УДК: 528.93 **DOI: 10.35595/2414-9179-2020-3-26-53-61**

П.М. Кикин[1], А.А. Колесников[2], А.М. Портнов[3], Д.В. Грищенко[4]

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ПРОСТРАНСТВЕННО-ВРЕМЕННЫХ ЭКОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

АННОТАЦИЯ

Состояние экологических систем наряду с их общими характеристиками практически всегда описывается показателями, изменяющимися в пространстве и времени, что приводит к существенному усложнению построения математических моделей для прогнозирования состояния таких систем. Одним из способов, позволяющих упростить и автоматизировать построение математических моделей для прогнозирования состояния таких систем, является использование методов машинного обучения. В статье приводится сравнение традиционных и основанных на нейронных сетях алгоритмов и методов машинного обучения для прогнозирования пространственно-временных рядов, представляющих данные экосистем. Анализ и сравнение проводились среди следующих алгоритмов и методов: логистическая регрессия, случайный лес, градиентный бустинг на деревьях решений, SARIMAX, нейронные сети долгой краткосрочной памяти (LSTM) и управляемых рекуррентных блоков (GRU). Для проведения исследования были подобраны наборы данных, имеющих как пространственную, так и временную составляющие: значения численности москитов, количество заражений лихорадкой денге, физическое состояние деревьев тропической рощи, уровень воды в реке. В статье рассматриваются необходимые действия по предварительной обработке данных, в зависимости от используемого алгоритма. Также в качестве одного из параметров, которые могут помочь формализовать выбор наиболее оптимального алгоритма при построении математических моделей пространственно-временных данных для используемых наборов, была вычислена колмогоровская сложность. По результатам проведенного анализа даются рекомендации по применению тех или иных методов и конкретных технических решений в зависимости от особенностей набора данных, описывающего конкретную экосистему.

КЛЮЧЕВЫЕ СЛОВА: экосистемы, пространственно-временные показатели, LSTM, SARIMAX, прогнозирование

[1] Санкт-Петербургский политехнический университет Петра Великого (СПбПУ), ул. Политехническая, д. 29, 195251, Санкт-Петербург, Россия; *e-mail:* **it-technologies@yandex.ru**
[2] Сибирский государственный университет геосистем и технологий, ул. Плахотного, д. 10, 630108, Новосибирск, Россия; *e-mail:* **alexeykw@mail.ru**
[3] Московский государственный университет геодезии и картографии, Гороховский пер., д. 4, 105064, Москва, Россия; *e-mail:* **portnov@miigaik.ru**
[4] Сибирский государственный университет геосистем и технологий, ул. Плахотного, д. 10, 630108, Новосибирск, Россия; *e-mail:* **mr_divis@mail.ru**

Pavel M. Kikin[1], **Alexey A. Kolesnikov**[2], **Alexey M. Portnov**[3], **Denis V. Grischenko**[4]

NATURAL LANGUAGE PROCESSING SYSTEMS FOR DATA EXTRACTION AND MAPPING ON THE BASIS OF UNSTRUCTURED TEXT BLOCKS

ABSTRACT

The state of ecological systems, along with their general characteristics, is almost always described by indicators that vary in space and time, which leads to a significant complication of constructing mathematical models for predicting the state of such systems. One of the ways to simplify and automate the construction of mathematical models for predicting the state of such systems is the use of machine learning methods. The article provides a comparison of traditional and based on neural networks, algorithms and machine learning methods for predicting spatio-temporal series representing ecosystem data. Analysis and comparison were carried out among the following algorithms and methods: logistic regression, random forest, gradient boosting on decision trees, SARIMAX, neural networks of long-term short-term memory (LSTM) and controlled recurrent blocks (GRU). To conduct the study, data sets were selected that have both spatial and temporal components: the values of the number of mosquitoes, the number of dengue infections, the physical condition of tropical grove trees, and the water level in the river. The article discusses the necessary steps for preliminary data processing, depending on the algorithm used. Also, Kolmogorov complexity was calculated as one of the parameters that can help formalize the choice of the most optimal algorithm when constructing mathematical models of spatio-temporal data for the sets used. Based on the results of the analysis, recommendations are given on the application of certain methods and specific technical solutions, depending on the characteristics of the data set that describes a particular ecosystem.

