# FORECASTING THE DISTRIBUTION OF DISEASES IN TROPICAL ZONES USING MACHINE LEARNING METHODS

*Alexey A. Kolesnikov*[(1)], *Pavel M. Kikin*[(2)]

[(1)]Siberian State University of Geosystems and Technologies, Novosibirsk

[(2)]Peter the Great St. Petersburg Polytechnic University, St. Petersburg

**Abstract:** Infection with tropical parasitic diseases, according to WHO, has a huge impact on the health of more than 40 million people worldwide and is the second leading cause of immunodeficiency. The number of infections is influenced by many factors - climatic, demographic, vegetation cover and a number of others. The article presents a study and an assessment of the degree of influence of each of these factors, as well as a comparison of the quality of forecasting by separate methods of geo-informational analysis and machine learning and the possibility of their ensemble.

*Keywords: tropical diseases, fever, machine learning, geostatistics, neural networks.*

# ПРОГНОЗИРОВАНИЕ РАСПРОСТРАНЕНИЯ БОЛЕЗНЕЙ В ТРОПИЧЕСКИХ ЗОНАХ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

*Колесников А.А.*[(1)], *Кикин П.М.*[(2)]

[(1)] Сибирский государственный университет геосистем и технологий, г. Новосибирск

[(2)] Санкт-Петербургский Политехнический Университет Петра Великого, г. Санкт-Петербург

Заражение тропическими паразитарными болезнями, по данным ВОЗ, оказывают огромное влияние на здоровье более 40 миллионов человек во всем мире и являются второй по значимости причиной иммунодефицита. На количество заражений влияют многие факторы - климатические, демографические, растительный покров и ряд других. В статье представлено исследование и оценка степени влияния каждого из этих факторов, а также сравнение качества прогнозирования отдельными методами геоинформационного анализа и машинного обучения и возможности их ансамблирования.

*Ключевые слова: тропические болезни, лихорадка, машинное обучение, геостатистика, нейронные сети.*

**Введение.** Многие важные научные задачи связанны с обработкой данных, которые изменяются в пространстве и времени. В качестве примеров можно привести климатические, экологические, сейсмические исследования [1-3]. Корректно подобранные методы и алгоритмы моделирования и прогнозирования позволяют достоверно оценить тенденцию развития исследуемых показателей произвольных объектов и явлений. Автоматизировать процесс выбора алгоритмов, формул и их параметров способны технологии машинного обучения [4-6]. Формирование путей взаимодействия между технологиями геоинформационных систем и машинного обучения для пространственно-временного прогнозирования позволит разработать наиболее оптимальные решения для эффективного анализа и управления процессами и явлениями. Важно иметь информацию о методах и алгоритмах, которые хорошо работают на практике для обработки и прогнозирования данных максимально широкого диапазона пространственно-временных процессов [7-8]. Значительный прирост в качестве построения математических моделей пространственно-временных процессов дало развитие концепции Deep Learning, в частности, различных вариантов рекуррентных нейронных сетей, например, LSTM. Также широко распространенные сверточные нейронные сети (CNN) используют слои со сверточными фильтрами для извлечения локальных объектов посредством скользящего окна и может моделировать близлежащие или долгосрочные пространственные зависимости (архитектура SRCN) [9]. Для оценки применимости, точности получаемых математических моделей, универсальности методов машинного обучения для решения задач геоинформатики была выбрана задача прогнозирования распространения тропической болезни денге по данным ДЗЗ, представленных в виде временных рядов [10]. Денге - это инфекционное заболевание, передающееся от человека к человеку посредством москитов Aedes aegypti и Aedes albopictus, которые является основным переносчиком вируса в различных частях земного шара. По оценкам Всемирной организации здравоохранения (ВОЗ), ежегодно во всем мире регистрируется около 50-100 миллионов случаев заболевания лихорадкой денге, а две пятых населения мира подвержены риску, а денге или DHF / DSS затронули более ста стран. эпидемии. С 1950 года было зарегистрировано более 500 000 случаев госпитализации и около 70 000 случаев смерти детей; уровень заболеваемости среди детей достигает 64 на

1000 человек населения. Согласно анализу глобального распространения вируса денге, число инфекций в год оценивается в 390 миллионов, из которых почти 96 миллионов являются симптоматическими. По оценкам, число инфекций денге резко возросло за последние 50 лет, что привело к огромному воздействию на здоровье человека во всем мире. Регионы распространения включают страны Юго-Восточной Азии, Латинской Америки, Африки, где лихорадка денге была гиперэндемической в течение десятилетий и представляет собой серьезную проблему [11–14].

