

Научная статья

УДК 630*43, 004.8

doi: 10.34987/vestnik.sibpsa.2023.81.32.016

Ретроспективный анализ лесных пожаров в Красноярском крае с использованием науки о данных и методов машинного обучения

Сергей Валерьевич Бабёнышев¹

Евгений Николаевич Матеров²

Татьяна Анатольевна Миловидова³

Сибирская пожарно-спасательная академия ГПС МЧС России, Железногорск, Россия

¹ <https://orcid.org/0000-0002-4298-2036>

² <https://orcid.org/0000-0002-3916-0613>

³ <https://orcid.org/0000-0003-0513-2881>

Автор ответственный за переписку: Евгений Николаевич Матеров, materov@gmail.com

Аннотация. В статье анализируются данные о лесных пожарах в Красноярском крае за период 2011-2023 гг. с учетом их географического распределения. В работе рассмотрены: (1) разведочный анализ данных о пожарах, включая антропогенное воздействие на пожарную активность и изменений в пожарах с течением лет; (2) построение модели машинного обучения, которая предсказывает, является ли термоточка критической, т.е. находится ближе 5 км от населенного пункта или площадь пожара более 200 га. Новизна работы заключается в эффективном использовании языка программирования R, наделенного расширенными возможностями для анализа данных, моделирования методами машинного обучения и визуализации. Использование методов обработки данных, рассмотренных в статье, может повысить точность обнаружения и прогнозирования изменений пожарной активности, последующего мониторинга и крупномасштабного управления пожаротушением. Подходы, разобранные в статье, могут быть применены и к другим регионам Российской Федерации.

Ключевые слова: лесные пожары, машинное обучение, моделирование

Для цитирования: Бабёнышев С. В., Матеров Е. Н., Миловидова Т. А. Ретроспективный анализ лесных пожаров в Красноярском крае с использованием науки о данных и методов машинного обучения // Сибирский пожарно-спасательный вестник. 2023. № 4 (31). С. 148-159. <https://doi.org/10.34987/vestnik.sibpsa.2023.81.32.016>.

Благодарности: авторы выражают глубокую благодарность Валерию Васильевичу Ничепорчуку, старшему научному сотруднику Института вычислительного моделирования СО РАН, за предоставленные данные, поддержку и полезные обсуждения.

Original article

RETROSPECTIVE ANALYSIS OF WILDFIRES IN THE KRASNOYARSK TERRITORY BY MEANS OF DATA SCIENCE AND MACHINE LEARNING METHODS

Sergey V. Babenyshev¹

Evgeny N. Materov²

Tatyana A. Milovidova³

Siberian Fire and Rescue Academy EMERCOM of Russia, Zheleznogorsk, Russia

¹ <https://orcid.org/0000-0002-4298-2036>

² <https://orcid.org/0000-0002-3916-0613>

³ <https://orcid.org/0000-0003-0513-2881>

Corresponding author: Evgeny Materov, materov@gmail.com

Abstract. The article analyzes data on wildfires in the Krasnoyarsk Territory for the period 2011-2023, taking into account their geographical distribution. The article examines: (1) exploratory analysis of fire data, including anthropogenic impact on fire activity and changes in fires over the years; (2) building a machine learning model that predicts whether the heat point is critical, i.e., it is located closer than 5 km from a populated area or the fire area is more than 200 hectares. The novelty of the work lies in the effective use of the R programming language, endowed with advanced capabilities for data analysis, modeling by machine learning and visualization methods. The use of data processing methods discussed in the article can improve the accuracy of detecting and predicting changes in fire activity, subsequent monitoring and large-scale firefighting management. The approaches discussed in the article can be applied to other regions of the Russian Federation.

Keywords: wildfires, machine learning, modeling

For citation: Babenyshev S.V., Materov E.N., Milovidova T.A. Retrospective analysis of wildfires in the Krasnoyarsk Territory by means of data science and machine learning methods // Siberian Fire and Rescue Bulletin. 2023. № 4 (31): 148-159. (In Russ.). <https://doi.org/10.34987/vestnik.sibpsa.2023.81.32.016>.

