

Google Earth Engine and machine learning for Earth monitoring

Alexander Uzhinskiy *

Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia

E-mail: auzhinskiy@jinr.ru

Hyperspectral images are a unique source for obtaining many kinds of information about the Earth's surface. Modern platforms support users to perform complex analyses with a collection of images without the use of any specialized software. Google Earth Engine (GEE) is a planetary-scale platform for Earth science data & analysis. Atmospheric, radiometric and geometric corrections have been made on number of image collections at GEE. While working with raw data, it is possible to use build-in GEE function to filter data and create composites to get cloud score threshold and the percentile. It is also possible to use custom algorithms for atmospheric corrections. There are over 200 satellite image collections and modeled datasets. Some collections have a spatial resolution of up to 10 meters. GEE has the JavaScript online editor to create and verify code and Python API for advanced applications. All that made GEE very convenient tool for different Earth monitoring projects. Over the last decades there has been considerable progress in developing a machine learning methodology for a variety of Earth Science applications involving trace gases, retrievals, aerosol products, land surface products, vegetation indices, fire and flood tracking, ocean applications, and many others. In this report, we will review basic GEE functions and practice, some examples of successful applications, and our experience in environmental monitoring.

*The 6th International Workshop on Deep Learning in Computational Physics (DLCP2022)
6-8 July 2022
JINR, Dubna, Russia*

Google Earth Engine и машинное обучение для мониторинга Земли

Александр Ужинский*

Объединенный институт ядерных исследований, Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6

E-mail: auzhinskiy@jinr.ru

Гиперспектральные спутниковые снимки – уникальные источник разнообразной информации о поверхности Земли. Раньше для работы с ними требовались профильные знания и навыки, но сейчас современные платформы предоставляют пользователям возможность выполнения различных операций и проведения анализа коллекций изображений спутниковых программ с высокой эффективностью и без использования специализированного программного обеспечения. Google Earth Engine (GEE) – платформа, позволяющая работать с сотнями различных спутниковых программ и наборов данных. GEE содержит как исходные данные, так и обработанные – для которых были произведены атмосферные, радиометрические и геометрические коррекции. В GEE есть онлайн JavaScript редактор, удобный для проведения анализа в режиме реального времени, и программный интерфейс Python, позволяющий использовать возможности платформы в высоконагруженных проектах. Все это делает GEE удобным и эффективным инструментом для мониторинга Земли. Данные с космоснимков зачастую используются совместно с различными методами машинного обучения для реализации экологических проектов, например, отслеживания загрязнения воды, почвы и воздуха, анализа последствий и прогнозирования лесных пожаров и наводнений, контроля изменений в рельефе и использовании почв. В данной работе будет представлена общая информация о Google Earth Engine, приведены примеры наиболее удачных приложений, реализованных на ее основе, а также наш опыт использования данной платформы.

*The 6th International Workshop on Deep Learning in Computational Physics (DLCP2022)
6-8 July 2022
JINR, Dubna, Russia*

*Speaker

1. Введение

Наша планета на протяжении всего своего существования сталкивается с различными вызовами, однако в последние десятилетия значительно увеличилось количество проблем, связанных с деятельностью человека. Глобальные изменения климата, кислотные дожди, загрязнение воды, воздуха, почвы, огромные острова мусора, вырубка лесов, разрастание городских агломераций – это далеко не полный их список. Одним из инструментов отслеживания экологической ситуации на планете являются различные спутниковые программы. В настоящее время на орбите Земли находятся сотни спутников, фиксирующих различные параметры состояния поверхности. Некоторые программы существуют более 40 лет, что предоставляет уникальные возможности для исторического анализа. Если первые спутники осуществляли съемку в основном в видимых спектрах (красный, зеленый, синий), то широта возможностей современных продуктов просто поражает воображение. Существуют специализированные программы, направленные на решение специфических задач. Например, Sentinel5 [1], в основном ориентированная на глубокий анализ воздуха, осуществляет съемку в 7 спектральных каналах, что позволяет получить представление о концентрации CO, NO₂, SO₂ и прочих летучих соединений. Технические и функциональные возможности спутниковых программ сильно разнятся. Общедоступные могут иметь разрешение от тысяч до нескольких метров и период обновления данных от нескольких недель до нескольких дней. Коммерческие программы, например www.planet.com, предоставляют данные ежедневно с разрешением до полуметра.

