

Документация моделей.



Список моделей

- Модель детекции нефти.
- Модель распространения нефти.
- Модель распространения лесных пожаров.
- Модель наводнений.
- Модель предсказания пожаров.
- Другие модели.



Модель детекции нефти.



Основные компоненты.

- Постановка задачи.
- Входные и выходные данные.
- Данные для обучения.
- Варианты моделей.
- Описание лучшей модели.



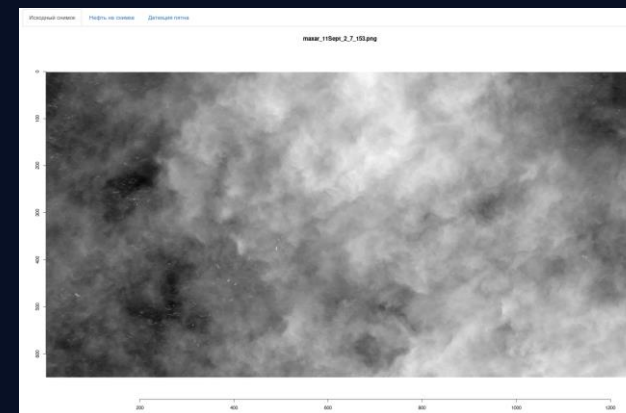
Постановка задачи.

- Есть снимки высокой точности, на которых может быть нефтяное пятно. Масштаб снимков может быть самым разным.
- Необходимо разработать модель, которая бы отмечала нефтяное пятно на снимках.
- С точки зрения ML это задача сегментации изображений.



Входные и выходные данные.

- Вход: снимок.
- Выход: маска.
Белым обозначена нефть



Данные для обучения.

- Качались снимки Махар, посвященные Wakashio Oil Spill.
- Применялся weak labelling (разметка на основе цветов областей, кластеризация с помощью HDBSCAN, выбор кластеров).
- Данные аугментировались, следуя **Krestenitis et al, Oil spill identification from satellite images using deep neural networks.**
- Суть аугментации:
 1. Нарезка до снимков 1250 на 650.
 2. Случайное изменение снимков в размере.
 3. Соккрытие участка размера 320 на 320 в случайном месте.
 4. И другие похожие преобразования.



Варианты моделей.

- Модели брались на основе работ **Krestenitis et al, Oil spill identification from satellite images using deep neural networks** и **Krestenitis et al, Early identification of oil spills in satellite images using deep CNNs**.

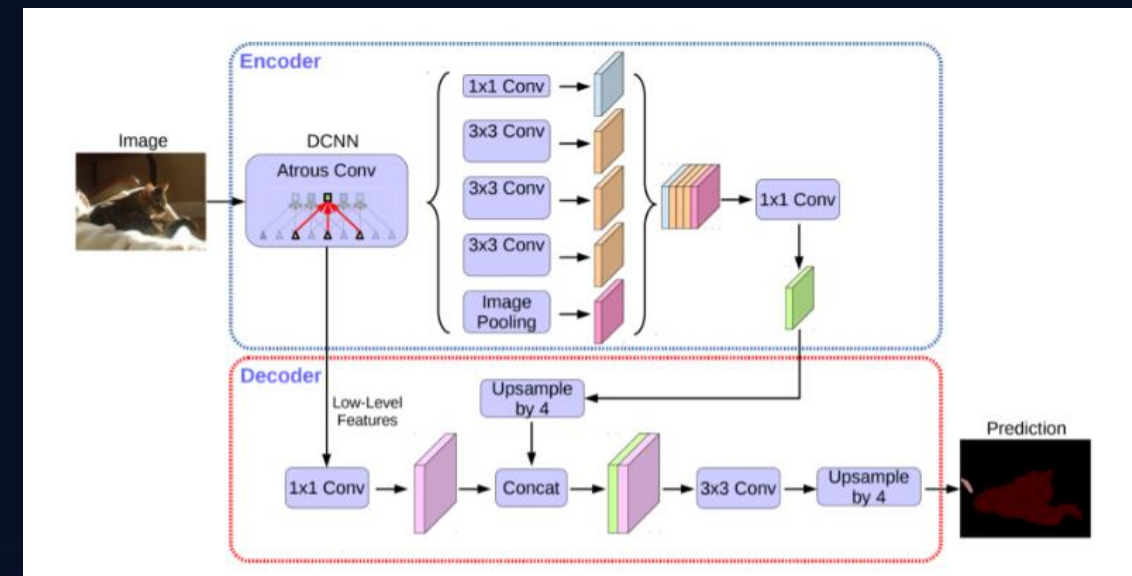
Название модели	Mean IoU
DeepLabV3+ с ResNet50	66.49%
DeepLabV3+ с MobileNetV2	62.81%
UNet	65.32%

- У **Krestenitis et al** на других данных получилось **65.06%**



Описание лучшей модели.

