалась их корреляция. Рассмотрим два снимка, у которых общими являются правый край одного и левый другого снимка. В таком случае производится расчет функции корреляции при пересечении этих краев только на 1 пиксель вдоль горизонтали, затем на 2 и так далее до заданного предела. Кроме того, также производится сдвиг вдоль вертикальной оси, а именно, при одном общем пикселе вдоль горизонтали сдвиг по вертикали изначально равен нулю, на следующем шаге сдвиг уже 1, то есть верхний пиксель левого изображения и нижний правого не учитываются в расчете функции корреляции, затем производится сдвиг на 2 пикселя и новый расчет корреляции и так далее.

Таким образом рассчитывается двумерная дискретная корреляционная функция пересекающихся частей соседних снимков. После чего максимальное значение определяется соответствующим истинному сдвигу, что позволяет совмещать изображения, не теряя информацию со снимков, но и не захватывая лишних пикселей, искажающих общую информацию о всей исследуемой территории.

Аналогичным образом производится сопоставление снимков, у которых общим являются верхний край одного и нижний другого снимка.

Корреляционная функция по всем пересекающимся точкам снимков слишком затратна по времени, поэтому было решено использовать для сравнения только каждый n-ый пиксель, где n принято равным 20. Это дало достоверные результаты при совмещении снимков, края которых имеют неравномерное распределение по цвету, например зеленая зона чередуется с песчаной. В случае относительно однородного покрытия, например только песчаной области, данное решение, ускоренное по времени, дает неверный результат.

Для решения возникшей проблемы была увеличена контрастность краев изображений. Производился поиск медианной яркости по каждому из трех цветов системы RGB, после чего производилась следующая обработка пикселей: все 3 составляющие его цвета сравнивались с медианными значениями, если они были меньше, то они домножались на уменьшающий коэффициент(<1), в противном случае на увеличивающий(>1). Такое решение позволило резко увеличить контрастность, что дало возможность сопоставлять даже достаточно однотонные изображения при небольших затратах на расчет корреляционной функции.

Обработка производилась только с краями изображений, так как в противном случае, часть изображения, не участвующая в совмещении, может существенно изменить медианное значение цветов пикселей, что очевидно может привести к неверным результатам.

Пример работы по совмещению соседних снимков с увеличенной контрастностью краев приведен на рисунке 2. Несмотря на резкое изменение резкости, можно заметить, что дороги с обеих сторон от контрастной части также переходят в дороги в этой контрастной части.

На рисунке 3 приведен результат работы этого алгоритма без демонстрации контрастной части.

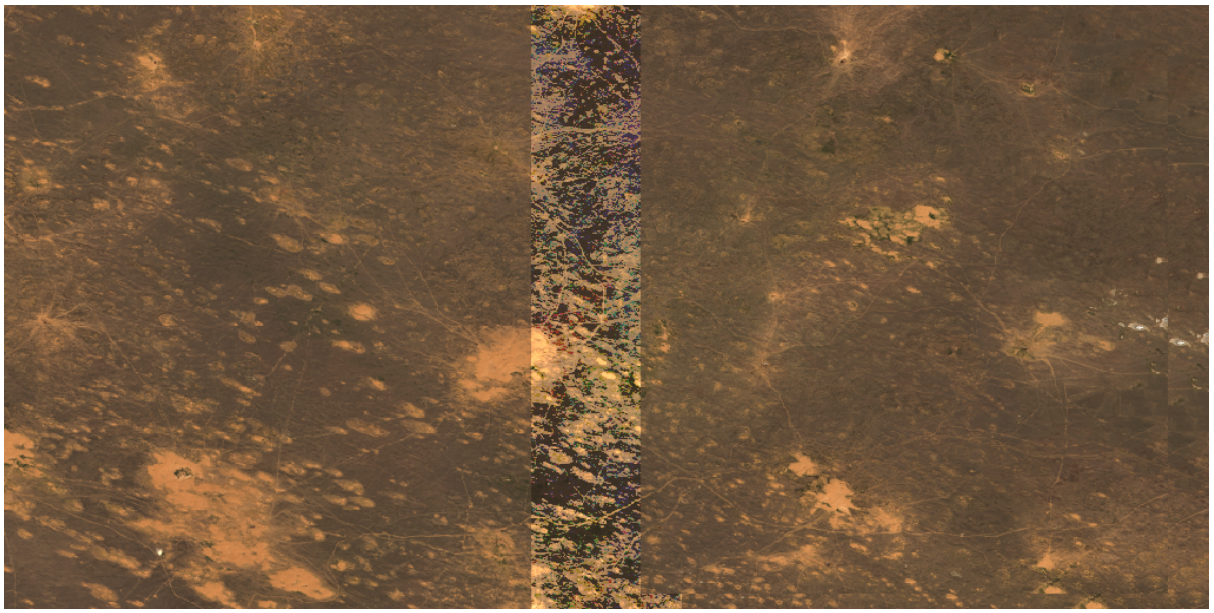

Рис. 2. Результат работы алгоритма по совмещению двух соседних снимков с демонстрацией увеличения контрастности краев.