Ранее при реализации проектов с использованием данных дистанционного зондирования земли требовалось знание специализированного программного обеспечения. Сама подготовка данных занимала значительно больше времени, чем непосредственно анализ. Например, для решения задачи отслеживания изменения лесного массива в определенном районе необходимо было произвести целый ряд операций, требующих профильных навыков и значительных временных и вычислительных затрат. Нужно было определиться с исследуемой областью и загрузить из открытых источников снимки подходящей программы, например Landsat [2]. Из-за наличия облачности пришлось бы загрузить всю коллекцию для создания композитных изображений. В результате необходимо было бы обработать более 1ТВ данных, только чтобы получить качественные изображения области размером со средний территориальный центр. Затем пришлось бы произвести атмосферную коррекцию, нормализацию и прочие процедуры по подготовке изображения. В конце можно было бы вычислить индекс вегетации, например, NDVI, и проанализировать его изменение во времени и пространстве. Подобная задача могла занимать месяцы, особенно с учетом того, что выполнялась бы на персональном компьютере. Используя возможности, которые предоставляют современные платформы, можно решить подобную задачу за час, а визуализация результата займет несколько секунд.

Google Earth Engine (GEE) [3] – платформа для анализа данных в планетарном масштабе. Ее запуск был анонсирован в 2012 году и приурочен к 40-летию миссии Landsat. GEE предоставляет для ученых и исследователей свободный доступ к мощностям Google и более чем 60Pb снимков различных программ и геопространственных данных. Для некоммерческих проектов доступ к платформе бесплатный. Наиболее удобным интерфейсом для знакомства с возможностями платформы является online JavaScript

редактор. Он позволяет не только запускать код и отображать карты и объекты на них, но и создавать различные графики, диаграммы, таблицы, анимированные изображения и многое другое. Для реализации проектов, требующих серьезных вычислений и манипуляций с данными, существует программный интерфейс Python. Платформа существует уже 10 лет и имеет исчерпывающую документацию, охватывающую все аспекты работы с ней, большое количество примеров реализации различных задач, обучающие курсы для пользователей с разным уровнем подготовки и развитое сообщество.

В каталоге данных GEE представлены как растровые данные спутниковых программ типа Landsat, Sentinel, Modis, NOVA и прочие, так и векторные данные, позволяющие получить, например, информацию по границам административных областей или составу и количеству населения. Кроме этого, в каталоге содержится информация по топологии поверхности земли, климатические и погодные данные. В общей сложности более 200 программ и программных продуктов доступны для анализа. На рисунке 1 приведены примеры изображений программ, доступных в GEE.

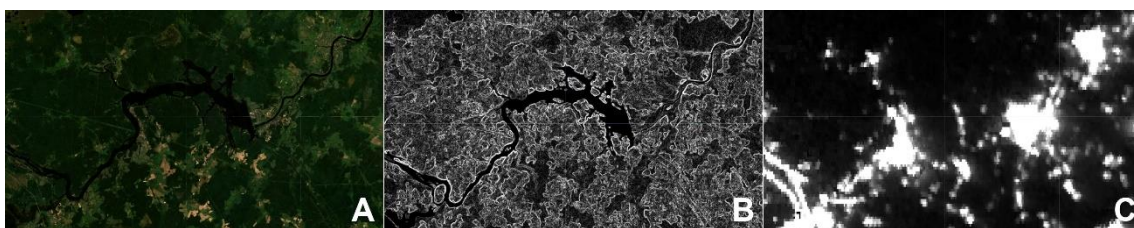


Рис. 1. Примеры снимков программ в GEE. А - USGS Landsat 8 Level 2. В - SRTM Digital Elevation Data Version 4. С - VIIRS Nighttime Day/Night Band Composites Version 1.

В GEE существует множество различных объектов и методов для работы с ними. Image – представляет собой набор растровых данных, привязанных к одному временному периоду. Image collection – набор одиночных изображений, над которыми можно производить различные манипуляции: фильтрацию, сортировку, агрегацию и прочее. Feature и feature collection – аналог предыдущих объектов, но содержат векторные данные. Reducer – применяется к объектам для агрегации данных. Например, можно получить медианное значение пикселей в определенной области изображения или среднее значение пикселей коллекции изображений за интересующий временной период см. рис 2. Reducer – один из основных инструментов при решении задач классификации. Machine Learning – алгоритмы машинного обучения, доступные для использования на наборах данных. Более 1000 различных типов данных, операторов и функций доступны в GEE и подробно описаны в документации платформы.