- **DeepLabV3+** --- SOTA для задачи сегментации изображений.
- Есть возможность смотреть на снимки на разных масштабах за счёт **atrous convolution** и **atrous spatial pyramid**.
- Использование предобученной нейронной сети для задачи классификации в качестве **backbone (ResNet50)**.
- Современная Encoder-Decoder архитектура.



Модель распространения нефти.



Основные компоненты.

- Постановка задачи.
- Входные и выходные данные.
- Базовая модель.
- Модель для течений: формализация.
- Модель для течений: данные.
- Модель для течений: выбор модели.
- Модель для течений: лучшая модель.
- Модель волн.
- Модель испарения.
- Модель эмульсификации.
- Модель вертикального транспорта.
- Объединённая модель.



Постановка задачи.

- Есть снимки или аналоги, на которых выделено нефтяное пятно.
- Необходимо смоделировать как может распространяться нефтяное пятно с течением времени.
- Необходимо смоделировать как на поведение пятна влияют разные параметры.
- Необходимо совершать прогнозирование прошлого.



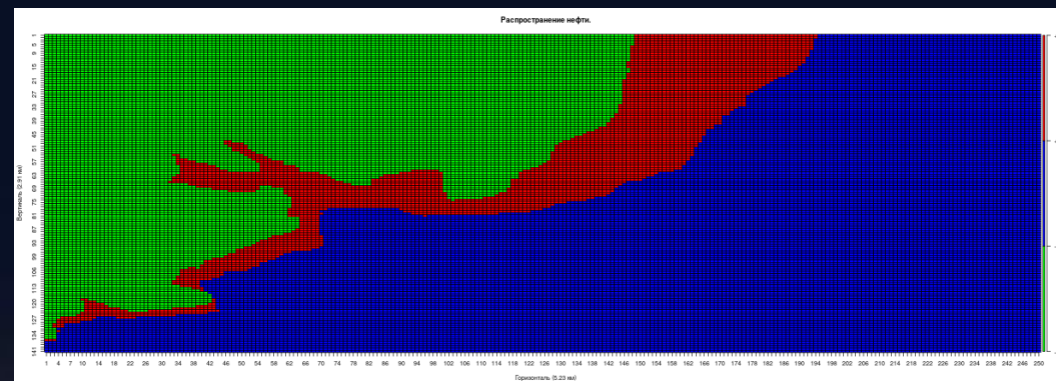
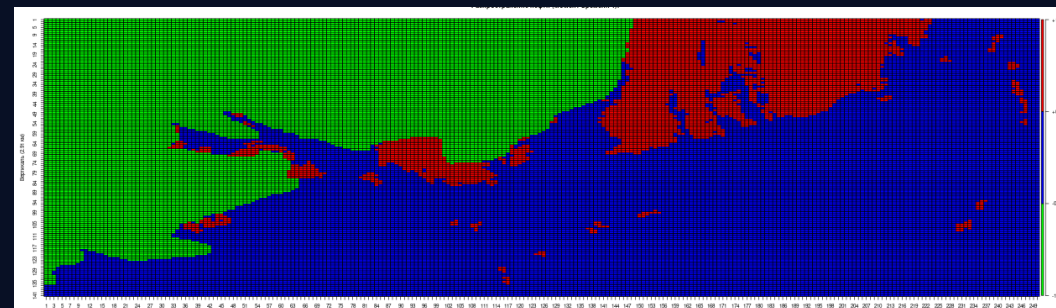
Входные и выходные данные.

Входные данные:

- Время.
- Скорость ветра (вектор)
- Температура воды.
- Другие параметры специфичные для моделей-компонент.

Выходные данные:

- Пересчитанное пятно на снимке на конкретный момент времени.



Базовая модель.

- Уравнение адвекции-диффузии на концентрацию.
- Параметры: скорость ветра и коэффициент диффузии.
- Пересчёт коэффициента диффузии по температуре (см. E. Ao-Leong et al, Modeling the BP Oil Spill of 2010: a simplified model of oil diffusion in water).
- Граничные условия соответствуют обнаруженному пятну.
- Дискретизация и сведение к системе ОДУ.
- Решение системы ОДУ методом Адамса.
- Скорость ветра может зависеть от времени и от пространства.
- Учёт суши проводится отдельным скриптом.
- Обратная задача у УРЧП не разрешима, у приближения системой ОДУ разрешима.
- Модель основывается на статье **Duran et al, Horizontal transport in oil-spill modeling.**