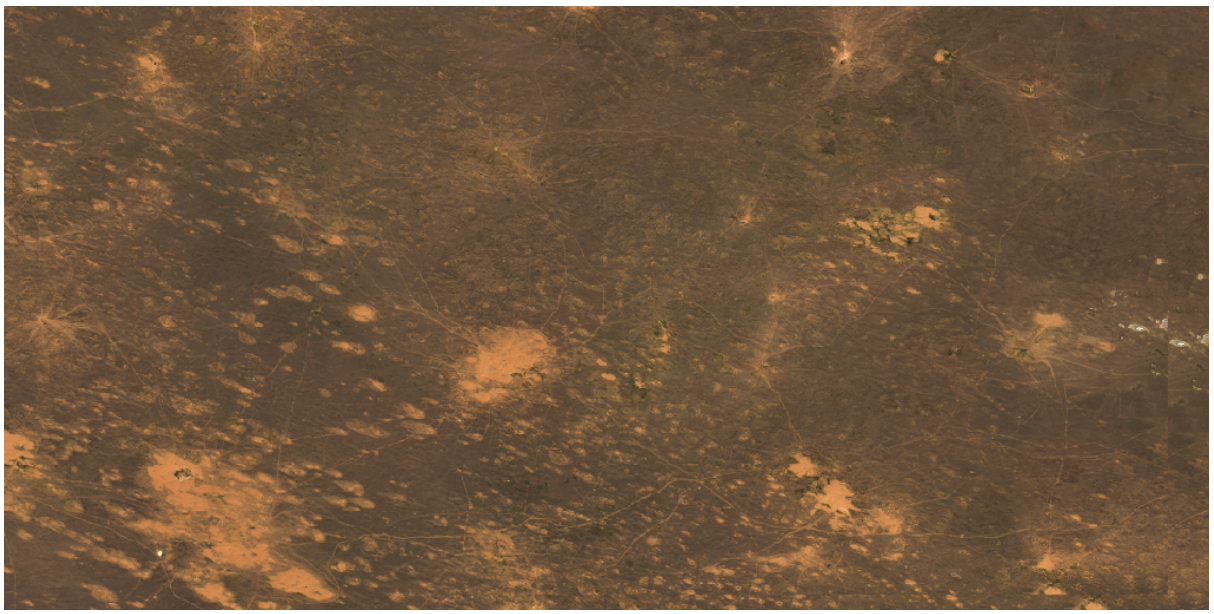

Рис. 3. Результат работы алгоритма по совмещению двух соседних снимков.

1.2. Обработка спутниковых снимков с целью коррекции областей облаков

Для наблюдения за поверхностью Земли необходима полная видимость, которая не всегда возможна по причине наличия облаков, закрывающих часть исследуемой области. Для решения этой проблемы используется несколько снимков одного и того же сектора в разные дни в течение месяца.

Для каждой позиции рассматриваются пиксели с нескольких изображений, среди которых отбрасываются те, которые определены как облако или тень. Если кроме облаков и теней ничего нет - приоритет достается теням. Если среди оставшихся есть пиксель, принадлежащий к интересующему классу, то выбирается он. Среди пикселей одного класса всегда выбирается наиболее яркий, но не превышающий верхнего порога. Это позволяет бороться с тенями и облаками, которые

не были детектированы. Кроме того, это снижает резкость переходов по яркости на одном изображении.

Таким образом, получается изображение, содержащее все объекты интересующего класса, которые были замечены в ходе наблюдения в течение месяца, демонстрирующее всю территорию, которую можно было увидеть в течение месяца наблюдений.

На рисунке 4 представлено изображение, полученное в результате обработки 6 изображений просматриваемого участка поверхности, снятых в течение одного месяца.