Acknowledgments: the authors express their deep gratitude to Valery Vasilyevich Nicheporchuk, Senior Researcher at the Institute of Computational Modeling SB RAS, for the data provided, support and useful discussions.

1. Введение

Как явление, лесные пожары являются важным компонентом экосистемы, но также могут привести к значительным экономическим потерям, серьезному загрязнению воздуха, человеческой смертности и ущербу окружающей среде. Ежегодно в России происходят десятки тысяч лесных пожаров, ущерб от которых исчисляется миллиардами рублей [1]. Обработка исторических данных с целью выявления тенденций и аномалий играет важную роль в предотвращении пожаров. Современные подходы в области информационных технологий позволяют внедрять передовые алгоритмы для прогнозирования распространения лесных пожаров. Существует также необходимость перехода в МЧС России на свободно распространяемые программные инструменты, в частности, основанные на экосистемах R и Python.

В то время как потребности служб пожаротушения требуют краткосрочной и среднесрочной оценки пожарного риска, научные исследования также фокусируются на долгосрочных изменениях условий, ведущих к увеличению риска лесных пожаров. Актуальность этих результатов становится решающей для надежности практических прогностических моделей из-за быстрых изменений базовых условий. Авторы статьи [2] рассматривают результаты геопространственного анализа, основанного на долгосрочных спутниковых наблюдениях, и указывают факторы, способствующие повышению риска возникновения лесных пожаров. Другой проблемой прогнозного моделирования является преобладающее влияние антропогенного фактора, что приводит к ситуации, когда большинство пожаров происходит в районах проживания и деятельности человека. Хотя эти пожары могут оказывать наибольшее прямое экономическое воздействие и были тщательно изучены, в том числе методами спутникового мониторинга (см., например, [3]), лесные пожары, которые возникают вдали от человеческого жилья, имеют тенденцию к росту и наносят большой экологический ущерб.

2. Методы исследования

В статье изучаются природные пожары с учетом географического распределения в Красноярском крае с использованием современных программных средств. В задачи данной статьи входят: общий анализ данных о лесных пожарах за период 2011-2023 года, включая географическое расположение пожаров, динамику их количественных показателей во времени и моделирование географически распределенных данных о пожарах на основе алгоритмов машинного обучения. В статье были использованы данные Информационной системы дистанционного мониторинга Федерального агентства лесного хозяйства [4] о лесных пожарах в Красноярском крае.

Для обработки, визуализации и моделирования географических данных мы используем свободно распространяемый язык программирования R. Язык R [5] изначально задумывался его авторами как язык программирования для задач математической статистики, но сейчас, благодаря наличию огромного количества библиотек, расширяющих базовые возможности языка, он используется для самых разных задач и хорошо зарекомендовал себя в применении к географическим данным: см., например, [6] или [7]. Одним из главных преимуществ использования языка R является то, что это свободное программное обеспечение, хорошо конкурирующее с платными продуктами. Альтернативой языку R для автоматизированной работы с географическими данными является язык программирования Python (см. [8]).

Ядром современных средств языка R для задач Data Science является коллекция библиотек *tidyverse* [9], объединенных общей структурой представления данных и принципами проектирования. Библиотеки, включенные в *tidyverse*, имеют интуитивно понятный интерфейс и общую грамматику, удобную для обработки и представления данных. Основными библиотеками для работы с географическими данными, используемыми в работе, являются *sf* и *terra*. Мы использовали фреймворк *tidymodels* для работы с моделями машинного обучения [10], который представляет собой интегрированный, модульный, расширяемый набор библиотек, облегчающий создание прогнозирующих моделей, основанных на принципах машинного обучения.

Статью можно рассматривать как продолжение работ [11, 12], где были рассмотрены возможности R в применении к географическим расчетам и ситуациям, связанных с пожарами.

