Желаемая секция: машинное обучение в задачах наук о Земле.

**Восстановление приповерхностной влажности атмосферы над океаном по данным сопутствующих метеорологических измерений с применением методов машинного обучения**

**Reconstruction of atmospheric surface humidity over the ocean from relevant meteorological measurements using machine learning methods**

*Вострикова С.А.1*

Email: vostrikova.sa@phystech.edu

Номер телефона: 89878750480

*Криницкий М.А.1,2*

Email: krinitsky@sail.msk.ru

Номер телефона: 89261416200

1Московский физико-технический институт,

Институтский пер., 9, Долгопрудный, Московская обл.

2 Институт океанологии им. П.П. Ширшова Российской академии наук

Нахимовский пр., 36, Москва

*Vostrikova S.A.1*

Email: vostrikova.sa@phystech.edu

Phone number: 89878750480

*Krinitsky M.A.1,2*

Email: krinitsky@sail.msk.ru

Phone number: 89261416200

1Moscow Institute of Physics and Technology,

Institutsky per., 9, Dolgoprudny, Moscow region.

2Shirshov Institute of Oceanology, Russian Academy of Sciences

36 Nakhimovskiy pr., Moscow

Влажность воздуха в приповерхностном слое атмосферы над океаном является ключевым климатическим параметром, оказывающим значительное влияние на процессы переноса влаги и тепла между океаном и атмосферой, а также на динамику атмосферных процессов в целом. Анализ метеорологических данных, собранных в течение XX века, показывает разреженность рядов измерений влажности в пространстве и времени [1]. Международный массив данных о характеристиках океана и атмосферы (ICOADS [4]) указывает на недостаточную плотность измерений в начале XX века по сравнению с более поздними периодами [1], что создает сложности для адекватного анализа климатических тенденций относительной влажности.

Представленные в литературе методы восстановления временных рядов влажности [6, 7] зачастую демонстрируют ограниченную точность, основываясь преимущественно на статистических и эвристических подходах. Исследование направлено на повышение качества решения этой задачи за счет применения методов машинного обучения.

В качестве первого, наиболее простого подхода, мы решали задачу в формулировке аппроксимации относительной влажности по данным сопутствующих измерений атмосферного давления, температуры воздуха, скорости и направления ветра, температуры поверхности океана, а также наблюдений количества и типов облачности на трех ярусах. Кроме этого, в составе сопутствующих переменных используется код погоды по стандарту ВМО и расчетная высота солнца. В исследовании использованы модели машинного обучения следующих типов: линейная регрессия, дерево решений [3], случайный лес [2], градиентный бустинг [5] и полносвязная искусственная нейронная сеть [8]. Для повышения территориальной и временной специфичности разрабатываемых моделей мы провели исследование для каждого 2-градусного квадрата и каждого сезона по отдельности. Для обучения и применения моделей машинного обучения мы применяли библиотеку scikit-learn и пакет, реализующий модель CatBoost [5]. Для каждого вида модели мы оптимизировали гиперпараметры с применением библиотеки байесовской оптимизации Optuna [9]. На основе полученных результатов были построены карты пространственного распределения ошибок моделей, которые позволили выявить регионы с высокой и низкой точностью аппроксимации влажности.

Результаты работы подтверждают эффективность алгоритмов машинного обучения для восстановления данных относительной влажности по сопутствующим измерениям и наблюдениям и позволяют повысить точность восстановления климатических рядов по сравнению с тривиальными подходами, например, усреднением величины по регионам.

Анализ пространственных карт ошибок моделей показал, что наибольшие ошибки аппроксимации наблюдаются в сезоны и в регионах с высокой изменчивостью влажности (карты разброса значений относительной влажности показаны на рис. 1). Примененный в работе подход, основанный на разделении данных по квадратам и сезонам, позволил учесть региональные особенности и сезонные колебания влажности, а также снизил ошибки аппроксимации.



Рис. 1*.* Карты разброса значений относительной влажности по сезонам

Анализ неопределенностей меры качества и целевой переменной, проведенный с использованием подхода бутстреп, позволил оценить надежность полученных результатов и выявить ключевые области для дальнейшего улучшения результатов. Неопределенность меры качества оказалась существенно низкой для всех моделей. Тривиальная модель, линейная регрессия и Catboost показали наименьшую неопределенность целевой переменной.

Сравнительный анализ моделей выявил, что Catboost и случайный лес продемонстрировали наилучшие результаты в данной работе, что делает их лучшими моделями для аппроксимации влажности. Catboost показал наивысшее среднее качество на отложенной выборке, а случайный лес показал наилучшее среднее качество по bootstrap-выборкам. Модели демонстрируют наиболее стабильные результаты, с низкими значениями неопределенности как меры качества, так и целевой переменной, что подтверждает их способность к точному восстановлению относительной влажности.

Исследование выполнено в рамках Соглашения № 075-03-2024-117 от 17.01.2024 Московского физико-технического института.

**Литература**

1. *Александрова М. П., Гулев С. К.* Реконструкция долгопериодной изменчивости облачности и радиационных потоков над мировым океаном: Дис. на канд. физико-математических наук: 1.6.17. М.: ИО РАН, 2023.
2. *Adele Cutler, D. Richard Cutler John R. Stevens* Random Forests // J. Springer. Ensemble Machine Learning. 2012. DOI: 10.1007/978-1-4419-9326-7\_5.
3. *Barry de Ville* Decision trees // J. WIREs Computational Statistics: Volume 5, Issue 6. 2013. DOI: 10.1002/wics.1278.
4. Eric Freeman, Scott D. Woodruff, Steven J. Worley, Sandra J. Lubker, Elizabeth C. Kent, William E. Angel, David I. Berry, Philip Brohan, Ryan Eastman, Lydia Gates, Wolfgang Gloeden, Zaihua Ji, Jay Lawrimore, Nick A. Rayner, Gudrun Rosenhagen, Shawn R. Smith ICOADS Release 3.0: a major update to the historical marine climate record  *International Journal of Climatology.* 2017. № 5 (37). DOI: 10.1002/joc.4775.
5. *Hancock, J.T., Khoshgoftaar, T.M.* CatBoost for big data: an interdisciplinary review // J. Big Data 7. 2020. DOI: 10.1186/s40537-020-00369-8.
6. *Jiajun Guo, Liang Zhang, Ruqiang Guo* Relative humidity prediction with covariates and error correction based on SARIMA-EG-ECM model // J. Springer. Modeling Earth Systems and Environment. 2023. DOI: 10.1007/s40808-023-01738-x.
7. Luminda Niroshana Gunawardhana, Ghazi A. Al-Rawas, So Kazama An alternative method for predicting relative humidity for climate change studies // J. Meteorological application. 2017. DOI: 10.1002/met.1641.
8. *Marvin Minsky, Seymour Papert* Perceptrons: An Introduction to Computational Geometry// The MIT Press. 1969.
9. *Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Kayama* Optuna: A Next-generation Hyperparameter Optimization Framework // Conference: the 25th ACM SIGKDD International Conference. 2019. DOI: 10.48550/arXiv.1907.10902.