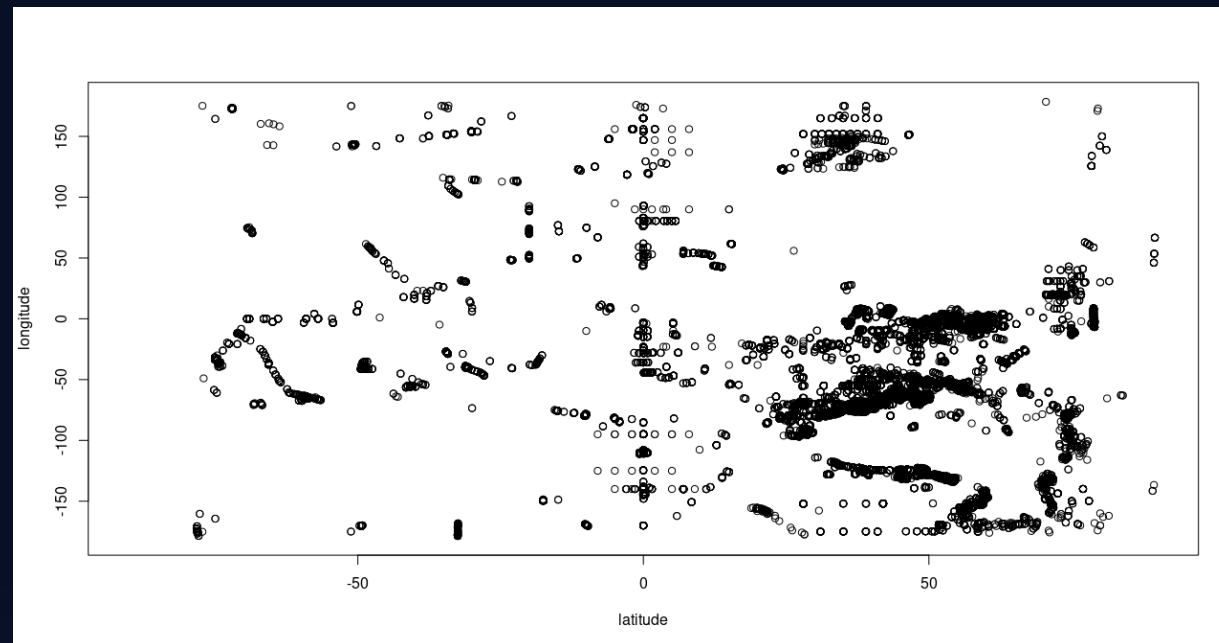
Модель для течений: формализация.

- Есть данные о течениях.
- Мы приближаем нашу точку некоторым кластером.
- Прогнозируем скорость каждой точки кластера на определённый момент времени.
- Скорость точки вычисляем как средневзвешенное по расстоянию от спрогнозированных скоростей кластера.
- Ключевой момент: модель прогнозирования временных рядов.



Модель для течений: данные.

- Данные о течениях
Woods Hole Oceanographic Institution.
- Короткие временные ряды.
- Зависимость от пространства.
- В каждой точке несколько признаков



Модель для течений: выбор модели.

Название.	SMAPE.
Бейзлайн.	51.17%
Catboost + HDBSCAN + SMA	55.41%
Catboost + HDBSCAN + SMA + корреляция	53.28%
Блендинг бейзлайна и HDBSCAN.	49.17%
EMA	49.38%
Эвристика.	50.46%
SMA	50.37%
Блендинг бейзлайна, HDBSCAN, EMA.	47.99%
VAR	51.29%
Блендинг бейзлайна, HDBSCAN, EMA, Catboost	47.91%
Блендинг бейзлайна, HDBSCAN, EMA, VAR.	46.21%

Модель для течений: лучшая модель.

- Включает эвристику.
- Включает модель, основанную на приближении по кластерам, построенным с помощью HDBSCAN (SOTA метод кластеризации).
- Включает EMA (хорошо подходит для прогнозирования одиночных небольших рядов).
- Включает VAR (хорошо подходит для прогнозирования нескольких связанных рядов).
- Модель является блендингом всех вышеупомянутых моделей.



Модель ВОЛН.

- Основная компонента: скорость ветра.
- Другие компоненты можно считать постоянными.
- Модель **Sverdrup, Munk, Bretschneider**.

$$H_s = 0,283\alpha \cdot \frac{W_s^2}{g} \tanh\left(\frac{0,0125}{\alpha} \left(\frac{gF}{W_s^2}\right)^{0,42}\right)$$

$$T_p = 7,54\beta \cdot \frac{W_s}{g} \tanh\left(\frac{0,077}{\beta} \left(\frac{gF}{W_s^2}\right)^{0,25}\right)$$

$$\alpha = \tanh\left\{0,53 \left(\frac{gH}{W_s^2}\right)^{0,75}\right\}$$

$$\beta = \tanh\left\{0,833 \left(\frac{gH}{W_s^2}\right)^{0,375}\right\}$$



Модель испарения.

- Основная компонента: скорость ветра.
- Другие компоненты можно считать постоянными.
- Модель из **M. Fingas, Evaporation Modeling.**

$$E = KC_s U^{7/9} d^{-1/9} Sc^{-r}$$



Модель эмульсификации.