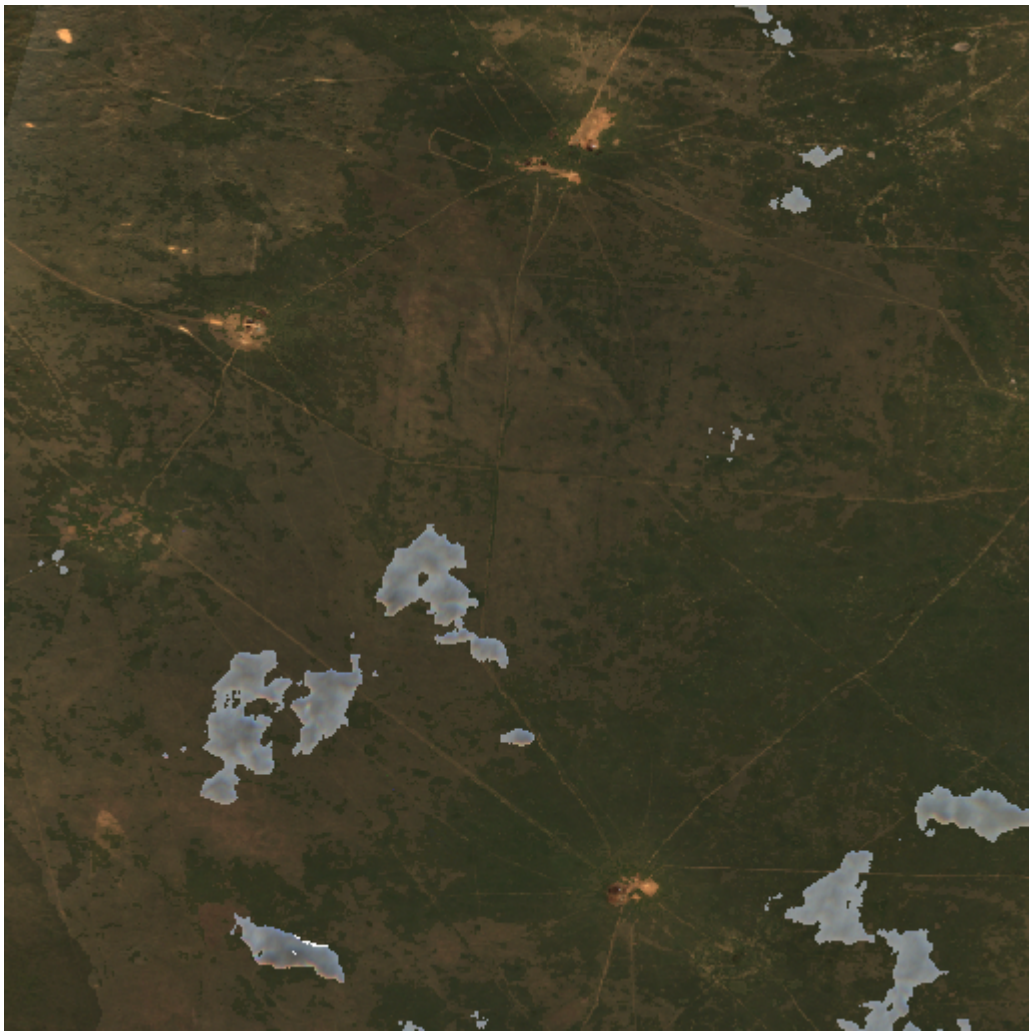

Рис. 4. Изображение, полученное в результате обработки 6 снимков одной территории, снятых за 1 месяц.

Как можно заметить, на данном изображении часть облаков осталась, хоть они и не являются абсолютно непрозрачными и через них даже можно что-то рассмотреть. Они остались в связи с тем, что не всегда конкретная часть территории безоблачна хотя бы на одном из 6 снимков.

Тем не менее, этот подход решает проблему закрытия облаками и дает хорошую видимость исследуемой территории, позволяя наблюдать общую картину. Закрыт совсем небольшой процент территории и от большей части облаков таким образом удается избавиться.

Глава 2. Детектирование почвы и песка

Задача детекции растительности посредством обработки спутниковых снимков уже давно имеет простое и эффективное решение, что позволяет наблюдать за областями с растительностью в республике Калмыкия. Однако, не все время на территории республики есть растительность, что уменьшает доступный период для наблюдения за ситуацией на территории республики.

Кроме того, учитывая проблему, связанную с распространением песка, становится актуальной задача классификации песка и плодородной почвы.

Сервис Sentinel Hub предоставляет возможность загрузить маску, на которой каждый пиксель относится к тому или иному классу, определенному алгоритмами детектирования сервиса. Имеются следующие классы: темная область, тени от облаков, растительность, голая земля, вода, низкая вероятность облаков, средняя вероятность облаков, высокая вероятность облаков, перистые облака, снег/лед.

2.1. Предобработка данных для обучения модели детекции почвы и песка

Для детектирования плодородной почвы и песка были использованы различные методы классификации библиотеки Python - scikit-learn. Посредством этих методов была произведена классификация пикселей спутниковых снимков, определенных сервисом Sentinel Hub, как голая земля, по двум типам: песку и плодородной почве.

