УДК 614.84, 004.8
doi: 10.34987/vestnik.sibpsa.2021.20.1.013

Прогнозирование временных рядов на основе методов машинного обучения в вопросах обеспечения природной и техносферной безопасности

Бабёнышев С.В., к. ф.-м. н.; Малютин О.С.; Матеров Е.Н. к. ф.-м. н.
ФГБОУ ВО Сибирская пожарно-спасательная академия ГПС МЧС России

Аннотация. Основная цель статьи – сделать обзор современных возможностей анализа и моделирования временных рядов на примерах прогнозирования количества пожаров и уровня подъема воды в реках с помощью современных методов машинного обучения в среде программирования R. Особенность данного моделирования состоит в возможности использования нескольких моделей одновременно, что позволяет автоматизировано выбирать модели с наименьшими погрешностями.

Ключевые слова: моделирование временных рядов, машинное обучение, количество пожаров

Time series forecasting based on machine learning methods for ensuring natural and technosphere safety

Babenyshev S.V., Ph.D. of Physical and Mathematical Sciences; Malyutin O.S.; Materov E.N., Ph.D. of Physical and Mathematical Sciences
FSBEE HE Siberian Fire and Rescue Academy EMERCOM of Russia

Abstract. The main purpose of the article is to give a review of some capabilities of time series analysis and modeling using examples of predicting the number of fires and the level of river flooding using modern machine learning methods in the R programming language environment. The peculiarity of this simulation is the possibility of using several models at the same time, which allows to automatically select models with the lowest bias errors.

Keywords: time series modeling, machine learning, number of fires

Введение

Для решения задач обеспечения природной и техногенной безопасности, а также мониторинга потенциально опасных территорий и объектов, в органах управления МЧС России различного уровня существует большое разнообразие различных программных систем и информационных ресурсов. В настоящий момент накоплен большой объем данных, который может быть использован для разработки автоматизированного математического и программного обеспечения решения задач управления мероприятиями по предупреждению и ликвидации чрезвычайных ситуаций включая моделирование и прогнозирование рисков их появления.

Для анализа больших объемов данных, имеющих сложный структурный характер, существует множество подходов. Современными инструментами такого рода анализа данных являются языки программирования, решающие многие задачи в Data Science, машинном обучении, big data и т.д. Лидирующие позиции среди таких языков программирования занимают языки Python, R и Julia. Python является наиболее популярным языком среди указанных, его основными качествами является простота изучения и широкая область применения. Язык R предоставляет богатые возможности библиотек статистической обработки данных, визуализации данных и возможности для статических и интерактивных отчетов. Язык Julia характеризуется высокой скоростью работы. Все три языка программирования имеют ориентированность на задачи машинного обучения и работу с большим объемом данных, и являются свободно распространяемыми программными

средами с открытым кодом. В данной статье использовался язык R, имеющий продуманную экосистему, основанную на библиотеках tidyverse, tidymodels и дочерних библиотеках. Настоящая работа продолжает цикл публикаций [1-3]. Основными источниками данных в статье являлись данные по пожарам в Красноярском крае и в Новосибирской области за последние 5 лет. Напомним, что учет пожаров и их последствий ведется на основании электронных баз данных, регламентированных приказом МЧС России [4].

Авторы выражают глубокую признательность за полезные обсуждения начальнику ФГБОУ ВО Сибирская пожарно-спасательная академия ГПС МЧС России А.А. Назарову, начальнику отдела прикладных исследований и инновационных технологий научно-технического центра ФГБОУ ВО Сибирская пожарно-спасательная академия ГПС МЧС России Н.В. Мартиновичу и старшему научному сотруднику ИВМ СО РАН В.В. Ничепорчуку.

1. Предварительный анализ и подготовка данных для работы с алгоритмами машинного обучения на примере статистики пожаров и их последствий

Язык программирования R является свободно распространяемым языком, актуальную версию которого можно загрузить и установить с сайта https://www.r-project.org/. Для работы с R также необходима IDE (интегрированная среда разработки), в качестве которой можно использовать RStudio (доступно по адресу: https://rstudio.com/) или Visual Studio Code (доступно по адресу: https://code.visualstudio.com/).

Первый этап работы с данными, – загрузка данных в R в виде таблиц (или фрейма данных). Данные в R можно загружать множеством способов: из Excel или CSV-файлов, из Google Sheets, а также непосредственно из баз данных. После загрузки данных, необходимо понять их структуру, определить типы переменных (столбцов таблицы), «уровни» факторных переменных, понять базовые статистические характеристики данных. Отметим, что язык R дает возможность работать с пропущенными данными, в отличие от некоторых иных программ для работы с электронными таблицами. Следующий этап заключается в преобразовании данных: выбора подмножества данных, группировки, получении базовых статистик, работы со сводными таблицами и т.д. Для понимания поведения данных необходима их визуализация, например, используя библиотеку ggplot2 и наследственные библиотеки. Замыкает цикл работы с данными моделирование (мы рассматриваем моделирование в параграфах 2 и 3 настоящей статьи) и составление отчета (примеры возможностей R для автоматизированного формирования отчетов будут рассмотрены в другой работе). Описанная цепочка работы с данными показана на рис. 1 (см. [5]).