KEYWORDS: ecosystems, spatio-temporal indicators, LSTM, SARIMAX, forecasting

ВВЕДЕНИЕ

Оценка потенциала и перспективы применения современных технологий машинного обучения и искусственного интеллекта в области экологии были представлены в 2000 г. на международной конференции Applications of Machine Learning to Ecological Modelling. Труды конференции описывают преимущества методов машинного обучения при работе с многомерными данными, обладающими нелинейностью, сложными взаимосвязями и разнородными ошибками в данных, а также описаны преимущества при совместном использовании различных типов экологических моделей для описания экосистемы, которые развиваются в структуре и поведении [*Knudby*, 2010].

Разнообразие актуальных алгоритмов машинного обучения, большое количество параметров и индивидуальные требования к подготовке исходных данных не позволяют специалистам, работающим с информацией в области экологических систем и явлений, эффективно их использовать и строить прогнозные математические модели [*Haupt*, 2009, *Olden*, 2008]. Таким образом, основной задачей исследования являлась формулировка

[1] Peter the Great St. Petersburg Polytechnic University (SPbPU), Polytechnicheskaya str., 29, 195251, St. Petersburg, Russia; *e-mail:* **it-technologies@yandex.ru**
[2] Siberian State University of Geosystems and Technologies, Plakhotnogo str., 10, 630108, Novosibirsk, Russia; *e-mail*: **alexeykw@mail.ru**
[3] Moscow State University of Geodesy and Cartography, Gorokhovsky lane, 4, 105064, Moscow, Russia; *e-mail*: **portnov@miigaik.ru**
[4] Siberian State University of Geosystems and Technologies, Plakhotnogo str., 10, 630108, Novosibirsk, Russia; *e-mail*: **mr_divis@mail.ru**

рекомендаций по подготовке и инжинирингу исходных данных и выбору конкретного, наиболее подходящего алгоритма на основе критериев оценки исходного набора данных.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЙ

Для проведения исследования требовались данные, в которых среди характеристик объекта или явления присутствуют пространственная и временная составляющие. Исходя из этих требований, были выбраны следующие наборы данных:

- численность москитов Aedes aegypti на территории западной и прибрежной частей Кении [*Ndenga,* 2017];
- количество заражений лихорадкой денге на территории Пуэрто-Рико и Перу[1];
- физическое состояние деревьев в низинной тропической роще на территории Пуэрто-Рико [*David,* 2006];
- уровень воды в бассейне реки Соан, Пакистан[2].

Сводные характеристики указанных наборов данных представлены в табл. 1. Каждый набор данных был дополнен показателем колмогоровской сложности, описывающим пространственно-временные ряды с целью возможности определения эффективности того или иного подхода машинного обучения [*Ульянов,* 2013; *Сметанин*, 2014].

Табл. 1. Характеристики используемых наборов данных
Table 1. Parameters of used datasets

Набор данных	Количество экзогенных показателей	Временной интервал данных	Количество наблюдений	Стационарность	Сезонность	Колмогоровская сложность
Численность москитов	3	06.2014–06.2016	200	Нет	Да	0.13
Количество заболеваний денге	24	1990–2010	1456	Да	Да	0.16
Состояние деревьев	9	1983–2000	3381	Да	Нет	0.03
Уровень воды	1	1980–2000	18263	Да	Да	0.13

В наборе данные, которые были использованы для прогнозирования численности москитов. Исходными параметрами являлись температура, влажность, количество осадков[3], средние значения индекса NDVI для пикселей спутникового снимка, окружающих анализируемые города. Графическое представление количества заражений, являющееся целевым параметром в используемом наборе данных, отображено на рис. 1.

