УДК 519.6

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАПОЛНЕНИЯ ПРОПУСКОВ В ДАННЫХ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ

Сарафанов М.И. (Университет ИТМО, ФГБУ «Государственный Гидрологический Институт»), **Никитин Н.О.** (Университет ИТМО), **Казаков Э.Э.** (ФГБУ «Государственный Гидрологический Институт»)

Научный руководитель – к.т.н., доцент Калюжная А.В.

(Университет ИТМО)

Аннотация. В работе представлены результаты разработки алгоритма, который на основе методов машинного обучения позволяет заполнять пропуски в данных дистанционного зондирования. На данных теплового зондирования (Sentinel-3) для территории Санкт-Петербурга произведена верификация модели.

Введение. Использование данных дистанционного зондирования Земли сегодня позволяет изучать пространственные системы в самых различных масштабах. Используя спутниковые снимки можно анализировать поведение климатической системы, следить за развитием отдельных биомов и экосистем, изучать состояние и выявлять зону влияния техногенных объектов и многое другое. Все это было бы невозможно без сложного оборудования, а также без хорошо адаптированных алгоритмов обработки данных дистанционного зондирования. Однако из-за специфики съемки некоторые продукты не имеют достаточного пространственного охвата, имеют слишком низкое пространственное разрешение или имеют большое количество пропусков, которые, как правило, вызваны облачностью. Например, для высоких широт облачность присутствует над территориями большую часть года, поэтому львиная доля спутниковых снимков в таких условиях поступает с пропущенными значениями. Вести мониторинг по данным зондирования в таком случае без применения алгоритмов восстановления пропущенных значений затруднительно.

Такие алгоритмы заполнения пропусков, как во временных рядах, так и в пространственных данных, существуют. Однако большинство из таких алгоритмов не имеют открытого исходного кода, требуют сложной подготовки данных или не учитывают особенности восстанавливаемого параметра.

Цель данной работы - разработка алгоритма заполнения пропусков в данных дистанционного зондирования с использованием спутниковых изображений температуры поверхности Земли (спутниковая система Sentinel-3) в качестве примера.

Основная часть. Для реализации алгоритма восстановления пропущенных значений в спутниковых снимках был использован язык программирования Python. Было решено использовать следующий подход: для каждого пропущенного пикселя строится своя модель, которая восстанавливает пропуск, опираясь на известные значения в пикселях на этом же снимке. Таким образом, предикторами выступают определенным образом выбранные известные значения в пикселях данного снимка. В качестве обучающей выборки используются предыдущие снимки для данной территории. Предсказания строятся на основе методов машинного обучения, таких как регрессия LASSO, k-ближайших соседей, случайный лес, метод опорных векторов.

Предикторы выбирались по трём стратегиям. Первая – в качестве предикторов выступают все известные значения пикселей на снимке. Данных подход требует больших вычислительных затрат, поэтому время работы алгоритма является очень большим. С другой стороны, потенциально, в данном случае мы вовлекаем большее количество информации в модель. Вторая стратегия – используются 100 случайно выбранных точек на снимке. В таком случае алгоритм работает быстро, но результат недостаточно точный. Третья стратегия – предикторами являются значения из пикселей, которые принадлежат тому же биому, что и

пропуск. Если известных пикселей из того же биома слишком много, используются 40 ближайших точек (по Евклидовой метрике) из данного биома. Такой подход позволяет добиться высокой точности наравне с небольшим временем работы. Минусом данной стратегии является необходимость получить матрицу (в данном случае – матрицу биомов), которая позволяла бы разбить пиксели на снимках на группы. С другой стороны, это позволяет использовать данные о внутренней структуре пространственных данных, что существенно повышает точность.
Для верификации модели были выбраны 6 снимков, для которых были сгенерированы пропуски разной формы и размера. Такие пропуски могли занимать от 3% до 90% от всей территории на снимке. В обучающую выборку вошло 350 спутниковых снимков. Для проверки точности алгоритма были использованы 3 метрики: средняя абсолютная ошибка, средняя медианная ошибка, квадратный корень из среднеквадратической ошибки.

Выводы. В результате проведенных исследований был реализован и верифицирован (на данных теплового зондирования со спутников Sentinel-3) алгоритм заполнения пропусков в пространственных данных.
Подготовленная реализация позволяет заполнять пропуски с учетом пространственной неоднородности территории, имеется возможность выбора определенного алгоритма машинного обучения и способа подбора гиперпараметров (поиск по сетке, случайный поиск по сетке или пользовательская настройка), а также стратегии подбора предикторов для модели.
В ходе экспериментов наиболее точным оказался метод опорных векторов с использованием предикторов из того же биома, что и пропущенные значения. В большинстве случаев ошибка модели не превышала 1°C.