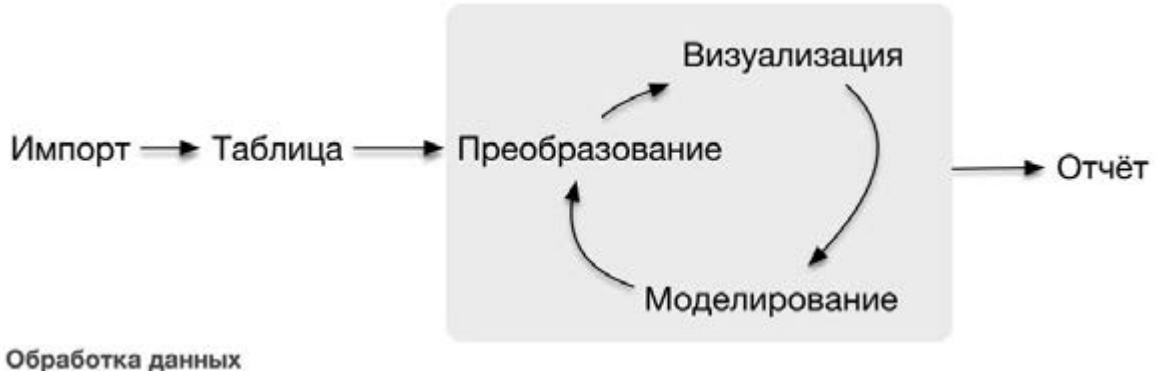

Рис. 1. Общая схема работы с данными в R

В качестве первого примера, рассмотрим ежедневное количество пожаров в 2019 году в Новосибирской области и в Красноярском крае по данным Главных управлений МЧС России и проведем небольшой сравнительный анализ.

На рис. 2 показано суточное количество пожаров в Красноярском крае и в Новосибирской области в 2019 году. Из графика видно, что оба региона имеют сходный характер роста пожаров:

1. большой всплеск количества пожаров, начинающийся с апреля по июнь, пик которого приходится на конец апреля – начало мая, объясняемый весенним горением травы и мусора;
2. небольшой осенний рост пожаров, приходящийся на октябрь – ноябрь.

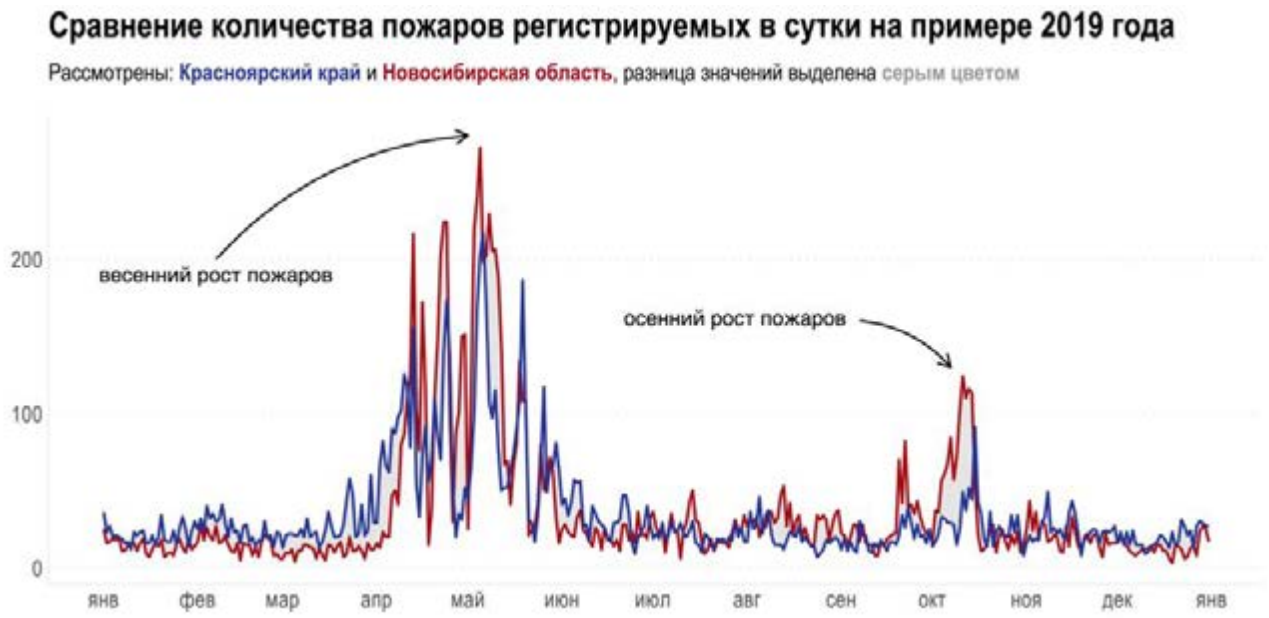

Рис. 2. Количество пожаров в Красноярском крае и в Новосибирской области в 2019 году, регистрируемых за сутки

Отметим, что такое ежегодное поведение графика количества пожаров, с различными вариациями, можно наблюдать и в других регионах, более того, оно зачастую является типичным.