[1] Center for Disease Control and Prevention. Электронный ресурс: https://www.cdc.gov (дата обращения 27.12.2019)
[2] Pakistan meteorological department (PMD). Электронный ресурс: www.pmd.gov.pk (дата обращения 27.12.2019)
[3] National Oceanic and Atmospheric Administration. Электронный ресурс: https://www.noaa.gov (дата обращения: 27.12.2019)

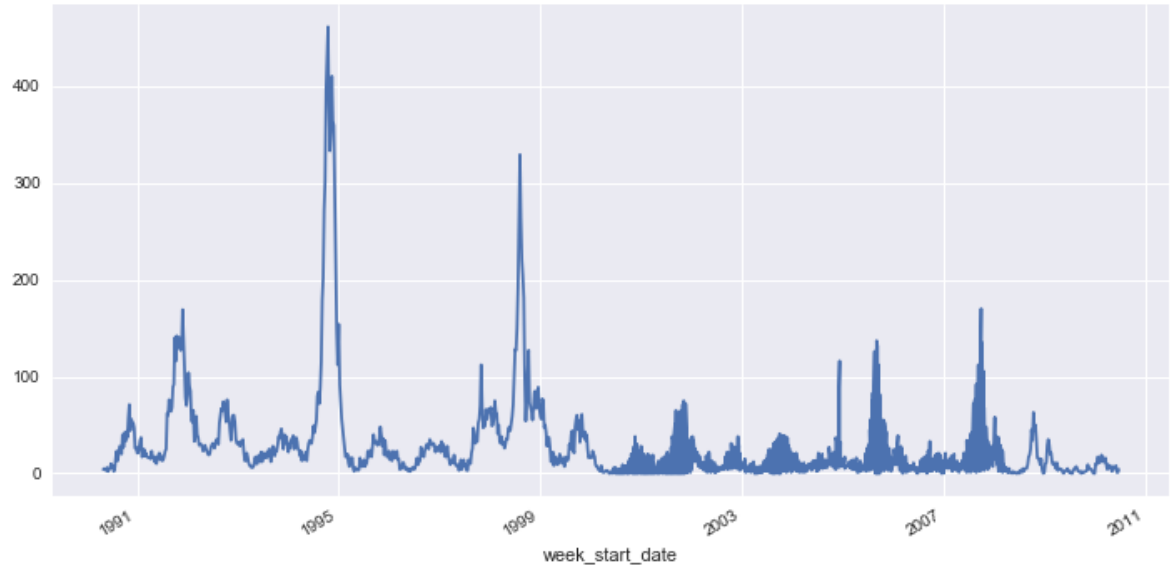

Рис. 1. График количества заражений лихорадкой денге
Fig. 1. Dengue fever count

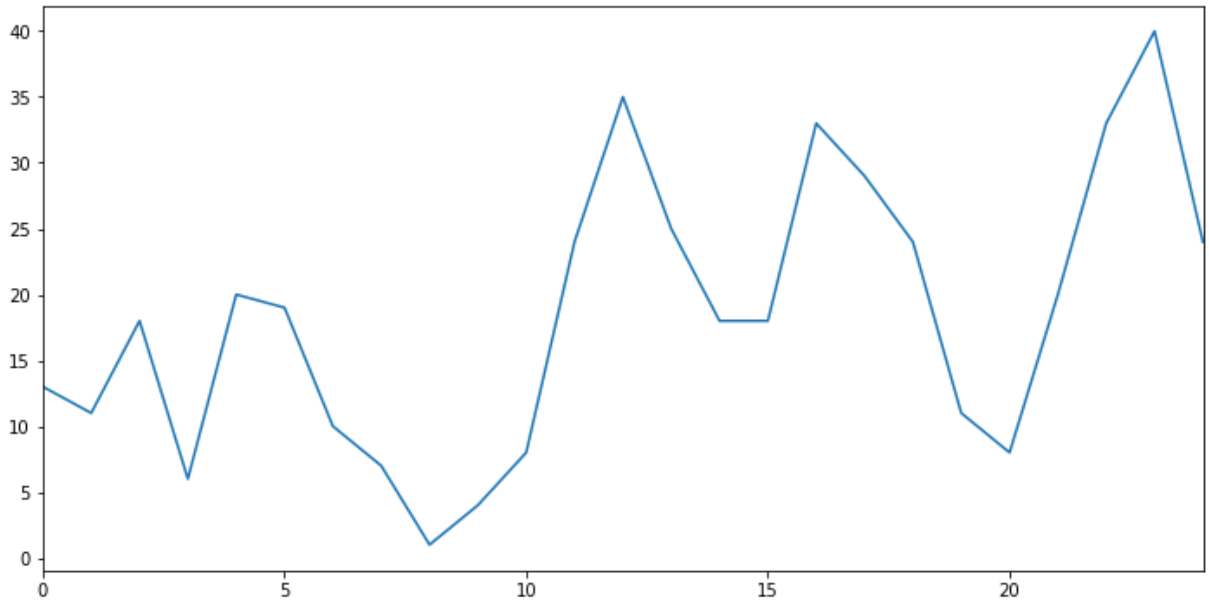

Рис. 2. График изменения численности москитов
Fig. 2. Mosquitoes count

Для данных, содержащих значения численности москитов дополнительно на указанные даты и для анализируемого района, из открытых источников (GlobeLand30, MODIS, WorldClim, ALOS Palsar) были собраны дополнительные характеристики — тип ландшафта, температура, NDVI, влажность, характеристики рельефа. Визуально изменение количества москитов за анализируемый период приведено на рис. 2.