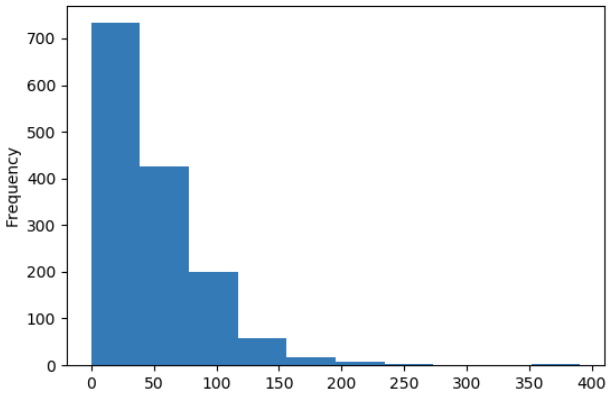
Основная цель этого исследования - изучить возможность и оценить качество результатов применения методов машинного обучения для анализа и прогнозирования болезни денге на основе измеренных параметров ДЗЗ, картографических данных и связанных статистических показателей.

**Методы и материалы.**

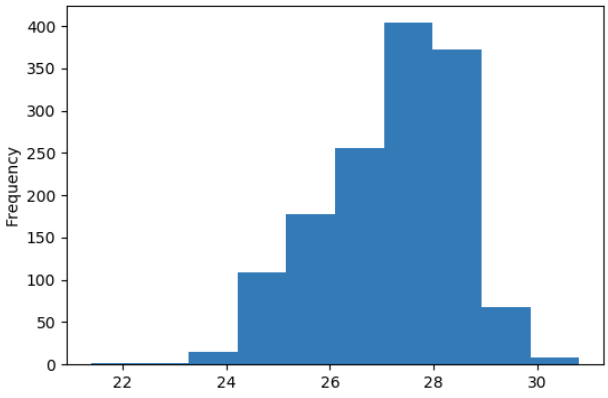
Исходной информацией для обучения алгоритмов являлись данные об окружающей среде собранные Центром по контролю и профилактике заболеваний (https://www.cdc.gov/Dengue), Национальным управлением по вопросам океана и атмосферы в Министерстве торговли США (http://www.healthmap.org/dengue/en), Министерством здравоохранения Филиппин (https://www.doh.gov.ph).

В качестве метрики оценки точности была использована среднеквадратическая ошибка, для ее расчета и демонстрации качества моделей на основе общего рейтинга использовался сервис drivendata.org. Список параметров включал в себя следующие категории исходных данных: аббревиатуры городов, дату измерений, текущие климатические показатели и их прогноз (температуру, влажность, количество осадков) по данным NOAA's GHCN, PERSIANN, NOAA's NCEP, значения индекса NDVI рассчитанные для прилегающих к городу пикселей спутникового снимка [15-17].

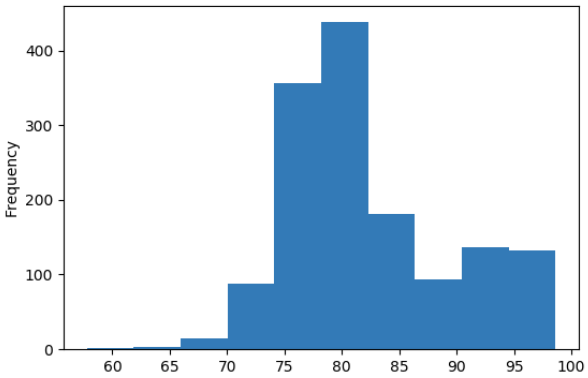
Визуальное представление ряда климатических параметров и временного охвата имеющихся данных приведено на рисунке 1.