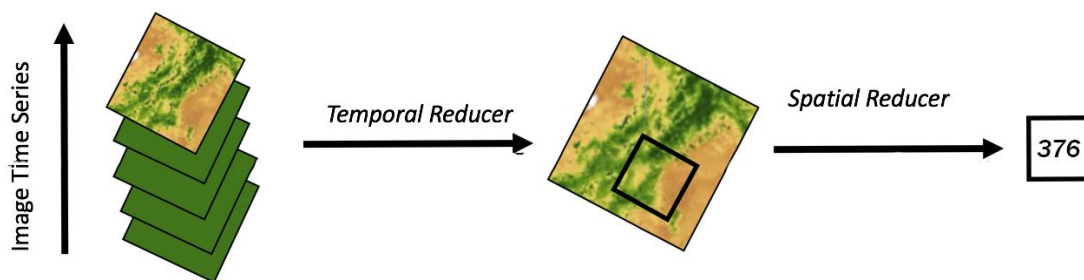


Рис. 2. Схема агрегации данных средствами GEE reducer во времени и пространстве.

Возможность создания изображений без облаков для Google Earth - одна из мотиваций появления GEE. Спутник не может зависнуть над определенной областью и ждать, пока пройдут облака, но при наличии достаточно большого количества наблюдений можно программно получить изображение без облаков и теней. В GEE существует несколько методов достижения положительных результатов в зависимости от используемых спутниковых программ: фильтрация данных, композитные изображения, попиксельные шаблоны замены и прочее. Если анализируемый период превышает несколько месяцев, то GEE позволяет получить отличные изображения без облаков см. рис. 3.



Рис. 3. Пример изображения программы USGS Landsat 8 Level 2. Слева – исходные данные. Справа – после применения функции попиксельной замены на основе канала с данными по облачности.

Подводя промежуточный итог, можно заключить, что GEE представляет собой удобный и эффективный инструмент для анализа геопространственных данных, который можно использовать для реализации различных проектов, связанных с мониторингом Земли.

2. Интересные проекты

GEE используется для отслеживания множества проблем: исчезновение лесов, деградация почв, таяние ледников, эпидемии, продовольственная безопасность, управление водными ресурсами, изменения климата, исчезновение различных видов живых существ и многих других. Рассмотрим наиболее интересные из них.

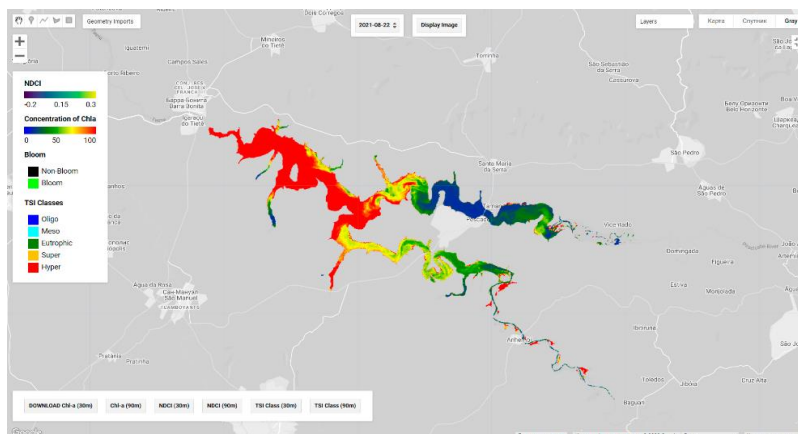


Рис 4. AlgaeMAP – состояние воды в реке Тиете в районе Барра Бонита, Бразилия, Август 2021

AlgaeMAP [4] – проект осуществляет контроль качества воды в Латинской Америки. В его рамках GEE используется как источник данных для моделей и инструмент

отображения информации по концентрации определенных типов водорослей в водоемах. Снимки программы Sentinel-2 применяют для вычисления NDCI (Normalized Difference Chlorophyll Index), который затем используется при решении задачи классификации. Любой желающий может получить как актуальную информацию, так и исторические сводки с 2015 года. Пример интерфейса приложения приведен на рисунке 4.