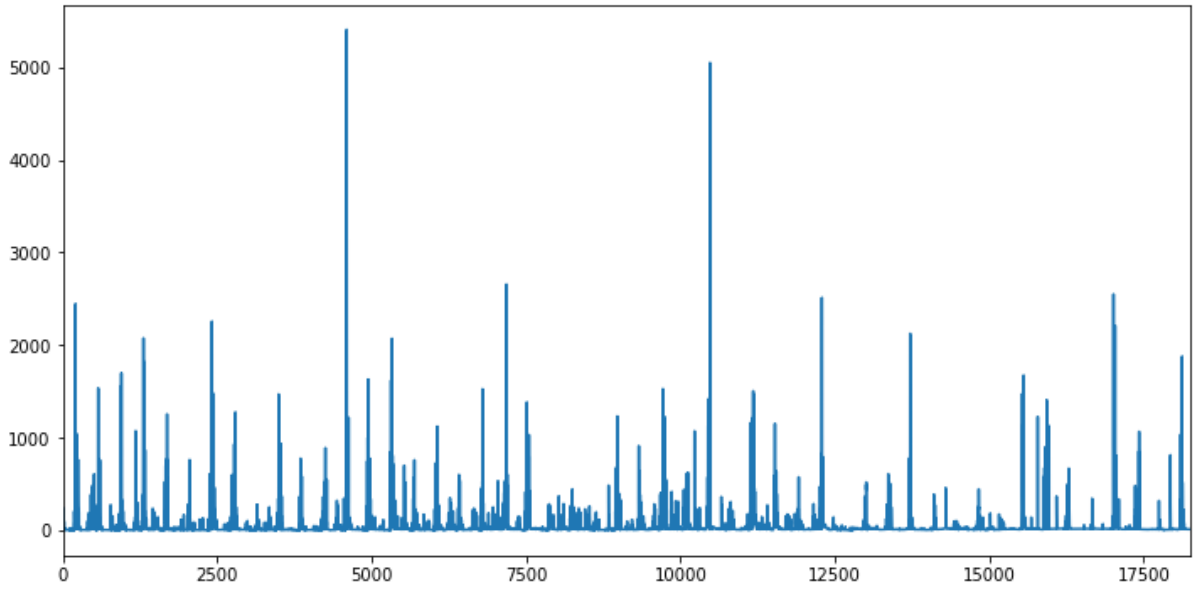

Рис. 3. График изменения уровня воды реки Соан
Fig. 3. Soan River water level

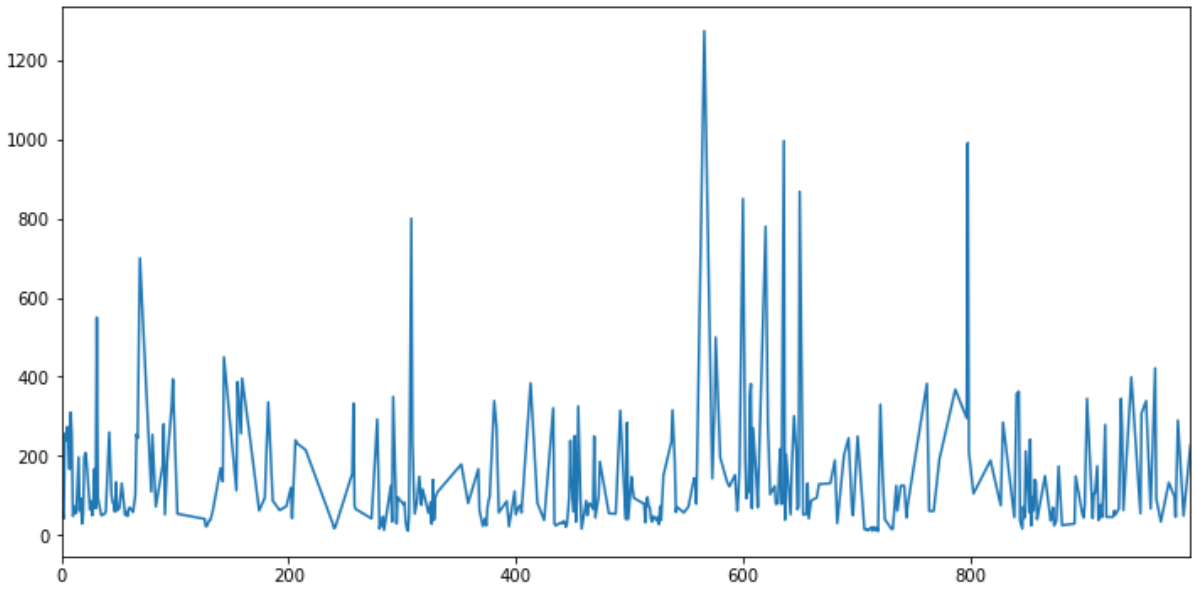

Рис. 4. График изменения сводной характеристики качества деревьев
Fig. 4. Graph of changes in the summary quality characteristic of trees

Построение математической модели прогнозирования уровня воды р. Соан базировалось только на значении количестве осадков. На рис. 3 показано изменение уровня вода за рассматриваемый период времени.

Оценка и прогноз изменения характеристик в используемом наборе данных основывался на годичных циклах, поскольку именно такой временной интервал наиболее оптимально отражает реакцию деревьев на изменения климата. Параметры представлены измерениями диаметра и высоты, состояния ствола и показателей освещённости кроны и структуры леса в целом. Графическое представление сводной характеристики качества деревьев отображено в виде графика на рис. 4.

Для построения прогнозов были использованы алгоритмы, которые наиболее успешно используются для работы с пространственно-временными данными:

- логистическая регрессия [*McCullagh*, 1989];
- random forest [*Liaw*, 2002];
- LightGBM, xgBoost, CatBoost;
- SARIMAX [*Arunra*j, 2016];
- LSTM (Simple, Stacked, Bi-Directional, CNN-LSTM) [*Hochreiter*, 1997] [Schuster, 1997], [*Chiu*, 2015];
- GRU [*Chung*, 2014].