3. Результаты

3.1 Разведочный анализ данных

Набор исторических данных о пожарах в Красноярском крае, использованный в статье, включает 21 696 записей, временной диапазон данных – с 2011 по 2023 год. Данные за 2023 год представлены по 8 мая. Наибольшее количество термоточек, соответствующих пожарам, было зафиксировано в 2019 году (3 517 случаев), наименьшее – в 2017 году (872 случая). Табличные данные базы данных содержат следующие наблюдаемые значения:

- географические координаты;
- расстояние до ближайшего населенного пункта;
- дата первого наблюдения;
- дата ликвидации пожара;
- лесное хозяйство;
- площадь, охваченная пожаром, гектары (всего).

Используя эти данные, легко рассчитать производные характеристики, например, продолжительность горения термической точки.

Наибольший интерес для исследования представляет географическое распределение пожаров и динамика их распространения с течением времени. Сначала мы нанесем на карту все термоточки, соответствующие лесным пожарам в Красноярском крае за исследуемый период, в виде точек на карте (см. Рис. 1).

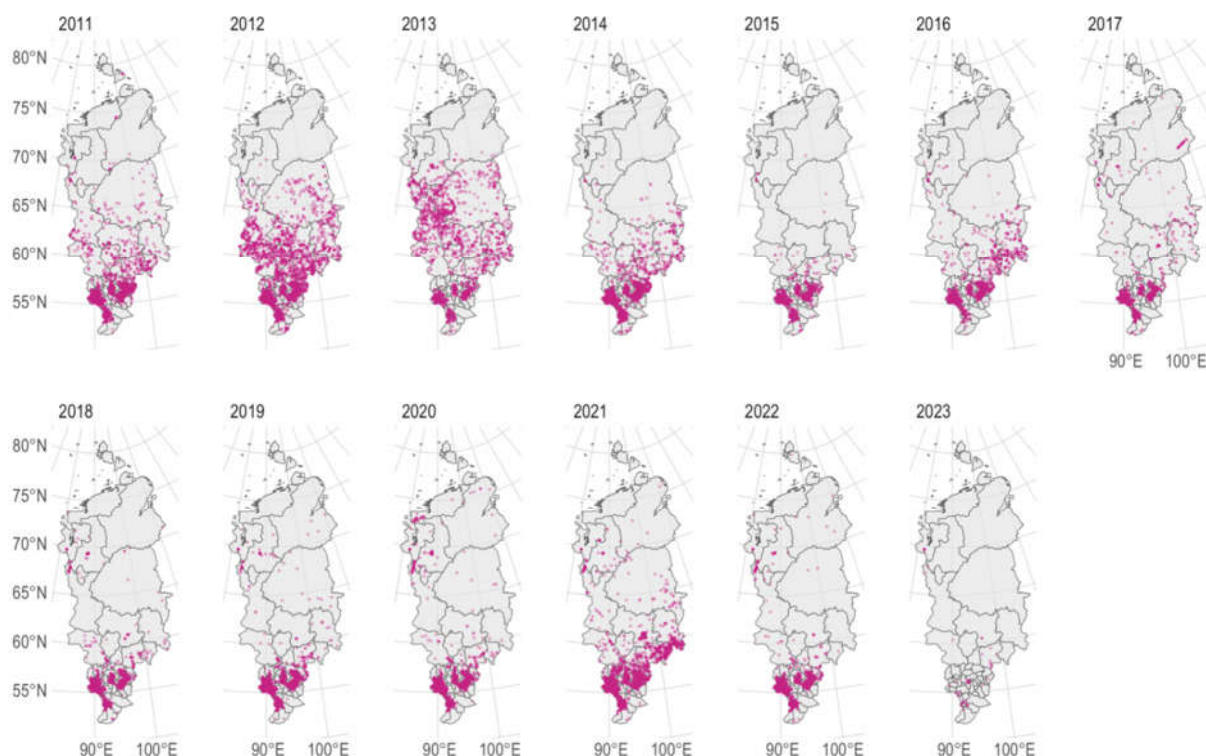


Рис. 1: Географическая карта: лесные пожары в Красноярском крае с 2011 по 2023 гг.