- Основная компонента: скорость ветра.
- Другие компоненты можно считать постоянными.
- Модель взята из **K. Aghajanloo et Al, Three dimensional numerical modeling of oil spill behavior in Marine Environment.**

$$\frac{dF_{wc}}{dt} = K_{wc}(U_{wind} + 1)^2 \left(\frac{1 - F_{wc}}{OC} \right)$$

- Если скорость ветра постоянно, то ОДУ решается явно.
- Если нет, то ОДУ решается приближёнными методами.



Модель вертикального транспорта.

- Основная компоненты: скорость ветра и коэффициент диффузии.
- Как и в базовой модели, они могут зависеть от времени и от пространства.
- УРЧП решается теми же методами, что и базовая модель.
- Проблема с определением в реальности начальных условий.
- Модель взята из **K. Aghajanloo et Al, Three dimensional numerical modeling of oil spill behavior in Marine Environment.**

$$\frac{\partial C}{\partial t} + \frac{\partial(uC)}{\partial x} + \frac{\partial(vC)}{\partial y} + \frac{\partial((-w_b)C)}{\partial z} = \frac{\partial}{\partial x} \left(D_h \frac{\partial C}{\partial x} \right) + \frac{\partial}{\partial y} \left(D_h \frac{\partial C}{\partial y} \right) + \frac{\partial}{\partial z} \left(D_z \frac{\partial C}{\partial z} \right) \pm S$$



Объединённая модель.

- Базовая модель берётся за основу.
- Модель вертикального транспорта выключается из-за сложности указания начальных условий.
- Скорость из модели течений вставляется с определённым весом в член адвекции в базовую модель.
- Скорость из модели волн вставляется с определённым весом в член адвекции в базовую модель.
- Модель испарения применяется к концентрации после базовой модели.
- Модель эмульсификации применяется к концентрации после базовой модели.



Модель распространения лесных пожаров.



Основные компоненты.

- Постановка задачи.
- Входные и выходные данные.
- Нахождение очагов.
- Определение типа леса.
- Модель пожаров кроны.
- Модель поверхностных пожаров.
- Объединённая модель.



Постановка задачи.

- Есть снимок с лесным пожаром на нём.
- Необходимо смоделировать поведение лесного пожара с течением времени.
- Необходимо смоделировать поведение лесного пожара в зависимости от различных условий.
- Предпочтение отдаётся наихудшему прогнозу, чтобы определить дойдёт ли пожар до жилого поселения.

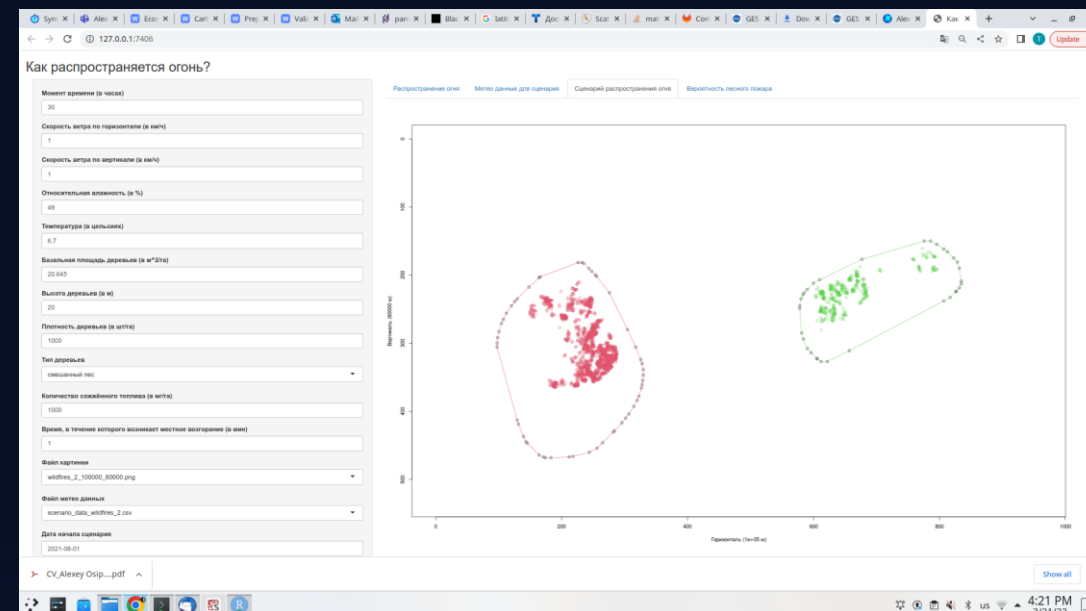


Входные и выходные данные.