Логистическая регрессия в данном исследовании была использована в качестве точки отсчёта для последующего сравнения метрик качества предсказания других методов. Random forest и другие алгоритмы, основанные на деревьях решений LightGBM, xgBoost, CatBoost и использующие градиентный бустинг в процессе обучения, на сегодняшний день остаются одним из самых надёжных и эффективных способов построения предсказательных моделей для табличных данных. Они обладают высокой скоростью обучения и устойчивы к ненормализованным числовым показателям, а также к наличию выбросов в данных. SARIMAX — наиболее изученный алгоритм прогнозирования числовых рядов с учётом сезонности и экзогенных параметров, куда были отнесены и данные о пространственном положении.

Также были использованы варианты рекуррентных нейронных сетей — LSTM и GRU. В обоих случаях тренировочные данные были преобразованы в числовой формат и нормализованы.

LSTM представляет собой развитие принципов рекуррентных нейронных сетей призванное решить их проблему с невозможностью «запоминания» долговременных зависимостей и работой с переменной величиной контекста. Архитектура LSTM спроектирована таким образом, чтобы «запоминать» взаимосвязи как на короткие, так и на длинные промежутки времени в анализируемых временных рядах, при этом используя механизмы «забывания», чтобы отсеивать несущественные взаимосвязи.