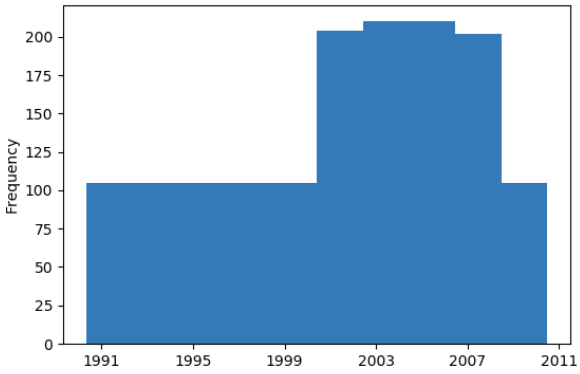

Среднее значение осадков (в мм)




Средняя температура




Относительная влажность




Временной интервал набора данных

Рис.1. Графическое представление ряда параметров исходного набора данных

В работе ставились две основные цели: построить модель точного прогноза для отслеживания тенденции развития эпидемии денге путем сравнения различных современных алгоритмов и технологий прогнозирования и оценка степени влияния различных показателей исходного набора данных на итоговый результат прогноза. С точки зрения первой цели анализировался достаточно широкий набор инструментов прогнозирования - алгоритмы и технологии в геоинформационных системах, типовых библиотеках машинного обучения, средствах обучения и использования искусственных нейронных сетей, программном обеспечении, позволяющем использовать алгоритмы машинного обучения с помощью графического интерфейса, без необходимости программирования.

В современных геоинформационных системах (ArcGIS Pro, GRASS GIS, SAGA GIS) для моделирования пространственного размещения показателей объектов или явлений используются следующие методы пространственного прогнозирования: IDW, методы локальных и глобальных полиномов, кригинг. Проблема этих методов состоит в том, что они рассчитаны на пространственное прогнозирование, не учитывающее в явном виде временной составляющей. Для того чтобы учитывать данные о времени измерений нужно отдельно использовать либо специализированные алгоритмы (например, ARIMA/SARIMA), либо информацию о годе, месяце, квартале, дне и т.д. преобразовывать в отдельные параметры с числовыми значениями и добавлять их в атрибутивную таблицу внесения предыдущих измерений параметров в качестве отдельных дополнительных колонок анализируемых данных. Используя подобные модификации данных возможно учитывать временную составляющую для обычных алгоритмов регрессионного анализа, например, линейной регрессии, случайного леса. В качестве типовых методов машинного обучения были использованы наиболее популярные технологии – градиентный бустинг на базе деревьев решений (в реализации xgBoost, LightGBM, CatBoost), случайные лес, метод ближайших соседей, линейная регрессия. Кроме традиционной реализации данных технологий и алгоритмов в библиотеках языка python (scikit-learn) представляла интерес оценка возможности их использования в программном обеспечении с графическим интерфейсом. Этот вариант важен тем, что далеко не все специалисты по геоинформационным системам являются программистами в достаточной степени, чтобы использовать все последние достижения в области искусственного интеллекта [18]. Использование графического интерфейса позволяет значительно расширить набор способов анализа данных разных типов. Для оценки качества такого подхода использовался

программный пакет (написать для чего этот пакет) Orange Lab (orange.biolab.si), позволяющий с помощью методов визуального программирования оперировать методами анализа данных и машинного обучения.