Для выявления влияния антропогенного фактора на количество пожаров на карту пожаров были нанесены населенные пункты края (Рис. 2). На рисунке хорошо видно, что пожары в южных районах происходили в основном вблизи населенных пунктов. Исследование показывает, что максимум плотности по количеству пожаров соответствует географическим координатам 55°40'57" с.ш. 90°10'26" в.д., что находится между населенными пунктами Красная Сопка и Средняя Берёзовка (ближайший крупный населенный пункт – Шарыпово).

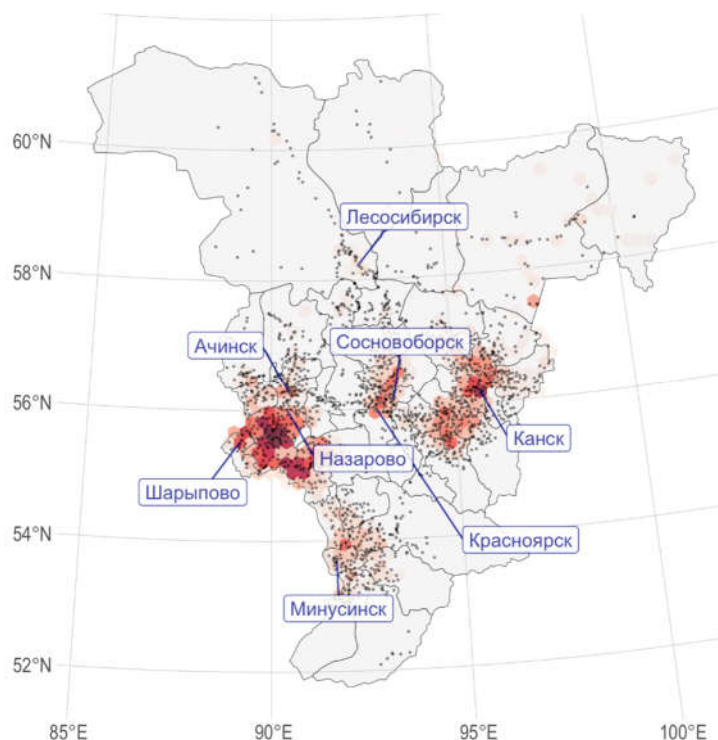


Рис. 2: Географическая карта: плотности лесных пожаров (чем темнее, тем больше пожаров) на юге Красноярского края с 2011 по 2023 гг. Точки соответствуют населенным пунктам края, крупные города региона подписаны

Отметим также, что чем севернее термическая точка, тем больше расстояние от термоточки до населенного пункта, что показывает Рис. 3.