GRU более простой вариант с точки зрения количества вычислений, LSTM архитектуры и при этом в большинстве задач показывающий аналогичные результаты по точности. Основываясь на принципах работы GRU вариант должен показывать более качественные результаты при небольших объёмах данных для обучения.

Для всех используемых алгоритмов категориальные признаки обучающих наборов кодировались методом One Hot Encoding. Последующая предобработка наборов данных выполнялась в зависимости от используемого алгоритма. Для логистической регрессии ко всем имеющимся данным применялась нормализация. В случае алгоритма SARIMAX обучающая выборка формировалась на основе значений дат и прогнозируемого показателя, все остальные параметры помещались в массив экзогенных параметров. Для всех используемых вариаций LSTM и GRU, кроме обязательной нормализации, производилось преобразование данных таким образом, чтобы в каждой записи дублировались данные о нескольких предыдущих шагах. Значение количества шагов, дописываемых таким образом к каждой записи, подбиралось экспериментально и для описываемых данных составляло 7 единиц для набора данных о количестве заболеваний денге, 4 — для набора данных о состоянии деревьев, и 5 — для данных о численности москитов и уровне воды.

Построенные вышеперечисленными методами математические модели обладают низкой степенью интерпретируемости, а также параметры самих нейронных сетей (количество слоёв и нейронов в них, порядок слоёв, функции активации и т.д.) требуется подбирать экспериментально, что приводит к повышенной сложности их использования на данных, подобных используемым в описываемом исследовании.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ И ИХ ОБСУЖДЕНИЕ

Результаты для всех описанных наборов данных приведены в табл. 2.

Для примера графическое представление результатов прогнозирования с помощью алгоритмов логистической регрессии, случайного леса, LightGBM, LSTM, GRU и истинного уровня (линия True) р. Соан приведено на рис. 5.

Табл. 2. Сводная таблица результатов
Table 2. Summary table of the results

MAE	Логистическая регрессия	Random forest	LightGBM	SARIMAX	LSTM Simple	LSTM Stacked	LSTM Bi-Direc-tional	GRU
Численность москитов	9.58	14.30	13.91	8.12	7.56	7.13	6.49	6.68
Количество заболеваний денге	37.08	26.56	26.05	34.66	30.56	30.05	28.89	30.45
Состояние деревьев	32.14	20.14	19.23	26.61	22.80	21.55	20.38	21.63
Уровень воды	14.16	27.60	29.60	14.71	13.16	12.83	11.99	12.40

