

**РАЗРАБОТКА МЕТОДА ВОСПОЛНЕНИЯ ДАННЫХ О ПРОЕКТНОЙ МОЩНОСТИ
ОБЪЕКТОВ СОЦИАЛЬНОЙ ИНФРАСТРУКТУРЫ**

Беляев А.В. (ИТМО),

Научный руководитель – эксперт Института дизайна и урбанистики Кудинов С.А.
(ИТМО)

Введение. Объекты социальной инфраструктуры играют ключевую роль в обеспечении качества жизни населения, а их оценка требует применения современных методов анализа. В этом контексте школы, благодаря своей высокой гомогенности и однородности, представляют собой оптимальный объект для апробации методов восполнения данных о проектной мощности. Современные алгоритмы, успешно применяемые для анализа визуальных данных, сталкиваются с ограничениями точности при работе с реальными пространственными данными. Цель данного исследования заключается в тестировании метода восполнения данных о проектной мощности объектов социальной инфраструктуры с использованием как реальных, так и сценарных данных, что особенно актуально в условиях увеличения объёма недоступной информации.

Основная часть. В исследовании используются геопространственные данные, представленные в виде файла `.geojson`, содержащего мультиполигональные объекты в метрической системе координат. Выбор данной системы обусловлен необходимостью корректного расчёта площадей и периметров полигонов. Модель, восполняющая данные о проектной мощности школ, состоит из двух компонентов: основного кода, осуществляющего обучение модели, и дополнительного кода, предназначенного для добавления геометрических атрибутов новым объектам после завершения обучения. Данные загружаются с помощью библиотеки `geopandas`, при этом выполняется предварительная обработка: очистка заголовков, проверка наличия ключевых переменных (например, этажность, площадь, периметр) и приведение данных к корректным типам, а также замена пропусков значениями `NULL`.

После подготовки данных производится вычисление геометрических характеристик мультиполигональных объектов. Среди рассчитываемых параметров – компактность, отношение площади к периметру, длина и ширина минимального описанного прямоугольника, количество вершин, координаты центроида и степень отклонения от выпуклой формы. При этом этажность берётся из исходных данных, поскольку её невозможно корректно вычислить исключительно по форме полигона. Далее данные разделяются на обучающую и тестовую выборки в соотношении 80:20, где ключевые геометрические параметры используются в качестве признаков, а вместимость школ – в качестве целевой переменной.

Для повышения точности прогнозирования применяются несколько моделей машинного обучения: `XGBoost`, `CatBoost`, `LightGBM`, `Random Forest` и `Gradient Boosting` [1-3]. Оптимизация гиперпараметров модели `XGBoost` осуществляется с помощью библиотеки `optuna`, что позволяет подобрать оптимальные значения для таких параметров, как глубина деревьев, скорость обучения и степень регуляризации. Предсказания, полученные различными моделями, объединяются с использованием ансамблевой техники методом взвешенного голосования, а качество модели оценивается по метрике `RMSE`. Эксперимент включает четыре сценария, в каждом из которых последовательно изменяются параметры модели, состав выборки и набор геометрических характеристик.

В ходе исследования разработаны обоснования и четыре сценария для оценки качества метода восполнения данных о проектной мощности школ как объектов социальной инфраструктуры. В базовом сценарии (сценарий 0) использовались три основных показателя – площадь, периметр и этажность, что приводило к недостаточной точности расчётов. Добавление шести дополнительных геометрических параметров в сценарии 1 улучшило результаты, однако достигнутой точности оказалось недостаточно. Сценарий 2 проводил анализ влияния координат центроидов на точность модели, а сценарий 3, предусматривающий изменение выборки для повышения однородности данных и минимизации влияния выбросов, показал наилучшие результаты. Дополнительные расчёты в сценарии 4, выполненные на платформе `SMILE` [4] посредством матрицы исключения показателей, позволили оценить вклад каждого параметра в итоговую точность модели.

Выводы. В сценарии 3, который продемонстрировал наилучшие результаты, были

исключены координаты центроидов, а выборка была сокращена за счёт исключения школ-интернатов и объектов с частичным представлением данных. В этом сценарии 62,6% объектов попадают в диапазон точности 0,9–1,1, а 77,7% – в диапазон 0,8–1,2. Несмотря на то, что сценарий 3 обеспечил максимальную точность, для достижения целевых показателей необходимо расширение выборки и применение метода к другим типам объектов социальной инфраструктуры, таким как детские сады. Таким образом, проведённое исследование демонстрирует практическую возможность использования комбинированного подхода с ансамблевыми методами для восполнения данных о проектной мощности, что имеет важное значение для планирования и развития объектов социальной инфраструктуры.

Список использованных источников:

1. Dabrock, K., Pflugradt, N., Weinand, J., Stolten, D. Leveraging Machine Learning to Generate a Unified and Complete Building Height Dataset for Germany. (2023) January 2023
2. Atik, M.E.; Duran, Z.; Seker, D.Z. Machine Learning-Based Supervised Classification of Point Clouds Using Multiscale Geometric Features. // ISPRS Int. J. Geo-Inf. 2021, 10, 187.
3. Lou W, Wang X, Chen F, Chen Y, Jiang B, et al. Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. PLOS ONE 9(1): e86703. 2014
4. Платформа SMILE (Simple Machine Learning Editor) [Электронный ресурс]. – URL: <https://smile.actcognitive.org/module/2313/> (дата обращения: 20.12.2024)