Перед применением методов классификации, данные, полученные с разных каналов (разных электромагнитных диапазонов), были обработаны для получения более точного результата. За основу был взят метод детекции растительности, заключающийся в применении вегетационного индекса NDVI и последующей кластеризации полученных данных. Индекс NDVI представляет собой формулу:

$$NDVI = \frac{NIR - RED}{NIR + RED} \text{ [2]} \quad (1)$$

Эта формула, примененная к каждому пикселю, создает новое изображение, яркость пикселя на котором будет зависеть от того, является ли конкретный пиксель частью растительной области, или нет.

Для предобработки использовались выражения аналогичные формуле (1), но с разными каналами данных. То есть, для формирования модели детекции тем или иным методом использовались выражения вида:

$$\frac{channelX - channelY}{channelX + channelY} \quad (2)$$

Где channelX и channelY два произвольных канала данных. Таким образом, подбирая наиболее результативные пары каналов, была увеличена точность детекции.

Для обучения были использованы пиксели со спутниковых снимков, которые являются или плодородной почвой, или песком. На областях, являющихся плодородной почвой, летом зеленая зона. Соответственно, осенью территорию, которая была зеленой летом, можно считать плодородной почвой. Таким образом, в результате наложения маски, полученной для летнего периода, на осенний, были получены тренировочные данные для одного из двух классов. Тренировочными данными второго класса - песка, являлись пиксели, расположенные в пустынной части республики.

Из изображений с плодородной почвой и песком загружаются пиксели, являющиеся либо плодородной почвой, либо песком. После этого образуется вектор, каждый элемент которого содержит в себе все 12 значений с разных каналов для одной точки поверхности. То есть, в одном элементе этого вектора содержится информация со всех четырех изображений, загруженных для конкретной области и содержащих в себе информацию с 12 каналов. И из таких

элементов состоит весь вектор. Кроме того, формируется вектор, который размечает каждую точку по принадлежности к классу плодородной почвы или классу песка.

В ходе работе были перебраны все пары, которые можно составить из имеющихся каналов. Каждая полученная по шаблону формула применялась к имеющимся данным, вследствие чего формировался вектор, содержащий полученные данные для каждой точки. После этого этот вектор, а также вектор, размечающий точки по принадлежности к классу, передавались в функцию, производящую обучение модели детектирования поверхности.

Таким образом, для каждого метода был получен список с парами каналов, для каждой из которых указывалась точность детекции, полученная в результате обучения модели на основе данных, рассчитанных по указанной выше формуле с подстановкой этих двух каналов. Обучение производилось на датасете, который содержал пиксели разных классов в соотношении 1:1. За точность принималась среднее между точностью детектирования плодородной почвы (процент пикселей плодородной почвы, определенных частью своего класса) и точностью детектирования песка.

Данный список был отсортирован по убыванию точности, после чего производилось обучение модели на основе вектора, содержащего данные, рассчитанные только на основе двух самых точных пар. То есть один компонент вектора имел размерность 2. Если точность становилась выше, чем при обучении на основе только одной самой точной пары, то вторая пара добавлялась в “основной состав”, и дальше производилось обучение на основе уже трех самых точных пар, с целью определить, рост или падение точности, чтобы добавить или отбросить еще одну пару соответственно. В том случае, если точность с присоединением пары снижалась, она отбрасывалась.

В результате такого отбора формировался набор пар каналов, который впоследствии применялся для обучения модели детекции типа поверхности. Такой подход существенно увеличивает точность детекции по сравнению с

использованием только одной пары или относительно использования всех каналов данных без предварительной обработки.

2.2. Метод опорных векторов

Метод опорных векторов [3] разделяет гиперплоскостью пространство признаков, в котором находятся объекты выборки, применяемой для обучения модели. В случае двух линейно разделимых классов гиперплоскость проводится так, чтобы расстояние от ближайших к ней точек каждого класса было максимально.

Зададим плоскость, разделяющую объекты разных классов, следующим уравнением:

$$(\vec{n}, \vec{x}) - b = 0 \tag{3}$$

Где $\vec{n}$ - вектор нормали к плоскости, $\vec{x}$ - вектор идущий из начала координат в точку плоскости, b - расстояние от плоскости до начала координат.

В том случае, если вектор $\vec{x}$ будет представлять собой вектор, соединяющий начало координат и элемент обучающей выборки, в зависимости от того, с какой стороны от гиперплоскости лежит этот элемент, значение выражения $(\vec{n}, \vec{x}) - b$ будет положительным или отрицательным.