Используя анализ данных и их визуализацию с помощью библиотек tidyverse, нетрудно получить информацию самого различного характера, например, распределение количества погибших в зависимости от времени суток.

Рис. 3. Распределение количества погибших (чем светлее ячейка, тем больше случаев, значение в ячейке соответствует количеству случаев) в зависимости от времени суток в Новосибирской области в 2016-2020 гг.

Из рис. 3, например, видно, что наибольшее количество погибших, приходится на ночной промежуток времени от 0:00 до 2:00.

2. Анализ временных рядов, связанных с вопросами природной и техносферной безопасности

Под временным рядом обычно понимается последовательность {yt} значений переменной, которые принимают значения через определенные (обычно регулярные) значения времени t. Область применения временных рядов очень обширна, временные ряды используются в сейсмологии, метеорологии, экономике, а также при регистрации значений любых датчиков. В пожарной безопасности в качестве временных рядов выступают, например, количество пожаров, регистрируемых в сутки/неделю/месяц. Временным рядам

посвящено большое количество литературы, в частности, работа с временными рядами в среде R описана в монографиях [6] и [7]. Некоторые вопросы работы с временными рядами в R, касающиеся количества пожаров рассматривались в статье [2] и в книге [8].

Обычно для описания временного ряда, имеющего стационарный характер и не имеющего долгосрочных циклических компонент, как в нашем случае, применяют аддитивную модель: $y_t = T_t + S_t + R_t$, где T_t – тренд, характеризующий долгосрочную тенденцию возрастания/убывания данных, S_t – сезонная компонента, которая показывает краткосрочные периодические изменения и R_t – остаток, являющийся необъясняемой величиной временного ряда, так называемый белый шум. Для декомпозиции временных рядов можно использовать различные методы. Одним из наиболее используемых является STL-разложение (от *Seasonal and Trend decomposition using Loess*). Пример графического STL-разложения показан на рис. 4.