Входные параметры:

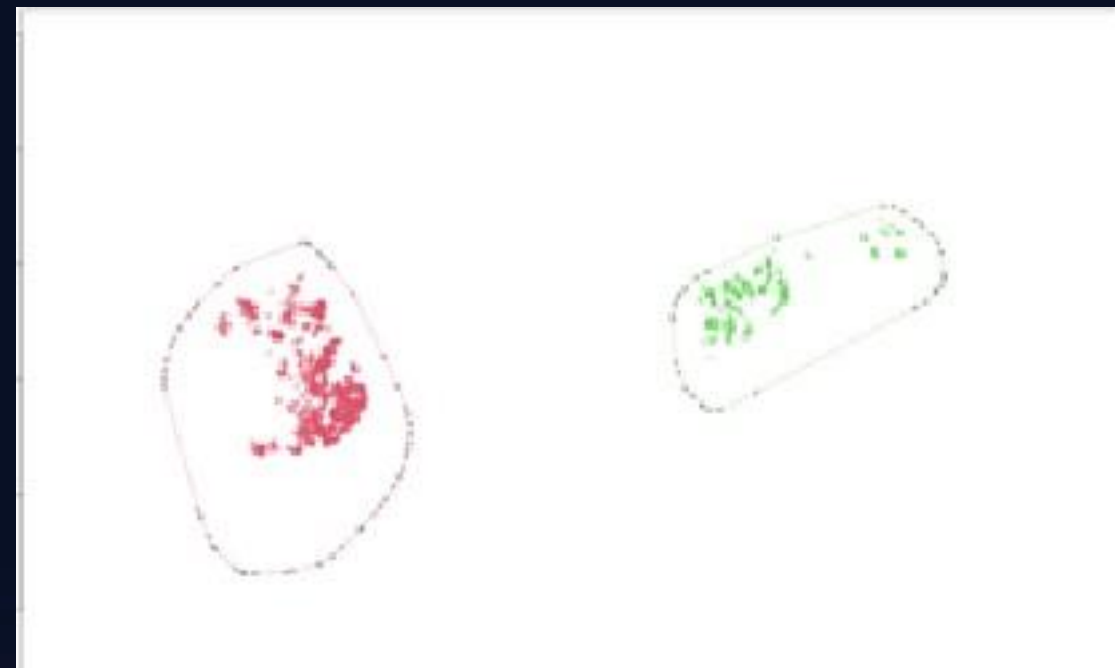
- Время.
- Скорость ветра (вектор).
- Относительная влажность.
- Температура.
- Тип деревьев.
- Характеристики леса (разные параметры).
- Наклон местности.

Выходные параметры:
снимок, на котором отмечена максимальная область распространения пожара для каждого очага.



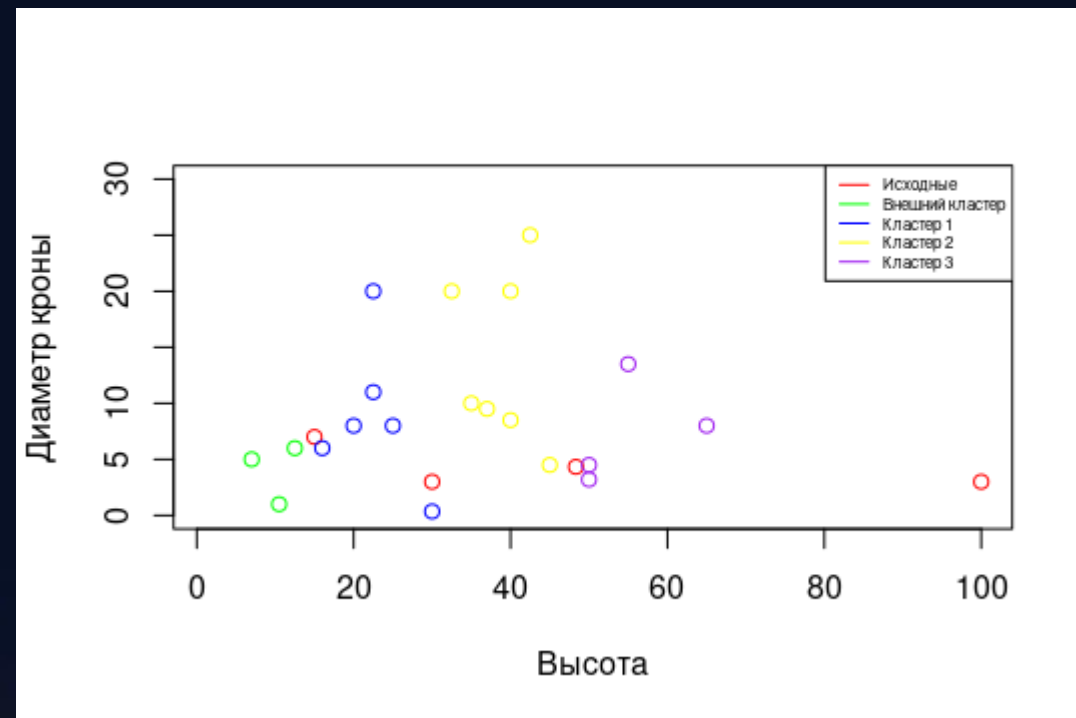
Нахождение очагов.