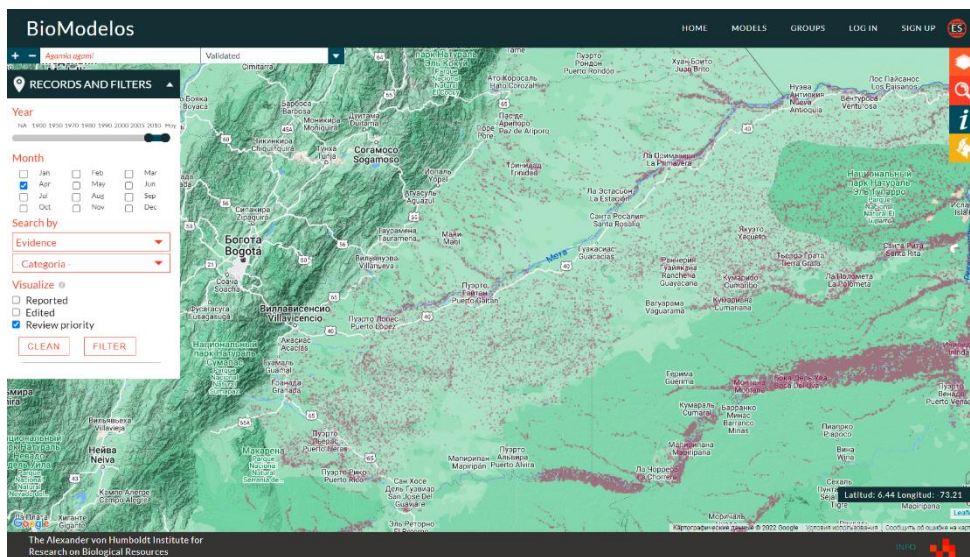


Рис 5. BioModelos – распространение вида Agamia Agami в Колумбии.

BioModelos [5] – система отслеживания биоразнообразия в Колумбии. Участники программы фиксируют случаи наблюдения различных представителей животного и растительного мира. Затем эта информация, совместно с данными по топологии и метриками, полученными с космоснимков, используется в специализированных моделях для составления прогноза по возможным ареалам обитания см. рис. 5. Подобные проекты реализуются и в Северной Америке [6].

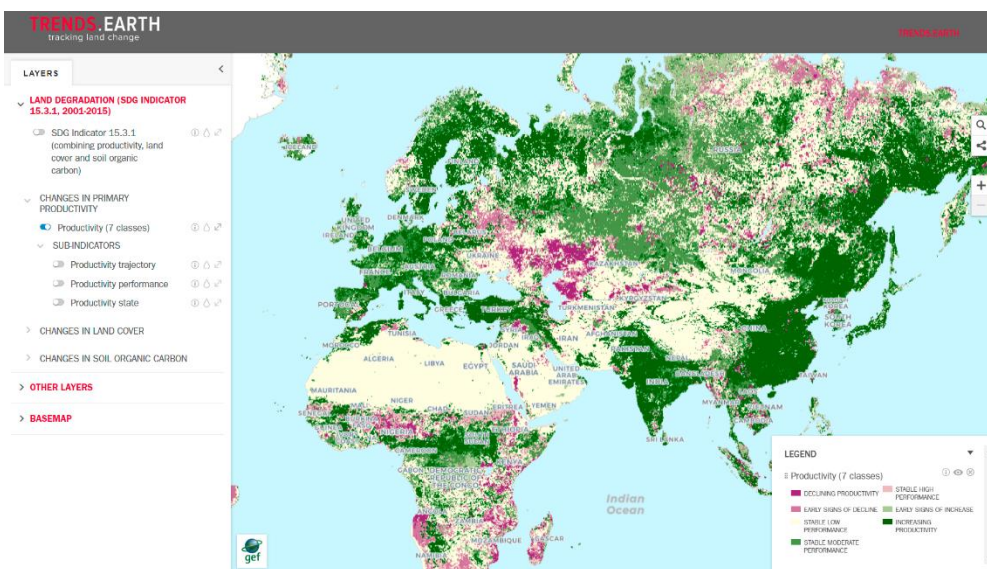


Рис. 6. Trends.Earth – классификация земли по типам использования

POS (DLCSP2022)021

Trends.Earth [7] – платформа контроля изменения статуса использования земель. Данные программ Landsat и Sentinel применяются совместно с алгоритмами машинного обучения для решения задач классификации. Пример интерфейса приведен на рисунке 6. Классификация по типам землепользования - популярное направление исследований [8, 9].

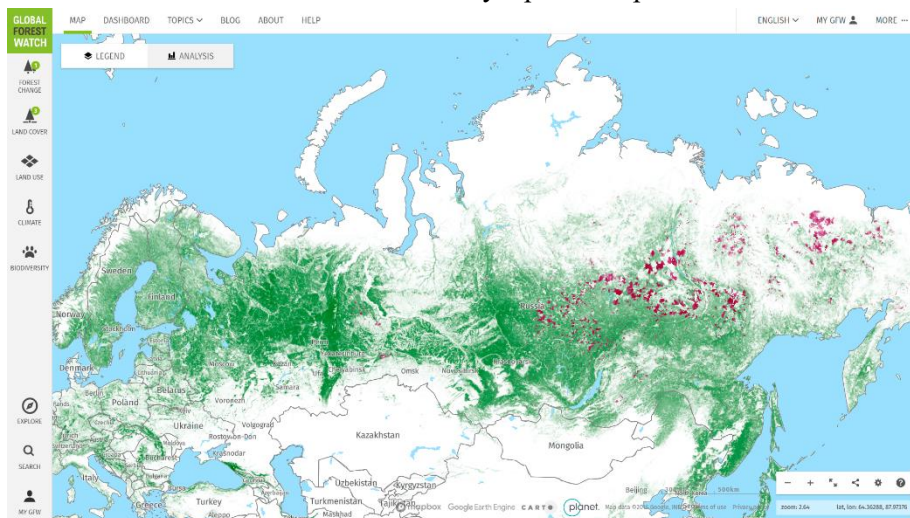


Рис 7. Интерфейс портала Global Forest Watch, на котором отображены потери в лесном массиве из-за пожаров за 2021 год.