Если линейно разделить два класса не получится, часть объектов будет находится относительно гиперплоскости со стороны другого класса. Для этого потребуется ввести понятие отступа от плоскости, которое будет равно значению выражения $(\vec{n}, \vec{x}) - b$ по модулю, но при этом будет иметь положительное значение, если класс объекта определен верно, и отрицательное, если неверно.

$$M(\vec{x_i}) = f((\vec{n}, \vec{x_i}) - b) \tag{4}$$

После этого поступаем с усложненной задачей аналогично задаче с линейно разделяемыми классами, то есть стремимся построить плоскость так, чтобы достичь максимального значения этого выражения.

2.3. Дерево решений

Дерево решений [4] – это математическая модель в виде графа, которая отображает точки принятия решений, предшествующие им события и последствия. Этот граф состоит из следующих элементов:

- узлы (вершины) - точки принятия решений, которые могут иметь несколько ветвей. На этом элементе происходит выбор ветви в зависимости от значения параметров конкретного объекта выборки.
- конечные узлы (листья) - представляют результат, то есть класс конкретного объекта в задаче классификации.
- ребра (ветви) - элементы, соединяющие узлы и описывающие вероятность события по данному сценарию.

Таким образом, в каждом узле есть набор условий, который определяет, на каком узле будет приниматься следующее решение. Такой переход между узлами будет происходить до тех пор, пока не будет достигнут конечный узел, сообщающий класс объекта.

2.4. Случайный лес

Случайный лес [5] - это множество деревьев решений, в результате работы которых будет определен тот класс объекта, который вернуло большинство деревьев. Деревья случайного леса строятся независимо по описанному далее алгоритму.

Из элементов обучающего множества выбирается подмножество, на основании которого строится дерево решений.

При построении каждого расщепления в узле дерева рассматривается случайное подмножество всех признаков объекта. Причем, для каждого узла свой случайный набор признаков.

Исходя из заранее заданного критерия, происходит выбор наилучших признаков и расщепление по ним. Как правило, дерево строится до тех пор, пока в каждом листе не будут объекты только одного класса.

2.5. Стохастический градиентный спуск

Стохастический градиентный спуск [6] - это итерационный метод для оптимизации целевой функции с подходящими свойствами гладкости. Его можно считать стохастической аппроксимацией оптимизации методом градиентного спуска, потому что он заменяет реальный градиент его оценкой, которая вычисляется из случайно выбранного подмножества данных.

Идея метода заключается в движении в направлении наискорейшего уменьшения/роста значения функции, которое задается градиентом. В связи с тем, что требуется найти конфигурацию, для которой количество определенных объектов класса будет максимально, будем рассматривать направление наискорейшего роста.

Рассмотрим обычный градиентный спуск. Пусть целевая функция имеет вид:

$$F(\overrightarrow{x}) : X \to R \tag{5}$$

$$F(\overrightarrow{x}) \to max(\overrightarrow{x}), \overrightarrow{x} \in X \tag{6}$$

Тогда, следующая точка, в которой окажется метод, рассчитывается по формуле:

$$x^{[j+1]} = x^{[j]} + \lambda^{[j]} \nabla F(\overrightarrow{x}^{[j]}) \tag{7}$$

Таким образом, происходит движение в направлении максимального роста. Аналогично и в методе стохастического градиентного спуска, но с той

разницей, что для градиента выбирается только несколько признаков каждый раз, а не все.

2.6. Квадратичный дискриминантный анализ

Квадратичный дискриминантный анализ [7] - метод для поиска квадратичной комбинации признаков с целью описать или разделить два или более класса.

Пусть тренировочные данные для каждого элемента выборки заданы в виде вектора признаков $\overrightarrow{x}$ и класса y, к которому принадлежит конкретный объект. Требуется решить задачу поиска предсказания класса y для объекта, заданного вектором признаков $\overrightarrow{x}$.

Метод квадратичного дискриминантного анализа предназначен для нормального распределения условных плотностей вероятностей $p(\overrightarrow{x}, y = 0)$ и $p(\overrightarrow{x}, y = 1)$ со средним и параметрами ковариации $(\overrightarrow{\mu_0}, \Sigma_0)$ и $(\overrightarrow{\mu_1}, \Sigma_1)$ соответственно. Отсюда предсказание на основе байесова оптимального решения определяет, что конкретный объект является представителем второго класса, если отношение правдоподобия больше порогового значения T:

$$(\overrightarrow{x} - \overrightarrow{\mu_0})^T \Sigma_0^{-1} (\overrightarrow{x} - \overrightarrow{\mu_0}) + ln|\Sigma_0| - (\overrightarrow{x} - \overrightarrow{\mu_1})^T \Sigma_1^{-1} (\overrightarrow{x} - \overrightarrow{\mu_1}) - ln|\Sigma_1| > T \quad (8)$$

Соответственно, в противном случае объект принадлежит первому классу.