Указанные технологии хорошо изучены и являются негласными стандартами при построении математических моделей, если же говорить о перспективах развития, то самым новым направлением для задач прогнозирования пространственно-временных данных являются нейронные сети. Нейронные сети обладают способностью обучаться на имеющихся данных, что имеет большое теоретическое и практическое значение для создания моделей анализа и прогнозирования временных рядов. Дополнительными плюсами этой технологии являются способность эффективно работать в таких нестандартных условиях как недостаточность понимания структуры системы, ошибки и недостаточность в экспериментальных данных. Несмотря на то, что нейронные сети являются нелинейными структурами, они позволяют аппроксимировать произвольную непрерывную функцию. Модель на основе нейронной сети возможно обучить таким образом, чтобы она с высокой достоверностью определяла дальнейшее развитие изучаемого процесса или явления в указанный период. Поскольку временные ряды большинства показателей явлений и процессов представляют собой непрерывные функции, то применение нейронных сетей при их прогнозировании является вполне оправданным и корректным. Процесс использования нейронных сетей строится на использовании библиотек python - tensorflow, theano, pytorch и некоторых других. Кроме того, есть возможность использования созданных моделей и скриптов с интерфейсом ГИС, поскольку во многих распространенных настольных геоинформационных системах языком разработки модулей также является python [19-23].

Говоря о исходных данных для обучения вышеописанных алгоритмов, нужно отметить, что пространственно-временное моделирование описывает и моделирует процессы и явления в четырех измерениях — в трёх пространственных измерениях и во времени. В идеальном случае моделирования необходимо отслеживать все степени свободы (все доступные параметры), но в непрерывных системах каждая точка в пространстве прибавляет дополнительные степени свободы, что в итоге приводит к бесконечному числу измерений. В таких случаях необходимо провести дискретизацию, тем самым уменьшив число степеней свободы до допустимого в компьютерном моделировании. Таким образом, перед тем как перейти непосредственно к созданию модели необходимо выполнить отбор параметров. Это может быть сделано как с помощью расчета числовых показателей корреляции и энтропии, так и визуально, с помощью диаграмм, либо на карте. Сравнение параметров между собой может быть выполнено как с помощью традиционного расчета значения корреляции (например, коэффициенты Пирсона, Спирмена), так и с помощью специализированных расчетных показателей, ориентированных на анализ пространственных данных и временных рядов. Пространственную корреляцию обычно измеряют с помощью индекса Морана, показывающего присутствует ли кластеризация объектов, либо они расположены хаотично. Расчет этого показателя реализован, например, в ArcGIS Pro, GRASS GIS, PySAL. Для анализа энтропии временного ряда наиболее универсальным является показатель Ляпунова. Также для это цели могут быть использованы коэффициент Хёрста, detrended fluctuation analysis. Наибольшее количество инструментов расчета показателей хаотичности временного ряда реализовано в библиотеке nolds (https://pypi.org/project/nolds/) для языка python. Также для оценки значимости конкретных параметров на результат предсказания были использован параметр feature importances

(присутствует в большинстве реализаций алгоритмов scikit-learn) и специализированный алгоритм для отбора параметров – Boruta []. При анализе имеющихся данных выявилась высокая корреляция между значениями температуры, влажности и количества осадков, полученных с метеостанций и по результатам анализа спутниковых измерений. По результатам выполнения оценки важности с помощью Boruta и feature importances наиболее приоритетными являются температура, значения индекса NDVI и сезон.

Наилучшим образом себя обычно проявляют следующие алгоритмы машинного обучения: случайный лес, градиентный бустинг на основе деревьев решений (в реализации xgBoost, LightGBM, CatBoost), нейронная сеть (в реализации Tensorflow и Keras) с двумя архитектурами, различающимися наличием скрытых слоев, SARIMA и ансамблирование результатов работы алгоритмов SARIMA и xgBoost. Эти алгоритмы и были выбраны для анализа данных. Для всех этих алгоритмов выполнялся подбор гиперпараметров и анализировались варианты с отсечением маловажных атрибутов (по результатам анализа корреляционных матриц, параметра feature importances, сводной таблицы Boruta).

**Результаты.** При проведении экспериментов также учитывались следующие особенности:

- для использования нейронных сетей выполнялась предварительная нормализация данных;
- в процессе обучения для алгоритмов в Orange Lab автоматически выполняется подбор гиперпараметров, а для реализации на python это настраивалось вручную;
- для обоих вариантов нейронных сетей использовался метод активации “relu” и оптимизация – “adam”
- скрытые слои для нейронной сети состояли из 5 и 13 элементов;
- для CatBoost использовались варианты с указанием только города в качестве категориального параметра и дополнительно добавлением к нему года, сезона и недели

Результаты качества построения математических моделей описанными алгоритмами приведены в таблице 1.