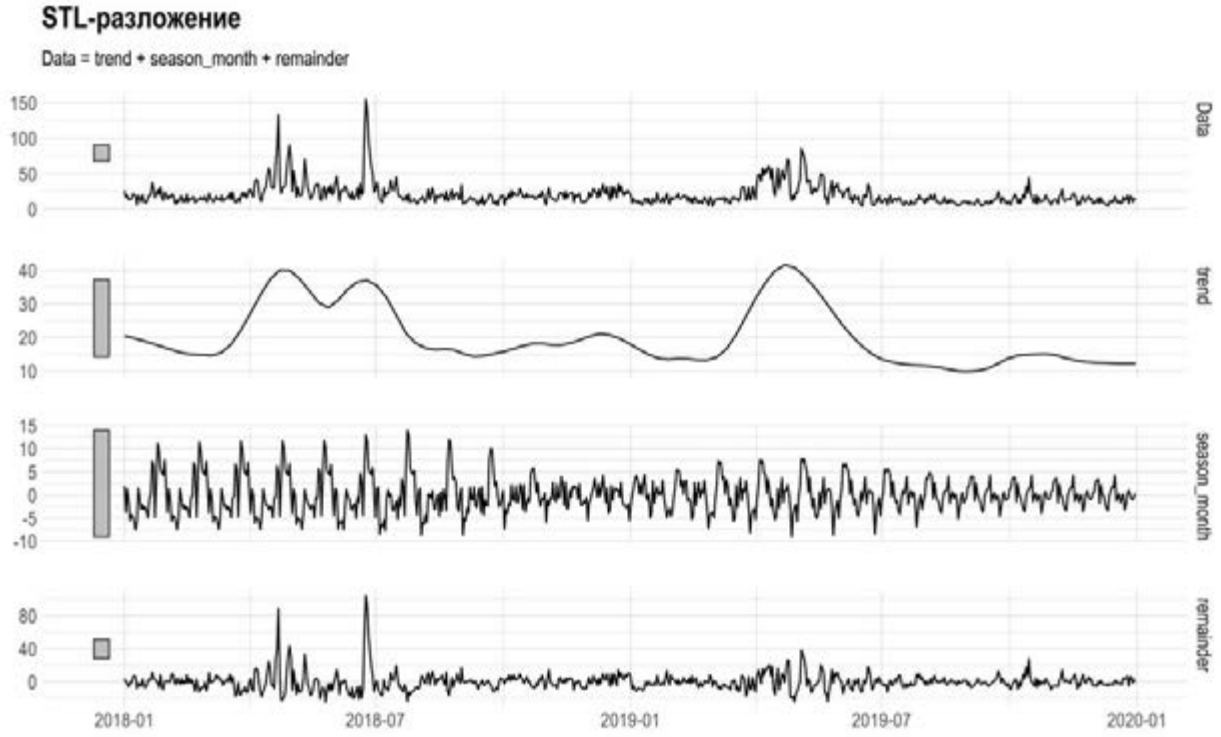

Рис. 4. STL-разложение временного ряда, соответствующего количеству пожаров в городской черте в Красноярском крае в 2018-2019 годах
Здесь: Data – исходные данные, trend – тренд, season_month – недельная сезонность, remainder – остатки.

Кратко остановимся на работе с временными рядами в актуальной библиотеке Prophet, разработанной лаборатории Facebook Open Source. (Библиотека доступна на: https://facebook.github.io/prophet/ для R и Python.) В алгоритме Prophet сезонная составляющая использует разложение в ряд Фурье, также, помимо нелинейного тренда и остатков, в аддитивную модель входит компонента, отвечающая за влияние выходных праздничных дней, а основной метод состоит в подгонке обобщенной аддитивной модели. В качестве примера, рассмотрим временной ряд, соответствующий количеству пожаров в Новосибирской области с 2016 по начало 2020 года, в который не входят объекты горения из категории трава и мусор (вне здания). Отдельные компоненты модели изображены на рис. 5.

Из рис. 5 видно, что модель имеет возрастающий тренд, есть годовые колебания, особенно выраженные в конце апреля – начале мая и осенью. Особый интерес представляет средняя часть графика (weekly), которая показывает, что максимальное количество пожаров приходится на выходные дни, а минимальное – на середину недели, данный факт верен и для пожаров в Красноярском крае за аналогичный период. Рис. 5 показывает, что рассматриваемый временной ряд имеет скачок, показывающий увеличение количества пожаров в 2019 году. Автоматическое обнаружение точек излома тренда такого характера также можно изучать с помощью библиотеки Prophet. Кроме того, библиотека Prophet позволяет оценить влияние праздничных дней на количество пожаров.