2.7. Экспериментальные результаты

Ниже представлены изображения, полученные в результате применения реализации метода коррекции областей облаков за август и за октябрь - рисунки 5 и 6 соответственно, а также маска, полученная в результате применения классификатора песка и почвы на основе модели обучения случайный лес к изображению за октябрь - рисунок 7.

Рис. 5. Изображение, полученное в результате улучшения видимости поверхности на снимках за август.

Рис. 6. Изображение, полученное в результате улучшения видимости поверхности на снимках за октябрь.

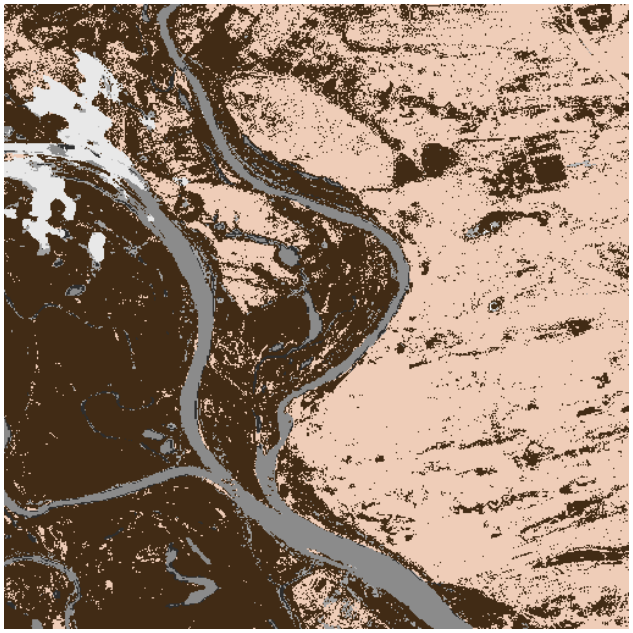

Рис. 7. Маска, полученная в результате применения классификатора песка и почвы на основе модели случайный лес к изображению, полученному за октябрь.

Далее представлена таблица, демонстрирующая, какая часть объектов плодородной почвы была определена принадлежащей этому классу.

Аналогичная информация представлена и для песка. Кроме того, в таблице указана F-score [8] (F-мера). Обучение моделей проводилось при соотношения числа объектов классов 1:1.

F-мера представляет собой общепринятую оценку качества работы алгоритма, являющуюся более точной, чем показатель точности метода, основанный на отношений количества правильных предсказаний к количеству всех предсказаний модели. Для его основы потребуется определить такие величины, как true positive, true negative, false positive, false negative.

Пусть y - истинная метка класса на конкретном объекте, а $\hat{y}$ - результат предсказания модели, тогда в следующей таблице (таблица 2) элементы, принадлежащие каждой ячейки будут пересечениями множеств истинных и предсказанных меток, где 1 - метка одного класса, условно положительный ответ, в данном случае плодородной почвы, а 0 - другого, условно отрицательный ответ, в данном случае песка.

	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive(TP)	False Positive(FP)
$\hat{y} = 0$	False Negative(FN)	True Negative(TN)

Таблица 2. Пересечения множеств положительных и отрицательных меток по условию и предсказанных классификатором.

Введем следующие величины: precision (точность) - доля объектов, определенных классификатором положительными и при этом действительно являющимися положительными, recall (полнота) - доля всех положительных объектов, которые детектировал классификатор.

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$Fscore = 2\frac{precision * recall}{precision + recall} \tag{11}$$

Метод	Часть объектов почвы, определенная моделью принадлежащей к этому классу	Часть объектов песка, определенная моделью принадлежащей к этому классу	F-score
Квадратичный дискриминантный анализ	83.40%	93.32%	87.75%
Метод стохастического градиентного спуска	86.87%	94.57%	90.35%
Дерево решений	91.40%	91.45%	91.42%
Случайный лес	94.58%	95.84%	95.18%
Метод опорных векторов	92.96%	96.32%	94.55%

Таблица 3. Точность предсказания различных методов классификации.

В результате обучения модели детекции к классу плодородной почвы классификатор относил не только ту почву, которая летом представляла собой

зеленую зону, но и часть той, которая ей не являлась. Модель достаточно хорошо детектируют песок и плодородную почву, однако в таком промежуточном случае, как почва, не являющаяся песком, но и не заполненная растительностью летом, она относится к плодородной.

Почва, являющаяся промежуточным типом между плодородной почвой и песком, нередко расположена рядом или внутри территории, озелененной летом, рядом с водоемами.Соответственно, от плодородной почвы она отличается меньше, чем от песка.