- Детекция пожара осуществляется с помощью эвристик (rule based model).
- Кластеры очагов вычисляются с помощью HDBSCAN.
- HDBSCAN выбран потому, что это SOTA алгоритм, который сам может вычислять число кластеров.
- Далее область распространения вычисляется индивидуально для каждого очага.



Определение типа леса.

- Поддерживаются несколько десятков типов леса.
- В модели **Alexander & Cruz** поддерживаются 4 типа леса.
- Данные типов леса берутся из открытых статей.
- Далее проводится интерполяция и экстраполяция, типы леса из модели **Alexander & Cruz** берутся за базовые.



Модель пожаров кроны.

- Пожары кроны распространяются гораздо быстрее чем лесные пожары.
- В работе **Alexander & Cruz, Evaluating a model for predicting active crownfire rate of spread using wildfire observations** описывается модель распространения пожаров кроны.
- Авторы сделали сравнение своей модели с другими, и пришли к выводу, что их модель лучше.
- В качестве модели пожаров кроны берётся модель из этой статьи.



Модель поверхностных пожаров.

- Классической моделью распространения являются модель **Rothermel** и её вариации.
- Модель взята из работы **J.H. Scott et al, Assessing crown fire potential by linking models of surface and crown fire behavior.**
- В качестве модели поверхностных пожаров берётся модель из этой статьи.
- Модель содержит несколько десятков параметров. Поэтому был проведён анализ чувствительности для выбора наиболее важных параметров. Остальные параметры были заменены на разумные значения по умолчанию.



Объединённая модель.

- Очаги вычисляются с помощью модели очагов.
- Определяется тип леса.
- Вычисляются другие параметры важные для моделей пожаров кроны и поверхностных пожаров.
- Применяется модель пожаров кроны. Она возвращает вероятность пожара кроны.
- Применяется модель поверхностного пожара. Считается средневзвешенное двух моделей, где вес привязан к вероятности.
- Считается выпуклая оболочка каждого очага. Эта выпуклая оболочка распространяется в соответствии с данными моделей. Сейчас это кастомный алгоритм распространения. В будущем можно решать уравнение адвекции.



Модель наводнений.



ОСНОВНЫЕ КОМПОНЕНТЫ.

- Постановка задачи.
- Входные и выходные данные.
- Описание медленной модели.
- Быстрые модели: основная идея.
- Быстрые модели: модель прогнозирования.
- Быстрые модели: модель распространения.



Постановка задачи.

- Есть река. Мы хотим смоделировать возможное наводнение, вызванное этой рекой.
- Желательно учитывать факторы плотин.



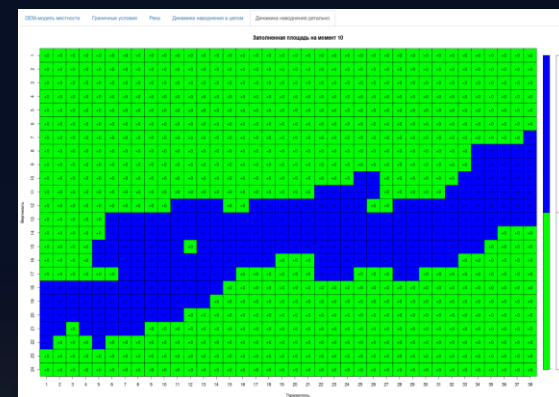
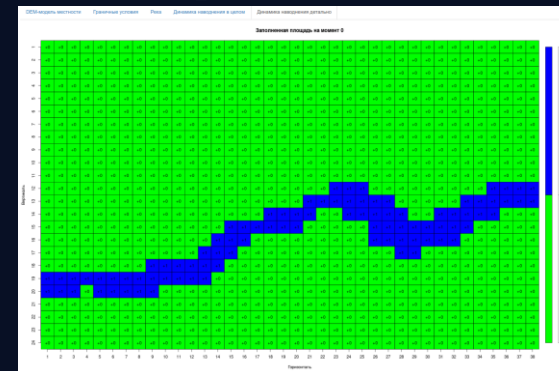
Входные и выходные параметры.

Входные параметры:

- Время
- DEM-модель местности (берётся из данных ASTER).
- Позиция реки на снимке.
- Позиция плотины на снимке.
- Коэффициент Гоклера-Маннинга
- Коэффициент шероховатости реки

Выходные параметры:

- снимок, на котором разлилась река



Описание медленной модели.

- **LISFLOOD-FP** это гибкая модель, которая описывается во многих статьях.
- Модель взята отсюда: [LISFLOOD-FP | School of Geographical Sciences | University of Bristol](#)
- В ней есть возможность учитывать множество факторов: плотины, мосты, дождь, ...
- Эта модель в целом выдаёт хорошие результаты, однако, она является медленной. Её ускорению посвящено достаточно много статей.
- Поэтому было решено взять LISFLOOD-FP за основу и построить быструю модель, которая бы имитировала её результаты.



Быстрые модели: основная идея.