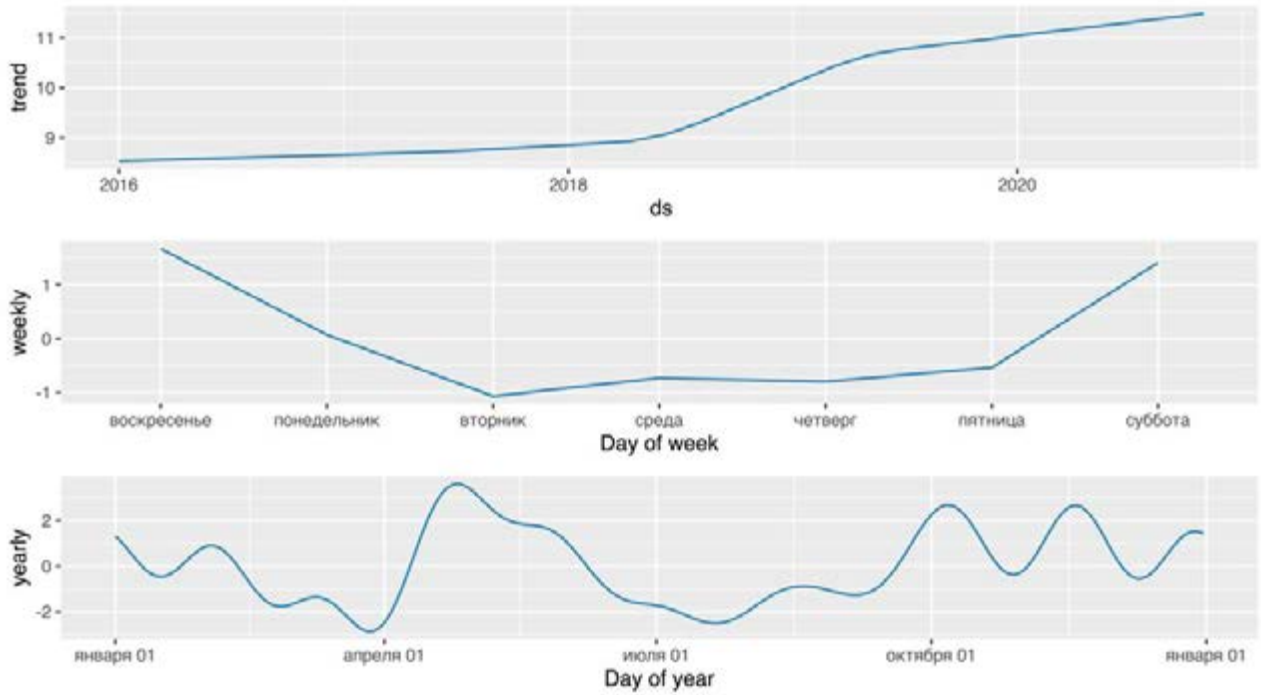

Рис. 5. Пример разложения Prophet-модели на компоненты

В заключении параграфа отметим, что для предварительной оценки временного ряда также важно иметь представление об аномалиях. Работа с аномалиями в R реализована, например, в библиотеке anomalize (https://github.com/business-science/anomalize), которая позволяет выявлять аномалии как в автоматическом режиме, так и используя настройки параметров, см. пример на рис. 6.

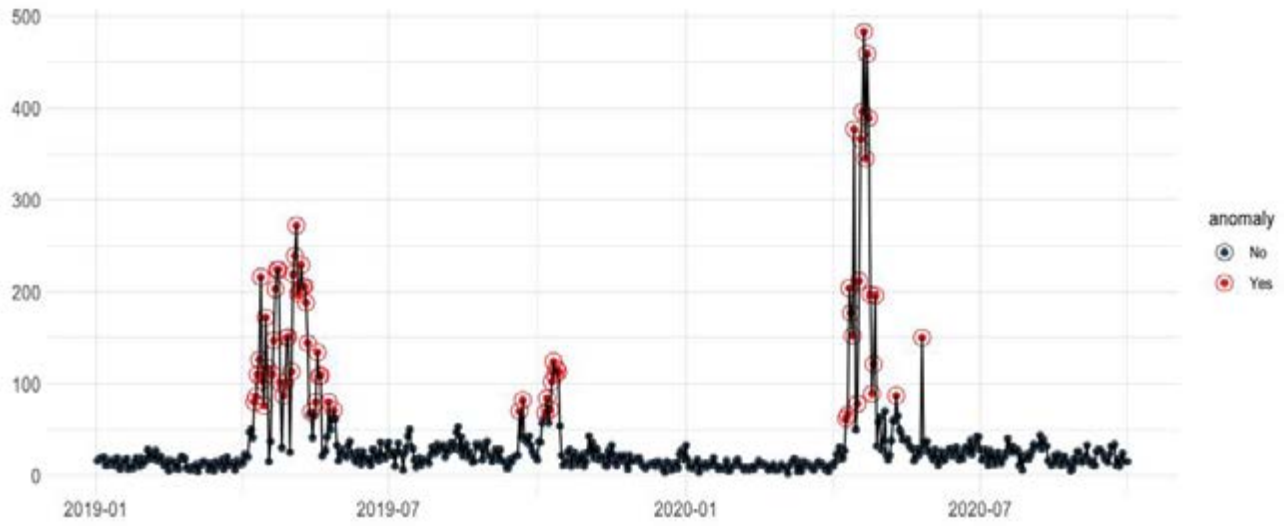

Рис. 6. Аномалии временного ряда, соответствующего количеству пожаров в Новосибирской области в 2019-2020 гг.

3. Использование методов машинного обучения для анализа временных рядов для прогнозирования явлений природной и техносферной безопасности

Как было сказано в начале статьи, язык R использует единую экосистему, основанную на библиотеках, разработанных ведущим специалистом RStudio Хадли Уикемом, ядро которых составляет tidyverse. В основе рабочей идеологии этих библиотек лежит сформулированный Уикемом принцип *tidy data* [5], применяемый для данных, представленных в виде прямоугольных таблиц:

- каждый столбец таблицы данных представляет собой переменную;

- каждая строка представляет собой наблюдение;
- искомое значение находится на пересечении столбца и строки.

Например, набор библиотек tidymodels предназначен для моделирования и машинного обучения с использованием принципов tidyverse, библиотека tsibble предназначена для работы с временными рядами и т.д. Ниже будет показана работа с высоко эффективной библиотекой modeltime (см. https://business-science.github.io/modeltime/), основанной на tidymodels, для моделирования временных рядов с помощью методов машинного обучения.

Весь поток операций в modeltime можно разбить на следующие 6 шагов, позволяющих выполнить:

1. Сбор данных и разделение их на обучающую и тестовую выборки.
2. Создание и подгонку нескольких моделей.
3. Добавление подогнанных моделей в таблицы моделей.
4. Калибровка моделей на тестовое множество.
5. Выполнение прогноза для тестового множества и оценка точности.
6. Корректировку моделей на полный набор данных и прогнозирование на будущие значения.

Кратко покажем реализацию этих шагов. В качестве исходных данных, иллюстрирующих работу в modeltime, используем данные, показывающие максимумы критического уровня воды рек по ближайшему гидрологическому посту в Российской Федерации за 2008-2015 года.