Таким образом, модель классифицирует тип поверхности на песок и почву, однако промежуточные случаи требуют более детального изучения. Учитывая необходимость борьбы именно с песком, модель можно считать успешной. Кроме того, это может позволить искать области, которые будут более отзывчивы к озеленению, чем песок, и которые особенно важно беречь, так как их состояние относительно критерия плодородности со временем может ухудшиться, в отличие от состояния песка.

Глава 3. Расчет площади территории, занятой растительностью

Прежде, чем приступать к расчету площади, занятой тем или иным типом поверхности, необходимо получить данные о всех пикселях, принадлежащих к интересующему классу. Сервис Sentinel Hub предоставляет возможность загрузить маску поверхности для выбранной территории, которая подразделяет элементы изображения на следующие типы: отсутствие данных, поврежденные данные, темная область, тень от облака, растительность, голая земля, вода, облака, лед/снег.

Затем производится процесс коррекции областей облаков посредством замены участков с облаками или тенями частями других снимков с того же места, но без облаков и теней.

Кроме того, следует учитывать границы Калмыкии. Вся загружаемая территория больше самой Калмыкии, и для того, чтобы это учесть, была создана маска территории республики, которая учитывается при подсчете количества пикселей конкретного класса объекта.

Учитывая все приведенное выше, производится подсчет всех пикселей определенного класса на территории республики за выбранный месяц, после чего результат переводится в единицы площади из расчета того, что один пиксель имеет длину стороны 20 м.

Ниже представлены результаты подсчета площади поверхности республики Калмыкии, занятой растительностью.

Дата	Площадь км^2
2018 август	881.68
2018 октябрь	8604.33
2019 август	1090.16
2019 октябрь	22001.32

2020 август	341.12
2020 октябрь	73.59

Таблица 4. Площадь, занимаемая растительностью в Калмыкии в зависимости от месяца.

Как можно заметить, при площади, равной 76100 км^2, республика имеет совсем немного растений и большую часть ее поверхности занимает земля, а именно - пустыня.

Есть сильное различие между площадью растительности летом и осенью в 2018 и 2019 годах. Это произошло в связи с тем, что большую часть сельскохозяйственных культур, выращиваемых на территории республики, представляют озимые сорта растений, которые высевают в конце лета или осенью.

Кроме того, если обратить внимание на одни и те же месяцы 2018 и 2019 годов, будет виден рост площади, занимаемой растениями. На территории республики проводятся мероприятия для обеспечения гарантии урожая сельскохозяйственных культур, например ввод в эксплуатацию и реконструкция земель регулярного и инициативного орошения. Производятся фитомелиоративные предприятия по закреплению открытых песков. Также организованы лесовосстановительные мероприятия в рамках федерального проекта “Сохранение лесов”.

Таким образом, удается снизить скорость распространения песка и увеличить территорию, занятую растениями.

В 2020 году произошло резкое падение площади исследуемой территории. Причиной этого является засуха. При отсутствии осадков суховей еще больше иссушал землю, в которой практически не осталось влаги. Ситуация усугубилась бесснежной зимой, в межсезонье дождей практически не было. Из-за отсутствия зимних морозов существенно увеличилась популяция саранчи, что еще сильнее ухудшило ситуацию.

Выводы

В ходе данной работы был проведен анализ возможностей существующего интерфейса обмена с сервисом, предоставляющим доступ к спутниковым данным Sentinel Hub для загрузки необходимых данных.

Разработан и реализован метод для автоматической загрузки спутниковых снимков с указанной области в указанный диапазон дат.

Разработан и реализован метод для коррекции областей облаков на изображениях поверхности.

Осуществлена классификация песка и почвы, как разных типов поверхности.

Произведен расчет территории, занятой растительностью.

Заключение

Реализованный метод для автоматической загрузки спутниковых снимков с указанной области в указанный диапазон дат позволяет иметь доступ к данным с любой требуемой территории. Также реализация данного метода позволяет настраивать набор интересующих каналов информации и вид, в котором они будут получены, что делает ее универсальной для любых задач, требующих загрузки данных сервиса Sentinel Hub.

Метод коррекции облаков позволяет увеличивать незакрытую облаками и их тенями видимую часть на изображении за счет создания из 6 снимков, сделанных за 1 месяц, одного изображения поверхности, с количеством облаков, как правило, меньше, чем на любом из этих 6 снимков. В крайнем случае, количество облаков будет равно их количество на наименее облачном снимке.

При реализации данного метода учитывалось то, что на одном снимке растительность может присутствовать, а на другом снимке той же территории за тот же месяц отсутствовать по причине плохой видимости в области ее нахождения и при этом отсутствия детекции этой части, как облака или по причине исчезновения растительности. Благодаря этому вся растительность, которая была зафиксирована в течение месяца, сохранена на итоговом изображении.