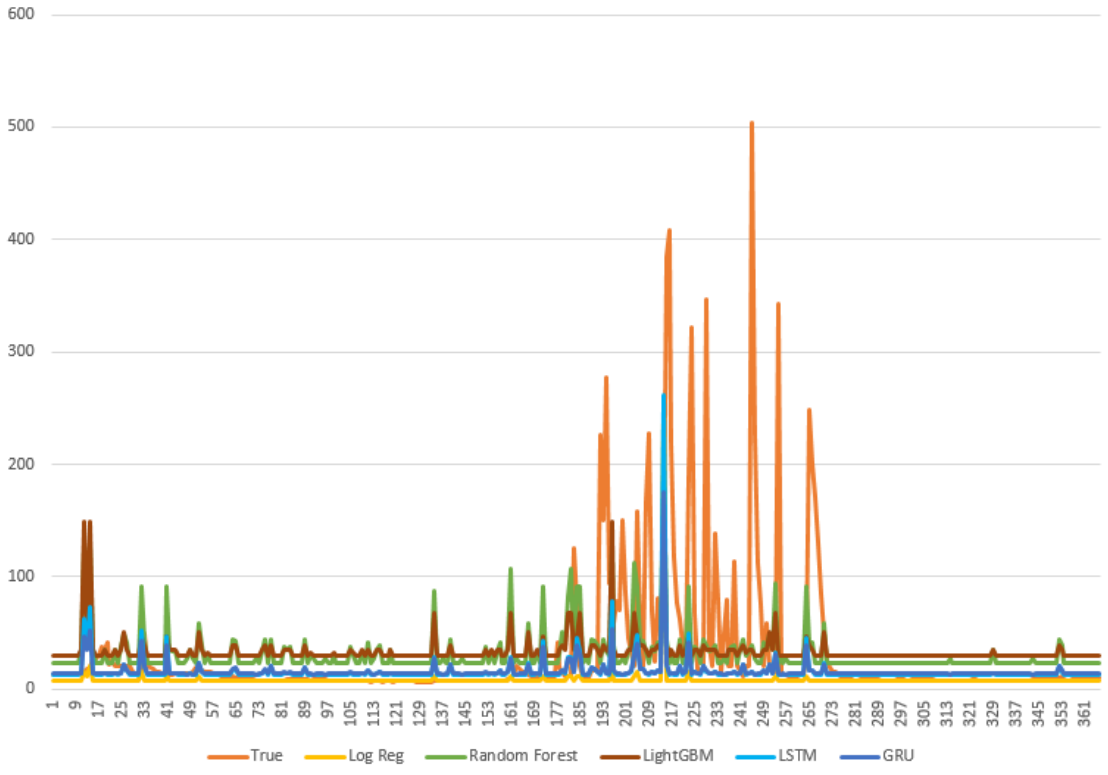

Рис. 5. Сравнение результатов прогнозирования уровня реки, полученного несколькими алгоритмами с истинными показателями
Fig. 5. Comparison of river level prediction results obtained by several algorithms with true values

ВЫВОДЫ

В результате анализа качества предсказаний, сделанных различными алгоритмами, было выявлено, что для пространственно-временных рядов с большим числом экзогенных параметров наилучшую предсказательную способность демонстрируют алгоритмы на основе деревьев решений. Для наборов данных с малым числом дополнительных параметров наилучшую предсказательную способность показывают рекуррентные нейронные сети.

Зависимость качества предсказания использованных алгоритмов от колмогоровской сложности исходных данных не была выявлена, что тем не менее может быть обусловлено недостаточно репрезентативной выборкой.

В дальнейших исследованиях планируется провести апробацию описанных алгоритмов на большем разнообразии исходных данных, а также применить модели Хольта-Уинтерса, библиотеку Facebook Prophet и интерпретируемые LSTM сети (IMV-LSTM).

Дополнительно планируется разработать набор методических указаний по формированию архитектуры рекуррентных нейронных сетей и подбору их гиперпараметров в зависимости от особенностей исходных данных.

Также нужны исследования, ориентированные на расчёт и формализацию величины перекрытия (обратных отсчётов) в пакетах данных (для используемых наборов данных наличие и величина перекрытия слабо влияла на точность используемых RNN).

СПИСОК ЛИТЕРАТУРЫ

1. *Сметанин Ю.Г., Ульянов М.В.* Построение кластерного пространства временных рядов: колмогоровская и гармоническая сложность. Научные труды Вольного экономического общества России, 2014. № 186 (186). С. 124−129.

2. *Ульянов М. В., Сметанин Ю. Г.* Подход к определению характеристик колмогоровской сложности временных рядов на основе символьных описаний. Бизнес-информатика, 2013. № 2. С. 49−54.

3. *Arunraj N.S., Ahrens D., Fernandes M.* Application of SARIMAX model to forecast daily sales in food retail industry. International Journal of Operations Research and Information Systems, 2016. V. 7 (2). P. 1−21. DOI: 10.4018/ijoris.2016040101.

4. *Chiu J., Jason P. C., Nichols E.* Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 2015. V. 4. P. 357−370.

5. *Chung J., Gulcehre C., Cho K.H., Bengio Y.* Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, 2014. 9 p. DOI: arXiv:1412.3555 [cs.NE].

6. *Clark D. B., Clark D. A.* Tree growth, mortality, physical condition, and microsite in an old-growth lowland tropical rain forest. Ecology, 2006. V. 87. P. 2132−2132. DOI: 10.1890/0012-9658(2006)87[2132:TGMPCA]2.0.CO;2

7. *Haupt S., Pasini A., Marzban C.* Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, 2009. 424 p. DOI: 10.1007/978-1-4020-9119-3.

8. *Hochreiter S., Schmidhuber J.* Long Short-Term Memory. Neural Computing, 1997. V. 9−8. P. 1735−1780. DOI: dx.doi.org/10.1162/neco.1997.9.8.1735.