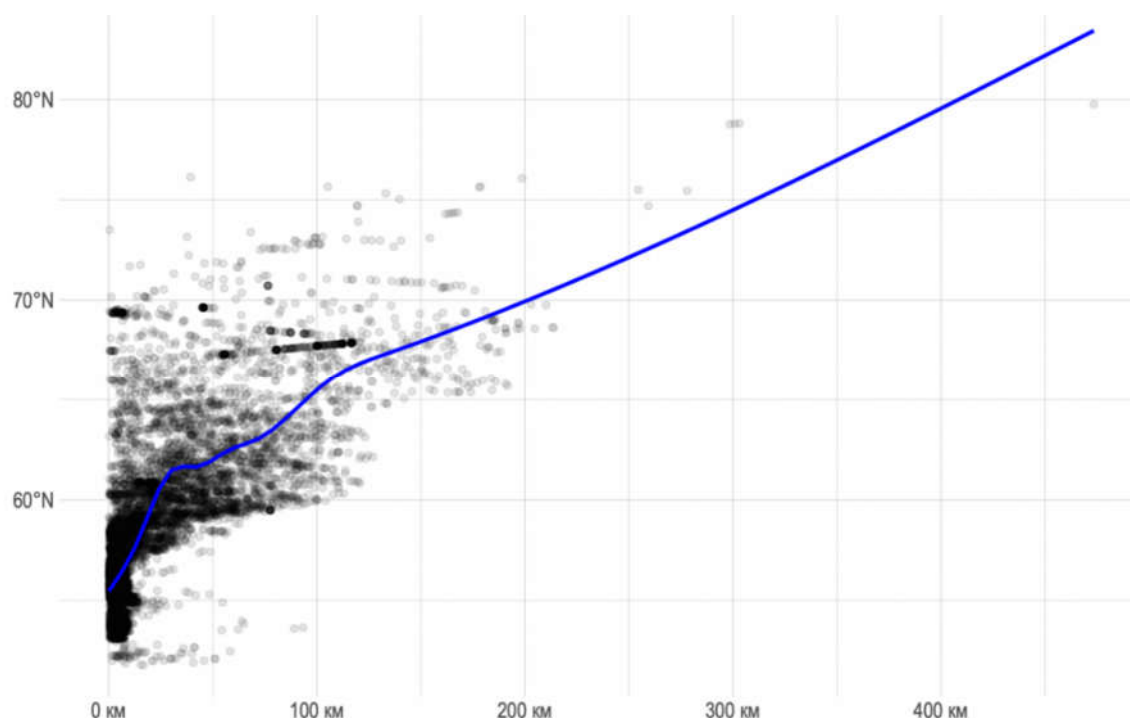


Рис. 3. Расстояние от фронта пожаров до населенных пунктов в Красноярском крае за период 2011-2023 гг. Каждая точка на графике представляет собой термическую точку, кривая синего цвета представляет собой сглаживание методом подгонки локальной полиномиальной регрессии алгоритмом LOWESS

3.2. Моделирование с помощью методов машинного обучения

Математические модели – это научные инструменты, которые могут описывать исследуемую систему и фиксировать взаимосвязи в предоставляемых им данных. Машинное обучение (ML) – это определенный класс математических методов и алгоритмов, который позволяет делать выводы на основе итеративного использования данных в процессе обучения, т.е. за счет поиска закономерностей определенного типа. Мы будем использовать подход, основанный на машинном обучении, для моделирования географически распределенных данных.

Основные расчеты в работе были выполнены с использованием *tidymodels*. Библиотеки *tidymodels*, используемые для ML-моделирования, основаны на принципах *tidy data* [13] и объединены единым подходом и грамматикой. Хорошим источником для введения в моделирование в R является книга [14].

В работе применяются методы машинного обучения к изучению лесных пожаров в Красноярском крае. Согласно приказу «Об утверждении Правил тушения лесных пожаров» [15] (вступил в силу с 1 марта 2023 г. и действует до 1 марта 2029 г.) решение о привлечении сил и средств единой государственной системы предупреждения и ликвидации чрезвычайной ситуации (РСЧС) муниципального (соответственно, регионального) уровня может быть принято при соблюдении не менее двух условий, в частности среди перечисленных условий:

- площадь лесного пожара – 25 Га и более в районе применения наземных сил и средств, 200 Га и более в районе применения авиационных сил и средств;
- лесной пожар действует в 5-километровой зоне от границы населенного пункта или объекта экономики, особо охраняемой природной территории (ООПТ).

Рассмотрим классификационную модель, определяющую в зависимости от остальных параметров, является ли точка критической (в нашем случае, ближе 5 км от населенного пункта или площадь пожара более 200 Га, это около 67% от всех данных), см. Рис. 4. Для моделирования мы будем использовать дополнительные данные: высоту в точке пожара над уровнем моря и растительный покров.