Global Forest Watch [10] — портал для отслеживания состояния лесного покрова в планетарных масштабах. Данные с космоснимков и алгоритмы машинного обучения используются для мониторинга и управления лесами, выявления незаконной вырубке, обнаружения пожаров и в целом обеспечения устойчивого развития. Пример интерфейса приведен на рис. 7. Оценка последствий, отслеживание и предсказание пожаров – само по себе является весьма востребованным направлением исследований, в котором достигнуты многообещающие результаты [11, 12].

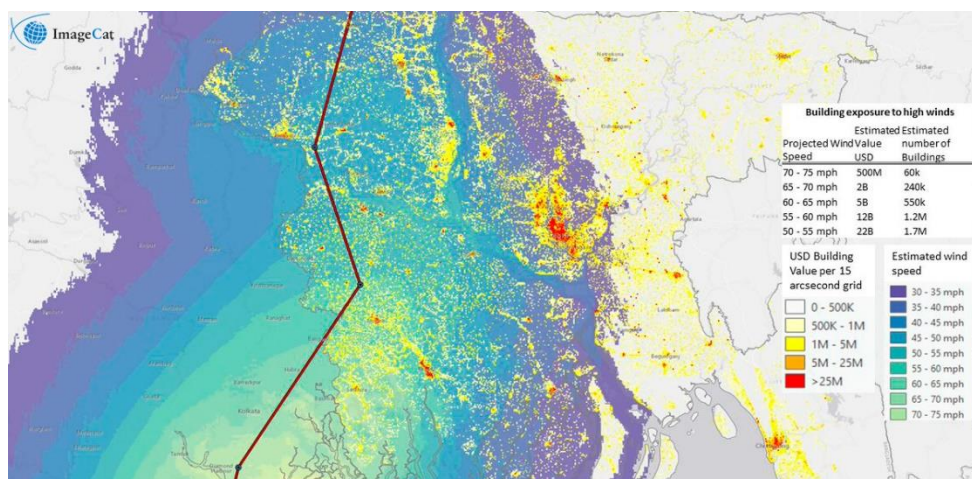


Рис 8. Данные программы ImageCat по зданиям, подвергшемся воздействию ураганного ветра из-за тропического циклона Амphan в Бангладеш.

ImageCat – платформа оценки рисков природных и техногенных катаклизмов и угроз. Проект существует уже много лет и использует различные данные, в том числе и снимки спутниковых программ совместно с различными статистическими и нейросетевыми

моделями см. рис. 8. Оценка рисков – еще одна популярная область исследований, включающая в себя множество различных направлений, например наводнения [13], таяние ледников [14], сейсмоактивность [15], эпидемии [16] и прочие. Ведутся проекты с применением GEE и в Объединенном институте ядерных исследований.

3. Наш опыт использования GEE

Мы используем GEE в рамках интеллектуальной платформы экологического мониторинга программы комиссии ООН по дальнему трансграничному переносу воздушных загрязнений ICP Vegetation [17]. Программа направлена на определение наиболее неблагополучных областей, создание региональных карт и улучшение понимания природы загрязнений. ICP Vegetation объединяет сотни ученых, которые проводят исследования в сорока трех странах Европы и Азии [18]. Участники программы собирают образцы мха и используют различные техники, например нейтронно-активационный анализ, чтобы получить данные по содержанию тяжелых металлов, азота, стойких органических соединений и радионуклидов. Разработка системы управления данными программы началась в 2016 году в Объединенном институте ядерных исследований. Проект, призванный автоматизировать типовые операции с данными и создание региональных карт, вбирал в себя все новые и новые подходы и технологии. В том числе был реализован механизм прогнозирования, основанный на применении машинного обучения совместно с данными дистанционного зондирования земли. Поскольку сбор данных участниками программы ограничен временными и географическими рамками, прогнозирование является важным элементом исследования, позволяющим заполнить пробелы в данных.