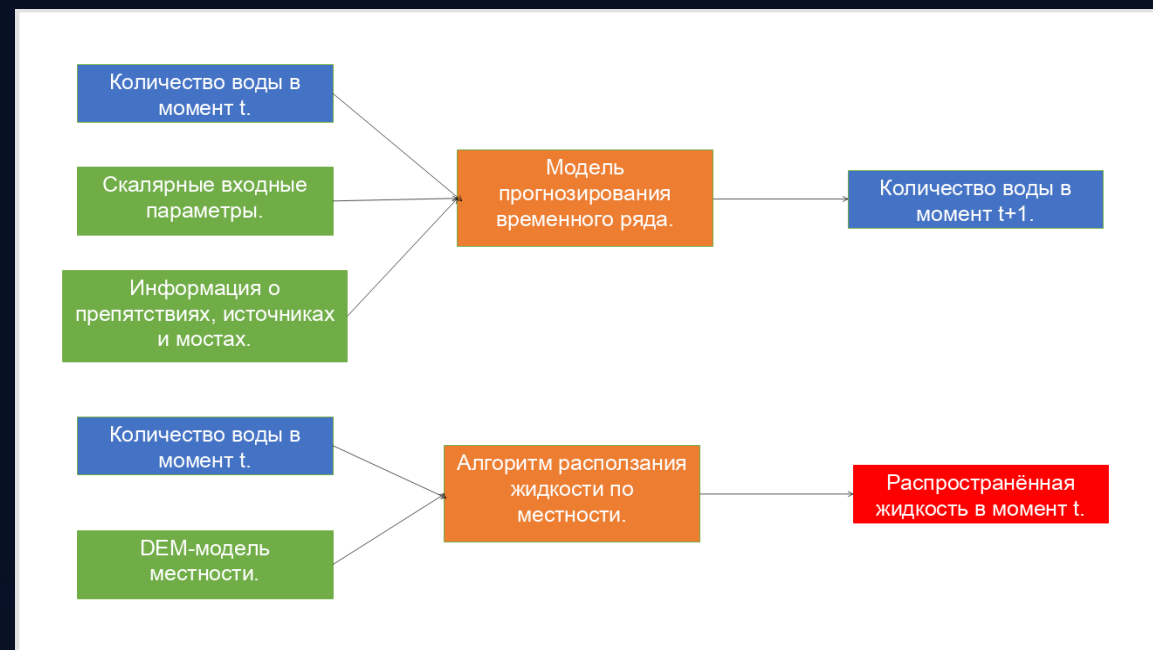
Две подзадачи:

- Задача прогнозирования временного ряда количества воды.
- Задача распространения фиксированного объёма жидкости.

Первая подзадача решается с помощью методов анализа временных рядов.

Вторая подзадача решается с помощью детерминированного алгоритма.

Для решения первой подзадачи строится датасет на основе результатов LISFLOOD-FP.



Быстрые модели: прогнозирование.

Название модели.	MAPE.
Catboost regression.	2.39%
Huber regression.	9.2%

Результаты **Catboost regression** лучше, но **Huber regression** является интерпретируемой моделью. Туда проще добавлять новые факторы.

Есть возможность вызывать обе модели (равно как и саму LISFLOOD-FP), но рекомендуется использовать **Huber regression**.



Быстрые модели: распространение.

Вот так выглядит детерминированный алгоритм распространения:

- У нас задача разлить определённое количество воды.
- Мы действуем итеративно, храня в памяти краевые точки реки.
- На каждой итерации мы выбираем одного соседа, исходя из DEM-модели местности и заполняем его.
- При определённых условиях алгоритм осушает неправильных соседей, начальная река не меняется.
- Все входные параметры учитываются в модели прогнозирования, модель распространения только распространяет воду наилучшим возможным способом.



Модель предсказания пожаров.



Основные компоненты.

- Постановка задачи.
- Данные.
- Обработка данных.
- Модель предсказания.
- Базовый вариант модели детекции.
- Улучшенный вариант модели детекции.



Постановка задачи.

- По данным снимкам необходимо оценить вероятность пожара в зависимости от различных условий.
- На данных снимках необходимо выделить область гарей.