Шаг 1. Разбиение на обучающую и тестовую выборку можно делать указав временной параметр, либо процентные соотношения (см. рис. 7).

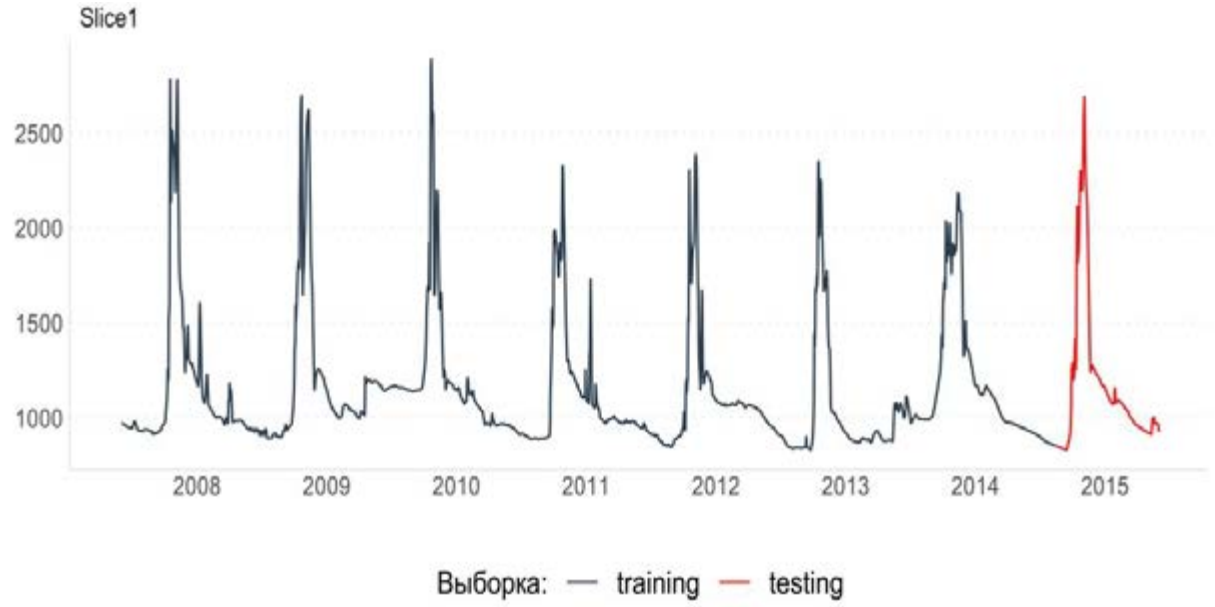

Рис. 7. Пример разделения временного ряда на две выборки: обучающую и тестовую

Шаг *2*. Следующим этапом является создание и подгонка моделей. Ключевая особенность modeltime заключается в возможности работы с несколькими моделями одновременно. Вот некоторые стандартные модели, встроенные в modeltime (полный список моделей можно получить на сайте библиотеки):

- ARIMA;
- линейная регрессия;
- экспоненциальное сглаживание;
- Prophet;
- MARS (*Multivariate Adaptive Regression Splines*);
- Elastic Nets;
- Random Forest.

Отметим, что modeltime позволяет комбинировать алгоритмы, значительно совершенствуя их, например, в модели Prophet можно уменьшить ошибки, используя известный алгоритм машинного обучения XGBoost, что дает новую модель, которая называется Prophet Boost.

Модели машинного обучения более сложны, чем автоматизированные модели. Эта сложность обычно требует рабочего процесса (иногда называемого конвейером в других языках программирования). Общий процесс протекает следующим образом:

- Создание типа модели, так называемого «рецепта» (*recipe*) предварительной обработки используя tidymodels.
- Создание спецификаций модели.
- Использование рабочего процесса для объединения спецификаций модели, предобработки и подходящей модели.

Шаг *3*. Модели прописываются и добавляются в единую таблицу моделей, в которой до включения можно настраивать параметры, а затем проходит их подгонка/масштабирование, проверка на соответствие и калибровка по отношению к тестовой выборке. Далее происходит оценка точности качества моделей на тестовой выборке используя различные показатели точности:

- MAE – средняя абсолютная ошибка;
- MAPE – средняя абсолютная процентная ошибка;
- MASE – средняя абсолютная нормированная ошибка;
- SMAPE – симметричная средняя абсолютная процентная ошибка;
- RMSE – среднеквадратическая ошибка;
- RSQ – показатель R^2.

Шаг *4*. Калибровка, по сути, – это способ определения доверительных интервалов и метрик точности, при этом калибровочные данные – это спрогнозированные значения и невязки, которые вычисляются на основе данных вне выборки.

Шаг *5*. Сформированные модели проверяются на тестовых данных и визуализируются (см. рис. 8).