9. *Knudby A., Brenning A., LeDrew E.* New approaches to modelling fish–habitat relationships. Ecological Modelling, 2010. V. 221 (3). P. 503−511. DOI: 10.1016/j.ecolmodel.2009.11.008.

10. *Liaw A., Wiener M.* Classification and Regression by Random Forest. R News, 2002. V. 2 (3). P. 18−22.

11. *McCullagh P., Nelder J.A.* Generalized linear models. 2nd edition. Taylor & Francis, 1989. 532 p.

12. *Ndenga B.A., Mutuku F.M., Ngugi H.N.* Characteristics of Aedes aegypti adult mosquitoes in rural and urban areas of western and coastal Kenya. PLoS One, 2017. V. 12 (12): e0189971. DOI: 10.1371/journal.pone.0189971.

13. *Olden J., Lawler J., Poff N.L.* Machine learning methods without tears: A primer for ecologists. The Quarterly Review of Biology, 2017. V. 83 (2). P. 171−193. DOI: 10.1086/587826.

14. *Schuster M., Paliwal K.K.* Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 1997. V. 45−11. P. 2673−2681. DOI: dx.doi.org/10.1109/78.650093.

REFERENCES

1. *Arunraj N. S., Ahrens D., Fernandes M.* Application of SARIMAX model to forecast daily sales in food retail industry. International Journal of Operations Research and Information Systems, 2016. V. 7 (2). P. 1−21. DOI: 10.4018/ijoris.2016040101.
2. *Chiu J., Jason P.C., Nichols E.* Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 2015. V. 4. P. 357−370.
3. *Chung J., Gulcehre C., Cho K.H., Bengio Y.* Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. 9 p. DOI: arXiv:1412.3555 [cs.NE].
4. *Clark D.B., Clark D.A.* Tree growth, mortality, physical condition, and microsite in an old-growth lowland tropical rain forest. Ecology, 2006. V. 87. P. 2132−2132. DOI: 10.1890/0012-9658(2006)87[2132:TGMPCA]2.0.CO;2.
5. *Haupt S., Pasini A., Marzban C.* Artificial intelligence methods in the environmental sciences. Springer Netherlands, Amsterdam, 2009. 424 p. DOI: 10.1007/978-1-4020-9119-3.
6. *Hochreiter S., Schmidhuber J.* Long short-term memory. Neural Computing, 1997. V. 9−8. P. 1735−1780. DOI: dx.doi.org/10.1162/neco.1997.9.8.1735.
7. *Knudby A., Brenning A., LeDrew E.* New approaches to modelling fish–habitat relationships. Ecological Modelling, 2010. V. 221 (3). P. 503−511. DOI: 10.1016/j.ecolmodel.2009.11.008.
8. *Liaw A., Wiener M.* Classification and regression by random forest. R News, 2002. V. 2 (3). P. 18−22.
9. *McCullagh P., Nelder J.A.* Generalized linear models. 2nd edition. Taylor & Francis, 1989. 532 p.
10. *Ndenga B.A., Mutuku F.M., Ngugi H.N.* Characteristics of Aedes aegypti adult mosquitoes in rural and urban areas of western and coastal Kenya. PLoS One, 2017. V. 12 (12): e0189971. DOI: 10.1371/journal.pone.0189971.
11. *Olden J., Lawler J., Poff N.L.* Machine learning methods without tears: a primer for ecologists. The Quarterly Review of Biology, 2017. V. 83 (2). P. 171−193. DOI: 10.1086/587826.
12. *Schuster M., Paliwal K.K.* Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 1997. V. 45−11. P. 2673−2681. DOI: dx.doi.org/10.1109/78.650093.
13. *Smetanin Y.G., Ulyanov M.V.* The design of cluster spaceoftime series: Kolmogorov and harmonious complexity. Scientific works of the Free Economic Society of Russia, 2014. V. 186 (186). P. 124−129 (in Russian).
14. *Ulyanov M. V., Smetanin Y. G.* An approach to characterizing the Kolmogorov complexity of time series based on symbolic descriptions. Business Informatics, 2013. V. 2. P. 49−54 (in Russian).