Таблица 1 Сводная таблица результатов оценки прогностических моделей

| Место | Модель | MAE |
|---|---|---|
| 1. | Ансамблирование SARIMA и xgBoost | 25.8 |
| 2. | Случайный лес, подбор гиперпараметров, отсечение параметров с корреляцией более 0.9 | 26.4 |
| 3. | Случайный лес, подбор гиперпараметров | 26.6 |
| 4. | Случайный лес в Orange | 26.6130 |
| 5. | Случайный лес, подбор гиперпараметров, отсечение атрибутов с рангом Boruta менее 20 | 26.9 |
| 6. | Случайный лес, подбор гиперпараметров, отсечение атрибутов с Feature importances <0.3 | 27.1 |

| | | |
|---|---|---|
| 7. | нейронная сеть с двумя скрытыми слоями | 27.4 |
| 8. | xgBoost, параметры по умолчанию | 27.9 |
| 9. | LightGBM, параметры по умолчанию | 28.7 |
| 10. | CatBoost, подбор гиперпараметров, 5 категориальных переменных | 29.561 |
| 11. | Линейная регрессия в Orange | 29.8173 |
| 12. | SARIMA | 30.3 |
| 13. | Keras, без скрытых слоев | 32.5 |
| 14. | KNN в Orange | 33.8774 |
| 15. | CatBoost, параметры по умолчанию, 4 категориальных переменных | 37.1 |
| 16. | CatBoost, параметры по умолчанию, категориальная переменная - город | 37.2 |

**Заключение.** По результатам проведенных исследований были сформулированы следующие выводы:

- для пространственно-временного прогнозирования с большим количеством параметров необходимо сочетание различных алгоритмов обработки данных методами бустинга или бэггинга;
- предобработка данных как правило является не менее сложным и продуктивным процессом, чем построение моделей;
- графические инструменты для обработки данных и построения моделей, такие как Orange уже практически не уступают по точности скриптам на языке python при этом превосходя их в скорости создания моделей, но необходимо отметить, что пока им не достает гибкости в обработке данных и представлении результатов;
- алгоритм случайного леса наряду с градиентным бустингом являются наиболее универсальными для задач пространственно-временного прогнозирования;
- в случае большого числа входов нейронной сети практически обязательны скрытые слои.

Также для увеличения точности и большей универсальности в дальнейших исследованиях планируется расширить исходные параметры открытыми данными о мониторинге москитов Ae. aegypti и Ae. Albopictus, дополнительными индексами, рассчитываемыми на основе спутникового мониторинга и для построения моделей временного ряда использовать нейронные сети архитектуры LSTM и ее вариаций.

## ЛИТЕРАТУРА


[1] *Москвичев В.В.* Риски развития и мониторинг социально-природно-техногенных систем – основа безопасности, стратегического планирования и управления промышленными регионами страны // Сборник трудов всероссийской конференции (29-31 августа 2017 г., г. Бердск). Новосибирск: ИВТ СО РАН, 2017, С.22-26.

[2] *Колесников А.А., Кикин П.М., Комиссарова Е.В., Грищенко Д.В.* Использование машинного обучения для построения картографических изображений // Международная научно-практическая