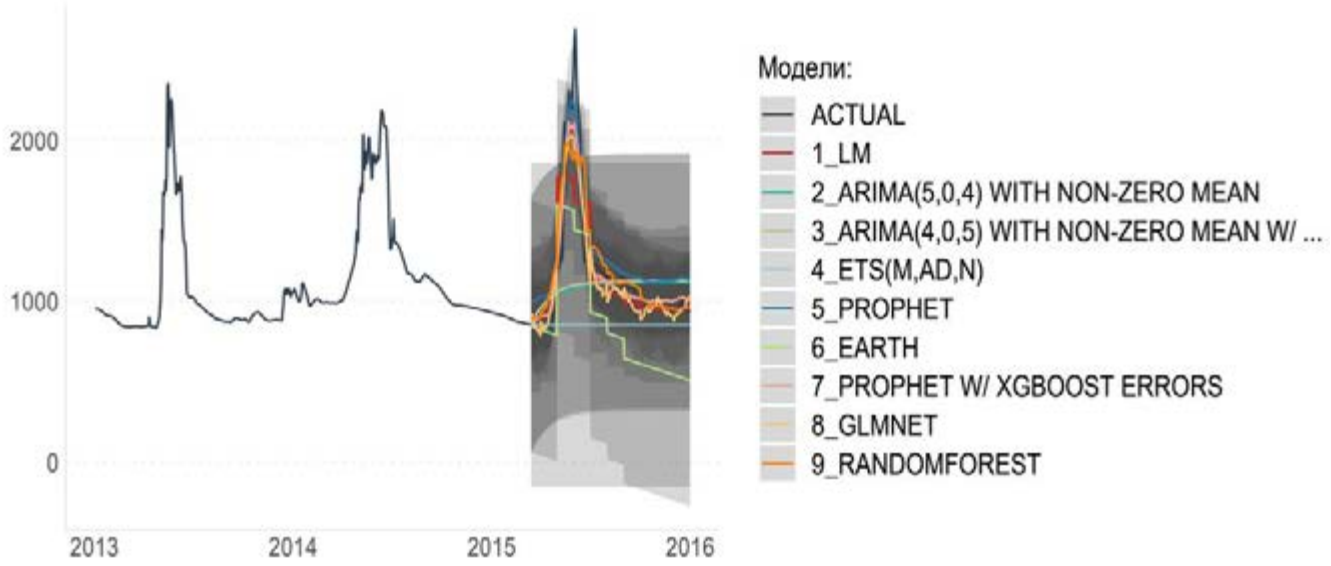

Рис. 8. Пример моделирования временного ряда на тестовую выборку

Также, составляется таблица ошибок, использующая рассмотренные выше показатели точности, пример такого рода таблицы показан ниже.

Таблица 2. Показатели ошибок проверки моделей на тестовом множестве

.model_id	.model_desc	.type	mae	mape	mase	smape	rmse	rsq
1	LM	Test	134.13	9.46	7.12	9.5	232.13	0.68
2	ARIMA(5,0,4) WITH NON-ZERO MEAN	Test	233.95	16.09	12.42	17.72	406.55	0.02
3	ARIMA(4,0,5) WITH NON-ZERO MEAN W/ XGBOOST ERRORS	Test	235.84	16.42	12.52	17.91	403.69	0.02
4	ETS(M,AD,N)	Test	325.55	22.08	17.28	27.24	514.37	0.02
5	EARTH	Test	344.41	29.04	18.29	35.21	399.23	0.61
6	PROPHET	Test	148.64	13.76	7.89	12.76	166.18	0.94
7	PROPHET W/ XGBOOST ERRORS	Test	74.83	5.79	3.97	5.75	117.65	0.94
8	GLMNET	Test	92.95	6.89	4.94	7.02	148.54	0.89
9	RANDOMFOREST	Test	106.79	8.06	5.67	7.7	177.35	0.82

Шаг *6*. Заключительный этап состоит в том, чтобы скорректировать модели, распространить их на полный набор данных и спрогнозировать будущие значения. Как видно из предыдущего шага, не все модели в нашем случае имеют достаточно хорошие показатели ошибок (в частности, показатель R^2 должен быть близок к единице), модели 1-4 и 6 можно удалить из-за низкой точности.