Это позволяет наблюдать за динамикой каждый месяц. Данное время можно считать хорошим периодом наблюдения для данной задачи, так как изменения растительности происходят сравнительно медленно и шаг в один месяц достаточен для наблюдения динамики.

В результате осуществления классификации песка и почвы, как разных типов поверхностей, были созданы модели, детектирующие два этих класса с высокой точностью для задач детекции. Лучшие результаты показали такие методы, как Случайный лес - F-мера=95.18% и Метод опорных векторов - F-мера = 94.55%. Это позволяет определять области, наиболее благоприятные

для посадки растений. Кроме того, эти области обладают повышенным приоритетом, так как при возникновении проблемы опустынивания, регион, столкнувшийся с ней, будет заинтересован в сохранении максимально большой части территорий, пригодных для роста растений.

Данная классификация может быть также полезна в том случае, если возникает потребность построить карту скоростей распространения песка. В некоторых регионах скорость роста пустыни достаточно велика, чтобы за несколько лет она была хорошо заметна на спутниковых снимках сервиса Sentinel Hub.

Полученная точность детекции является достаточно высокой, из чего следует, что имеет смысл модернизировать и развивать этот метод дальше. Разработанный алгоритм для создания классификатора почвы и песка может быть применен и к задачам классификации, нацеленным на другие классы поверхности.

При использовании не только данного метода детекции песка и почвы, но и методов, учитывающих взаимное расположение пикселей на изображении, есть вероятность еще больше увеличить точность классификации.

Еще одной отличительной чертой данного метода является то, что он использует обучение с учителем, в то время как для подобного рода задач популярны методы обучения без учителя. Например кластеризация методом K-means [1].

В результате расчета территории, занятой растительностью, была продемонстрирована способность таким образом получать возможность видеть общую картину происходящего на выбранной территории.

Это позволило собрать количественную информацию о растительности на исследуемой территории и подкрепить уже известную информацию о неблагоприятной ситуации точными расчетами.

Результаты, полученные в ходе расчета, согласуются с информацией о событиях, произошедших на исследуемой территории, что подтверждает правильность произведенных расчетов.

Благодаря реализованному методу расчета площади занятой территории, можно получать данные о размере площади, которую занимает любой заданный класс, зафиксированный на маске территории.

Список литературы

[1] Hartigan J.A., Wong M.A. Algorithm AS 136: A k-Means Clustering Algorithm // Journal of the Royal Statistical Society, Series C. 28(1), pp. 100-108

[2] Measuring Vegetation // NASA Earth Observatory, 2000

[3] Cortes C., Vapnik V. Support-Vector Networks // Kluwer Academic Publishers, Boston, Number 3, Volume 20, pp. 273 - 297, 1995

[4] Kalles D., Morris T. Efficient Incremental Induction of Decision Trees // Machine Learning 24(3), pp. 231-242, 1996

[5] Cutler A., Cutler D.R, Stevens J.R. Random Forests // Ensemble Machine Learning: Methods and Applications, pp. 157-176, Chapter 5, Springer, 2011

[6] Tsuruoka Y., Tsujii J., Ananiadou S. Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty // Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 477–485,Suntec, Singapore, 2-7 August 2009

[7] Ghojogh B., Crowley M. Linear and Quadratic Discriminant Analysis: Tutorial // 2019

[8] Sasaki Y. The truth of the F-measure // 2007

[9] Токарева О.С., Обработка и интерпретация данных дистанционного зондирования Земли. Учебное пособие // Издательство Томского политехнического университета, 2010

[10] Ball J.E., Anderson D.T., Chan C.S. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community // Journal of Applied Remote Sensing 11(04), 2017

[11]Barrile V., Bilotta G. An application of Remote Sensing: Object oriented analysis of satellite data // 2008

[12] Dietterich T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization // Machine Learning, 40(2), pp. 139-157, 2000

[13] Ferris M.C., Munson T.S. Interior-Point Methods for Massive Support Vector Machines // SIAM Journal on Optimization, 13(3), pp. 783-804, 2008
[14] Frasheri N., Cico B., Paci H., Bushati J. Use of Remote Sensing (Satellite Images) For Assessing the Environment Situation // 2010
[15] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning // Springer, 2008
[16] Li Y., Chen J., Ma Q., Zhang H.K., Liu J. Evaluation of Sentinel-2A Surface Reflectance Derived Using Sen2Cor in North America // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 1-25, 2018
[17] Sentinel Hub documentation // https://docs.sentinel-hub.com/api/latest/
[18] Supervised learning - scikit-learn 0.24.2 documentation // https://scikit-learn.org/stable/supervised_learning.html