После анализа образцов у нас имеется информация о концентрации различных элементов в местах сбора. Их координаты мы можем использовать, чтобы получать индексы различных спутниковых программ. Индекс включает в себя название спутниковой программы, данные которой используются, размер анализируемой площади, идентификатор спектрального канала, в которой осуществлялся съемка, и математической функции, применяемой к цифровой матрице полученного изображения. Мы используем программный интерфейс Python для работы с GEE, т.к. на первом этапе производится поиск наиболее перспективных индексов основываясь на коэффициенте корреляции Спирмена и данный процесс может занимать достаточно много времени. Используется более 40 различных коллекций, несколько вариантов функций агрегации и размеров анализируемых площадей. Наиболее перспективные индексы совместно с данными о концентрации металлов в точках сбора образцов используются для обучения моделей. Затем вычисляются индексы, покрывающие интересующую площадь сеткой с требуемой размерностью. Эти данные подаются в модель, и строится прогноз. В текущей реализации в зависимости от количества исходных данных используются статистические модели машинного обучения либо глубокие нейронные сети. При обучении на вход модели, кроме концентрации, могут подаваться как все индексы, так и определённый набор неколлинеарных индексов с высокой корреляцией с концентрацией металла. Решаются задачи регрессии и классификации, но последние более приоритетны в силу ряда причин. Во-первых, модельные данные в основном необходимы для построения карт, в которых изначально заложена градация уровней загрязнения. Во-вторых, критерии оценки точности при

решении задачи классификации прозрачнее, чем при решении задачи регрессии. В-третьих, точек отбора проб, в которых присутствует высокий уровень загрязнения, на порядок меньше, чем точек с нормальным уровнем, и при использовании методов балансировки обучающей выборки можно добиться лучших результатов. Подобный подход позволяет получить для некоторых элементов прогноз с точностью более 90%. Были достигнуты положительные результаты на региональном уровне, например, построены прогнозы загрязнения сурьмой в Норвегии, магнием в Сербии, ураном в Румынии [19]. На локальном уровне была создана модель загрязнения Белграда медью в летний и зимний период 2013-2014 [20]. В последнем исследовании [21] были использованы данные по концентрации тяжелых металлов в 281 точке в Московской, Владимирской и Ярославской областях, полученные в 2018-2019 годах. Данные с космоснимков использовались в качестве дополнительных данных при обучении моделей и в качестве основных для построения прогноза. Были опробованы различные модели: градиентный бустинг, многоуровневой перцептрон, сиамские нейронные сети. В результате для 9 элементов точность моделей превысила 89%. Для сурьмы точность модели, основанной на сиамских нейронных сетях, составило 93%. Прогнозы модели для центрального региона России на 2020 год (рис. 9) и на 2019 год в высоком разрешении были оценены экспертами в области мониторинга качества воздуха как реалистичные.

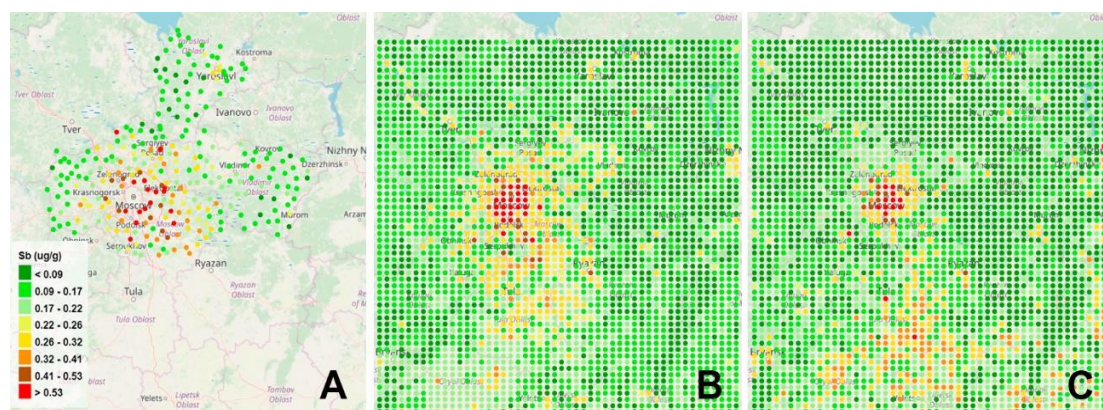


Рис. 9. А – концентрация сурьмы в центральном регионе России в 2019 году (реальные данные). В – прогноз модели на базе данных с космоснимков на 2019 год. С – прогноз модели на 2020 год.

В настоящее время ведутся работы по созданию глобальной модели загрязнения для европейской части Евразийского континента см. рис. 10.