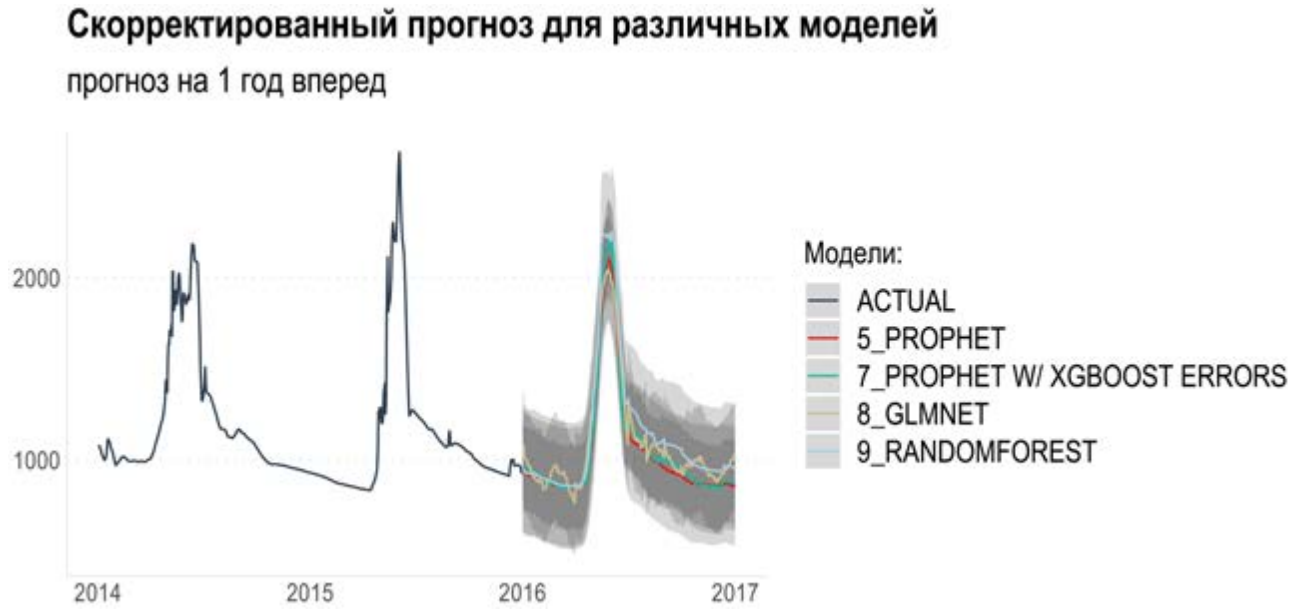

Рис. 9. Пример прогнозирования временного ряда с использованием скорректированных моделей машинного обучения

Отметим, что для улучшения точности прогноза сильно неструктурированных данных, например, при отсутствии явно выраженных сезонных компонент, можно использовать для моделирования нейронные (как правило, RNN, LSTM или CNN) сети, однако это выходит за рамки настоящей статьи, и обучение нейронной сети – процесс гораздо более трудоемкий, чем рассмотренное в работе моделирование, что может оказаться неэффективным с точки зрения временных затрат.

Заключение

В работе были рассмотрены некоторые стандартные задачи в работе с временными рядами: визуализация, поиск аномалий, STL-разложение. Также, в статье были рассмотрены методы прогнозирования временных рядов на основе современных алгоритмов машинного обучения для составления гидрологического прогноза, что публикуется впервые в применении к вопросам природной и техносферной безопасности. Исполь-

зуя возможности языка программирования R можно не только разрабатывать модели прогнозирования, но и в последующем делать на их основе актуальные аналитические веб-сервисы на основе R Markdown и Shiny для практического применения прогнозов, что представляется перспективным направлением повышения эффективности принимаемых решений органами управления подразделениями МЧС России при организации тушения пожаров и ликвидации ЧС.

Литература

1. Матеров, Е.Н. Использование языка программирования R в вопросах пожарной безопасности: обработка и визуализация данных. Сибирский пожарно-спасательный вестник. 2018. №4 (11). С. 60-66.
2. Матеров, Е.Н. Использование языка программирования R в вопросах пожарной безопасности: анализ статистики количества пожаров на основе теории временных рядов. Сибирский пожарно-спасательный вестник. 2019. №2 (12). С. 52-57.
3. Матеров, Е.Н. Использование языка программирования R в вопросах пожарной безопасности: анализ главных компонент. Сибирский пожарно-спасательный вестник. 2019. №1 (13). С. 49-53.
4. Приказ МЧС России от 24 декабря 2018 г. № 625 «О формировании электронных баз данных учета пожаров и их последствий».
5. Уикем, Х. Язык R в задачах науки о данных. Импорт, подготовка, обработка, визуализация и моделирование данных [Текст]. / Уикем Х., Гроулмунд Г. – Вильямс, 2018. Режим доступа: https://r4ds.had.co.nz/, свободный.
6. Hyndman, R. Forecasting: Principles and Practice [Текст]. / Hyndman R., Athanasopoulos G. – OTexts; 2018. Режим доступа: https://otexts.com/fpp2/, свободный.
7. Мастицкий С.Э. Анализ временных рядов с помощью R. [Электронная книга]. URL: https://ranalytics.github.io/tsa-with-r (дата обращения: 12.04.2020).
8. Бабёнышев, С.В. Математические методы и информационные технологии в научных исследованиях [Текст]: учебное пособие / С.В. Бабёнышев, Е.Н. Матеров – Железногорск: ФГБОУ ВО Сибирская пожарно-спасательная академия ГПС МЧС России, 2018. – 215 с.: ил.