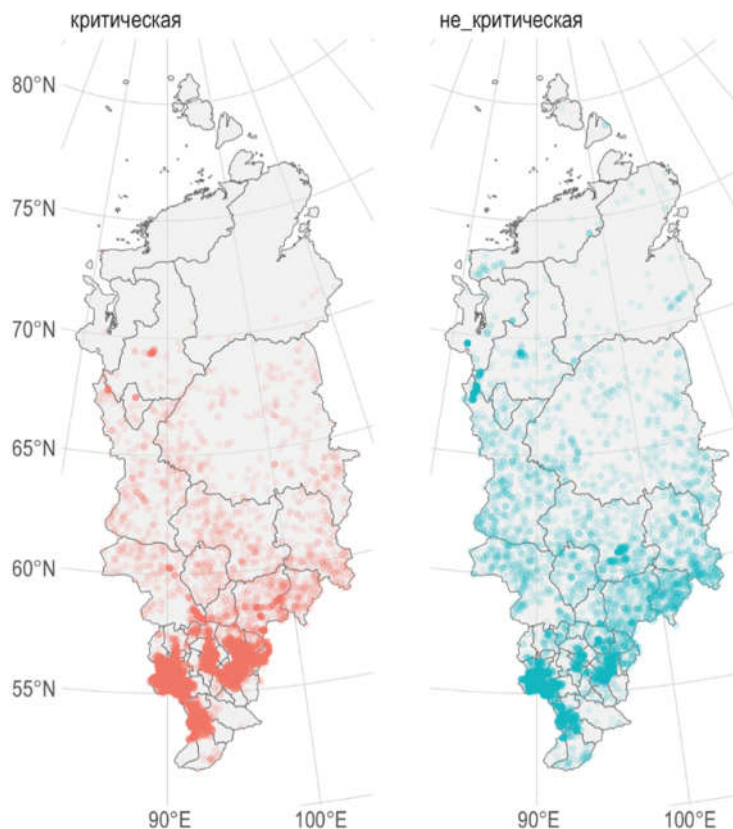


Рис. 4. Географическая карта: разделение точек с пожарами на критические и не критические

3.2.1 Разделение набора данных на обучающую и тестовую выборки

Рассмотрим основные этапы моделирования в *tidymodels*. Для этого разделим в первую очередь весь набор данных на две категории. Первый набор – так называемый *обучающий набор* составляет большую часть данных, используемых для разработки и оптимизации модели, и второй – *тестовый набор* для определения эффективности модели (в нашем случае в пропорции 75%/25%). При выборе этих наборов могут возникнуть следующие трудности: выборка может быть небольшой, значения могут быть непропорционально смещены относительно среднего значения, что может увеличить риск неэффективности прогнозирования. Для решения такого рода проблем *tidymodels* выполняет различные функции, например, проводится стратификация, чтобы снизить риск смещения. Стратифицированная случайная выборка выполнит разделение в каждом из этих подмножеств данных, а затем объединит результаты вместе.

Для повторной выборки пространственных данных в R можно использовать экспериментальную библиотеку *spatialsample* [16], которая позволяет формировать кластеры на основе широты и долготы на этапе разделения географических данных на обучающие и тестовые наборы, что в некоторых случаях может улучшить результат моделирования. В библиотеке *spatialsample* мы разделим весь набор на 7 кластеров по географическим признакам (см. Рис. 5).