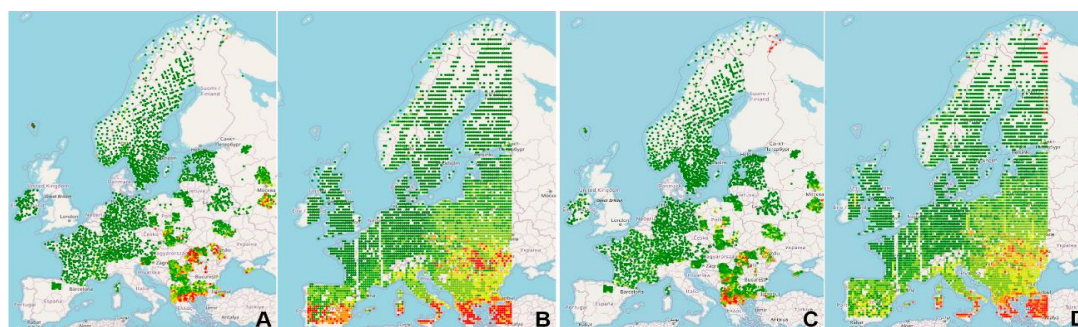


Рис. 10. Промежуточные результаты работы глобальной модели на данных атласа 2015-2016 гг. А – Fe – реальные данные. В - Fe – прогноз модели. С – Ni – реальные данные. D - Ni – прогноз модели.

Благодаря GEE и машинному обучению мы можем отслеживать изменения в ситуации с загрязнением воздуха намного оперативнее, получать детальную информацию в интересующих областях и в областях, где сбор образцов невозможен, а в перспективе даже частично автоматизировать процесс экологического мониторинга.

4. Заключение

Современные спутниковые программы — это великолепный источник данных для мониторинга состояния земли, платформа Google Earth Engine предоставляет удобные механизмы для работы с ними, а алгоритмы машинного обучения позволяют строить прогнозы и решать задачи классификации и кластеризации. Множество реализованных проектов и наш собственный опыт показывают, что использование GEE способно вывести на новый уровень исследования в области отслеживания загрязнения воды, почвы и воздуха, анализа последствий и прогнозирования лесных пожаров и наводнений, контроля изменений в рельефе и использовании почв и многих других. Дальнейшее развитие спутниковых программ, технологий обработки их данных и методов машинного обучения способствует исправлению общей экологической ситуации и появлению новых уникальных возможностей в области мониторинга Земли.

References

- [1] ESA. *Sentinel-5 Precursor: ESA's Atmospheric Chemistry and Pollution Monitoring Mission* // ESA Rep. — 2016 — SP-1332. — 86 pp., <https://esamultimedia.esa.int/multimedia/publications/SP-1332/SP-1332.pdf>.
- [2] Butcher G., Barnes C., Owen L. *Landsat: The cornerstone of global land imaging* // GIM International Magazine — Jan. 2019 — Vol. 1 — P. 31-35.
- [3] Gorelick N., Hancher M, Dixon M, Ilyushchenko S., Thau D., Moore R. *Google Earth Engine: Planetary-scale geospatial analysis for everyone* // Remote Sensing of Environment. — 2017. — Vol. 202. — P. 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- [4] Lobo, F.d.L.; Nagel, G.W.; Maciel, D.A.; Carvalho, L.A.S.d.; Martins, V.S.; Barbosa, C.C.F.; Novo, E.M.L.d.M. *AlgaeMap: Algae Bloom Monitoring Application for Inland Waters in Latin America.* // Remote Sens. — 2021. — Vol. 13. — 2874. <https://doi.org/10.3390/rs13152874>
- [5] Velásquez-Tibatá J, Olaya-Rodríguez MH, López-Lozano D, Gutiérrez C, González I, Londoño-Murcia MC. *BioModelos: A collaborative online system to map species distributions* // PLoS ONE — 2019 — 14(3): e0214522. <https://doi.org/10.1371/journal.pone.0214522>
- [6] Crego R.D., Jared A. Stabach J.A., Connette G. *Implementation of species distribution models in Google Earth Engine* // Diversity & distributions — 2022 — 28, no. 5 — P. 904-916. doi: 10.1111/ddi.13491
- [7] Gonzalez-Roglich, M., Zvoleff, A., Noon, M., Liniger, H., Fleiner, R., Harari, N., Garcia, C. *Synergizing global tools to monitor progress towards land degradation neutrality: Trends.Earth and the World Overview of Conservation Approaches and Technologies sustainable land management database* // Environ. Sci. Policy — 2019 — V 93, — P. 34–42. <https://doi.org/10.1016/j.envsci.2018.12.019>
- [8] Qu L, Li M., Chen Z., Liu W., Zhi J, Zhang L. *Mapping large area tea plantations using progressive random forest and Google Earth Engine* // Journal of Applied Remote Sensing — May 2022 — 16(2), 024509 <https://doi.org/10.1117/1.JRS.16.024509>