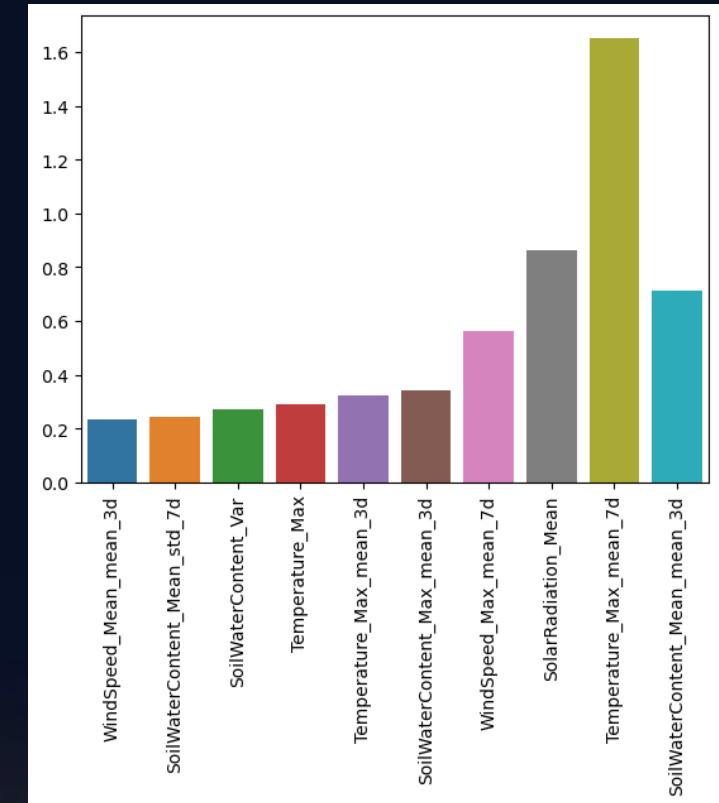
Данные.

- Используются открытые данные соревнования **Predicting Wildfires with Weather Forecast Data.**
- Признаки:
 1. Температура
 2. Относительная влажность.
 3. Скорость ветра.
 4. Атмосферные осадки
 5. Влажность почвы.
 6. Радиация почвы.



Обработка данных.

- Данные агрегируются в дневные значения различными способами (среднее, минимум, максимум, сумма, дисперсия).
- Считаются скользящие характеристики с окнами 3 и 7 дней, считаются среднее и среднеквадратичное отклонение.
- В итоге получается 66 параметров.
- Справа показаны наиболее значимые признаки согласно модели.



Модель предсказания.

- Использование бэггинга при обучении для исправления баланса классов (дисбаланс 65% на 35%).
- Логистическая регрессия (SGD classifier с log loss) с L2 регуляризацией.
- Есть возможность дообучения.

Метрика.	Значение.
Accuracy.	73.33%
ROC AUC.	64.81%
F1 score.	80%

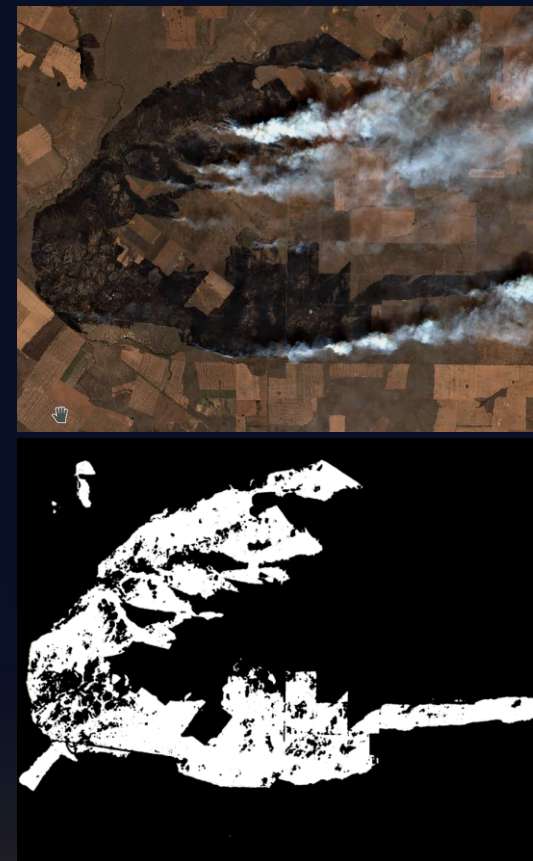
Базовая модель детекции.

- Берется комбинация
 1. масок, полученных из `delta_NBRSWR`
 2. масок, полученных на основе `max_pooling` канала `SWIR2` с окном 121 на 121 с шагом 1 (с добавлением нулей).
- Проблема этого подхода в случайных выделенных водных объектах, в плохом выделении сгоревшей области пожара внутри контура.



Улучшенная модель детекции.

- Берутся:
 1. маска по пороговому значению для NBRSWIR $(swir1 - swir2) / (swir1 + swir2)$.
 2. маска по пороговому значению для NDVI $(nir + red) / (nir - red)$.
 3. маска по пороговому значению для NBR $(swir2 - nnir) / (nnir + swir2)$.
- Маски объединяются через суммирующий фильтр размера 7 на 7.
- Сравнение с пороговым значением.



Другие модели.



Общий список других моделей.

- Химические аварии.
- Взрывы ТВВ.
- Взрывы пылевоздушной смеси.
- Пожары проливов ЛВЖ и ГЖ.
- Взрывы бытового газа.
- Охлаждение зданий при отключении отопления.
- Прогнозирование рудных месторождений.
- Выбросы CO₂ и CH₄.
- Заморозки.