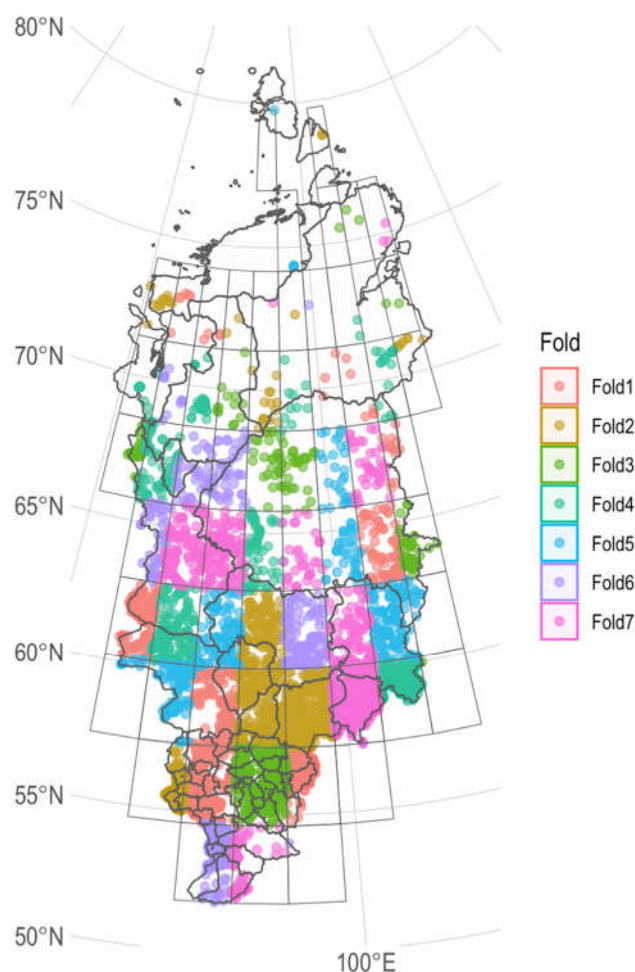


Рис. 5. Географическая карта: разделение точек кросс-валидации на блоки на основе библиотеки *spatialsample*

3.2.2 Настройка гиперпараметров модели и оценка результатов с помощью метрик

Данный шаг включает в себя преобразование значений предикторов, чтобы упростить их эффективное использование в модели для лучшего представления их важных характеристик, например, сокращение выборки для учета дисбаланса классов и т.д. Некоторые модели используют геометрические показатели расстояния, и, следовательно, числовые предикторы должны быть отцентрированы и масштабированы таким образом, чтобы все они были в одних и тех же единицах измерения. После этого выбирается модель: *tidymodels* предоставляет огромный выбор практически всех известных моделей [17], включая модели глубокого обучения (например, *TensorFlow* или *Torch*). Удобство фреймворка *tidymodels* заключается в том, что мы можем создать несколько моделей на основе одного и того же набора данных и затем сравнить их. Модель рассматривается на обучающей выборке, а таблица метрик проверяет, насколько хорошо выглядит результат моделирования (например, показатель метрики должен быть близок к 1 и т.д.).

Мы построим модель, основанную на алгоритме бэггинга (от англ. *bootstrap aggregating*) из библиотеки *baguette*. Идея бэггинга заключается в построении деревьев решений на основе обучающих *bootstrap*-выборок и дальнейшем усреднении коллективного прогноза, что позволяет уменьшить дисперсию прогноза [18].

3.2.3 Распространение модели на все множество и проверка точности

При распространении модели на тестовую выборку показатели пересчитываются заново и рассматривается матрица ошибок, показывающая точность модели. Кроме того, точность

моделирования может быть оценена с помощью так называемой *ROC*-кривой. Наконец, если модель показывает удовлетворительные результаты, учитывается значимость особенностей модели. После подгонки модели (распространения на весь набор) проверка метрики для нашей модели показывает результаты с точностью = 0,85 (доля данных, которые были предсказаны правильно), см. Рис. 6.