конференция «От карты прошлого – к карте будущего», 28 — 30 ноября 2017, г. Пермь – г. Кудымкар. С. 110-120.
[3] *Brown, F.J., Reed C.B., Hayes J.M., Wilhite A.D., Hubbard K.* A prototype drought monitoring system integrating climate and satellite data. Proceedings of the Pecora L5/land satellite information 1V/ISPRS commission I/FIEOS, 2002, Colarado, USA.
[4] *Breiman L.* Random forests. Machine learning, Т. 45, №. 1, 2001. С. 5–32.
[5] *Колесников А.А., Кикин П.М., Комиссарова Е.В., Грищенко Д.В.* Анализ и обработка данных ДЗЗ методами машинного обучения // Сборник материалов V Международной научной конференции "Региональные проблемы дистанционного зондирования Земли" (РПДЗЗ-2018), 2018, г. Красноярск. С. 130-134.
[6] *Goodfellow I., Bengio Y., Courville A.* Deep Learning. MIT Press. 2016, 800 с. ISBN: 9780262035613
[7] *Ничепорчук В.В., Чернякова Н.А.* Использование инфраструктур данных для оценивания рисков чрезвычайных ситуаций // Сборник трудов всероссийской конференции (29-31 августа 2017 г., г. Бердск). Новосибирск: ИВТ СО РАН, 2017, С.280-285.
[8] The Influence of Global Environmental Change on Infectious Disease Dynamics: Workshop Summary. Washington (DC): National Academies Press (US); 2014 №3. https://www.ncbi.nlm.nih.gov/books/NBK241611/ (дата обращения 11.06.2019).
[9] *Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A. L.* Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv preprint arXiv:1606.00915, 2016.
[10] *Kuno G.* Research on dengue and dengue-like illness in East Asia and the Western Pacific during the First Half of the 20th century // Reviews in medical virology. 2007. № 17(5):327–341. doi: 10.1002/rmv.545
[11] *Xiao JP, He JF, Deng AP, Lin HL, Song T, Peng ZQ,* Characterizing a large outbreak of dengue fever in Guangdong Province, China. Infectious diseases of poverty. 2016. // PubMed Central PMCID: PMCPmc4853873. doi: 10.1186/s40249-016-0131-z
[12] *Shepard DS, Undurraga EA, Halasa YA.* Economic and disease burden of dengue in Southeast Asia // PLoS neglected tropical diseases. 2013, №7(2):e2055 PubMed Central PMCID: PMCPmc3578748. doi: 10.1371/journal.pntd.0002055
[13] *Ooi E., Gubler D.* Dengue in Southeast Asia: epidemiological characteristics and strategic challenges in disease prevention // Cadernos de saude publica. 2009;25 Suppl 1:S115–24.
[14] *Halstead S.* Dengue in the Americas and Southeast Asia: do they differ? // Revista panamericana de salud publica. 2006. №20(6), С.407–415.
[15] *Haug S., Ostermann J.* A Crop Weed Field Image Dataset for the Evaluation of Computer Vision Based Precision Agriculture Tasks // Computer Vision - ECCV 2014 Workshops. Zurich: Springer, 2014, С. 105–116.
[16] *Peters, J.A., Walter-Shea A.E., Ji L., Vina A., Hayes M., Svoboda D.M.* Drought monitoring with NDVI-based standardized vegetation index. Photogrammetric Engineering and Remote Sensing, 2002. 68:7175.
[17] *Бериков В.Б., Пестунов И.А., Караев Н.М., Тевари А.* Распознавание гиперспектральных изображений с использованием кластерного ансамбля и частично контролируемого обучения // Сборник трудов всероссийской конференции (29-31 августа 2017 г., г. Бердск). Новосибирск: ИВТ СО РАН, 2017, С.60-65.
[18] *Bottou L.* Large-scale machine learning with stochastic gradient descent // Proceedings of COMPSTAT’ 2010, Springer, 2010, С. 177–186.
[19] *Badrinarayanan V., Kendall A., Cipolla R.* Convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, № 39 (12), С. 2481-2495.
[20] *Dai J., He K., Sun J.* Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation // IEEE International Conference on Computer Vision, 2015, С. 1635–1643.
[21] *Eigen D., Fergus R.* Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture // IEEE International Conference on Computer Vision, 2015, С. 2650–2658.
[22] *Hariharan B., Arbelaez P., Girshick R., Malik J.* Simultaneous detection and segmentation // European Conference on Computer Vision. Springer, 2014, С. 297–312.

[23] *Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A. C., Fei-Fei L.* ImageNet Large Scale Visual Recognition Challenge // International Journal of Computer Vision (IJCV), Т. 115, №. 3, 2015, С. 211–252.