- [9] Novais, J.J.; Lacerda, M.P.C.; Sano, E.E.; Demattê, J.A.M.; Oliveira, M.P., Jr. *Digital Soil Mapping Using Multispectral Modeling with Landsat Time Series Cloud Computing Based* // Remote Sens. — 2021 — 13, 1181. <https://doi.org/10.3390/rs13061181>
- [10] Shea K. *Measuring the Impact of Monitoring: How We Know Transparent Near-Real-Time Data Can Help Save the Forests* // In: Uitto, J.I., Batra, G. (eds) Transformational Change for People and the Planet. Sustainable Development Goals Series. Springer Cham — 2022. https://doi.org/10.1007/978-3-030-78853-7_18
- [11] Tavakkoli Piralilou, S.; Einali, G.; Ghorbanzadeh, O.; Nachappa, T.G.; Gholamnia, K.; Blaschke, T.; Ghamisi, P. *A Google Earth Engine Approach for Wildfire Susceptibility Prediction Fusion with Remote Sensing Data of Different Spatial Resolutions* // Remote Sens. — 2022 — 14, 672. <https://doi.org/10.3390/rs14030672>
- [12] Ban, Y., Zhang, P., Nascetti, A., Bevington A.R., Wulder M.A. *Near Real-Time Wildfire Progression Monitoring with Sentinel-1 SAR Time Series and Deep Learning* // Sci Rep. — 2020 — V. 10, 1322. <https://doi.org/10.1038/s41598-019-56967-x>
- [13] Pandey, A.C.; Kaushik, K.; Parida, B.R. *Google Earth Engine for Large-Scale Flood Mapping Using SAR Data and Impact Assessment on Agriculture and Population of Ganga-Brahmaputra Basin* // Sustainability — 2022. — 14, 4210
- [14] Liang D, Guo H, Zhang L, Cheng Y, Zhu Q, Liu X. *Time-series snowmelt detection over the Antarctic using Sentinel-1 SAR images on Google Earth Engine* // Remote Sensing of Environment — 2021. — Vol. 256, 112318. <https://doi.org/10.1016/j.rse.2021.112318>.
- [15] Akhoondzadeh M. *Advances in Seismo-LAI anomalies detection within Google Earth Engine (GEE) cloud platform* // Advances in Space Research — 2022 — Vol. 69, Is. 12. — P. 4351–4357. <https://doi.org/10.1016/j.asr.2022.03.033>.
- [16] Zarro, C.; Cerra, D.; Auer, S.; Ullo, S.L.; Reinartz, P. *Urban Sprawl and COVID-19 Impact Analysis by Integrating Deep Learning with Google Earth Engine* // Remote Sens. — 2022 —, 14, 2038. <https://doi.org/10.3390/rs14092038>
- [17] Uzhinskiy A. *Intelligent environmental monitoring platform* // CEUR Workshop Proceedings — 2021. — 2267 — P. 351- 358. <https://doi.org/10.54546/MLIT.2021.63.35.001>
- [18] Harmens H., Norris D.A., Steinnes E., Kubin E., Piispanen J., Alber R., Aleksiyenak Y., Blum O., Coskun M., Dam M., De Temmerman L., Fernández J.A., Frolova M., Frontasyeva M., González Miqueo L., Grodzinska K., Jeran Z., Korzekwa S., Krmar M., Kvietskus K., Leblond S., Liiv, S., Magnússon S.H., Mankovská B., Pesch R., Rühling Å., Santamaria J.M., Schröder W., Spiric Z., Suchara I., Thöni L., Urumov V., Yurukova L. & Zechmeister H.G. *Mosses as biomonitors of atmospheric heavy metal deposition: spatial patterns and temporal trends in Europe* // Environ Pollut — 2010 — 158, 3144–3156.
- [19] Uzhinskiy A., Ososkov G., Goncharov P., Frontsyeva M. *Combining satellite imagery and machine learning to predict atmospheric heavy metal contamination* // CEUR Workshop Proceedings — 2018 — 2267 — P. 351- 358.
- [20] Uzhinskiy A., Aničić Urošević M., Frontasyeva M. *Prediction of air pollution by potentially toxic elements over urban area by combining satellite imagery, Moss Biomonitoring Data and Machine Learning* // Ciencia e Tecnica Vitivinicola Journal — 2020. — Vol. 35, No. 12. ISSN:2416-3953.
- [21] Alexander Uzhinskiy, Konstantin Vergel *Central Russia heavy metal contamination model based on satellite imagery and machine learning* // Computer Optics, 47-1