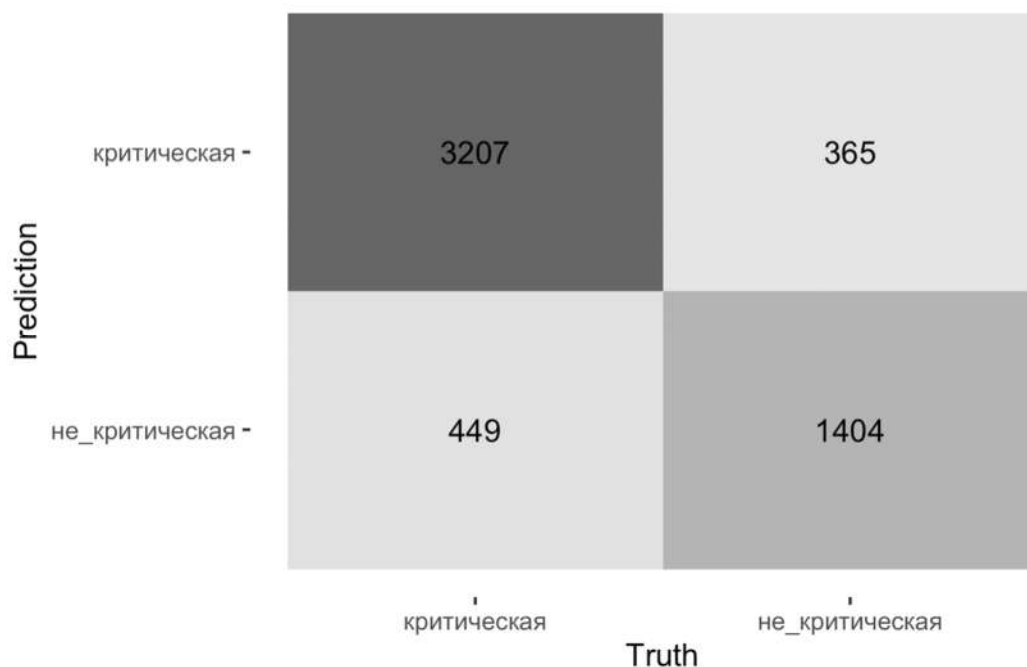


Рис. 6. Матрица ошибок модели в виде тепловой карты

Показатель $ROC_{auc} = 0,916$ (площадь под *ROC*-кривой) близок к 1, что также указывает на хорошую эффективность модели. Можно видеть, что наибольшую значимость в модели (см. Рис. 7) имеет широта (см. также Рис. 3).

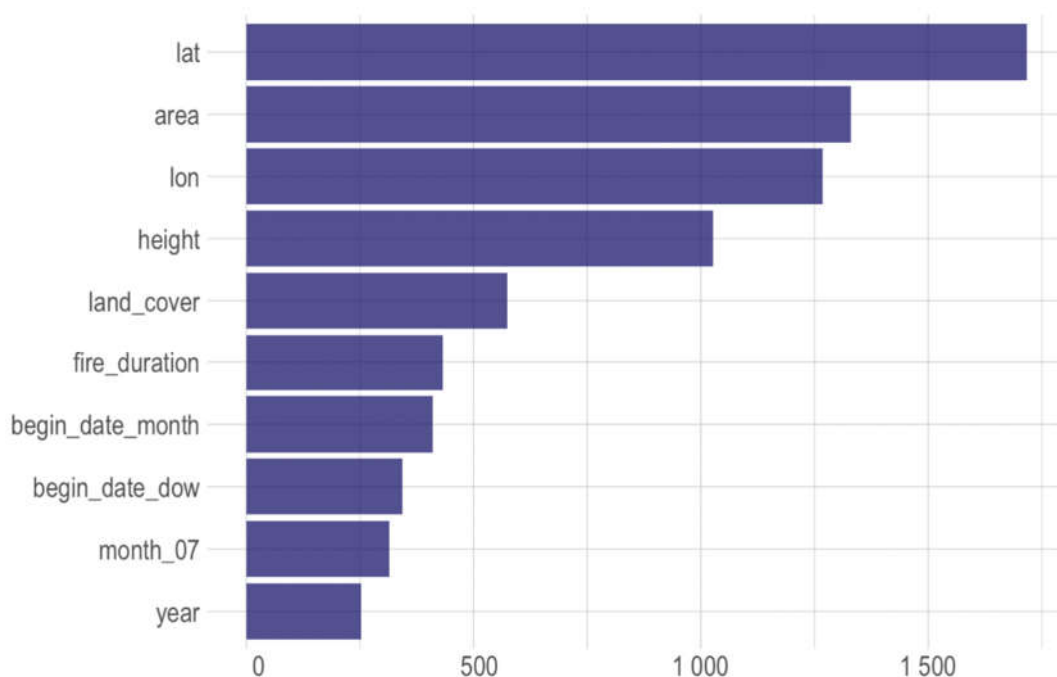


Рис. 7. Значимость предикторов модели

С помощью гексагональной карты покажем процент правильно классифицированных пожаров. Для наглядности на карту нанесены населенные пункты (см. Рис. 8).

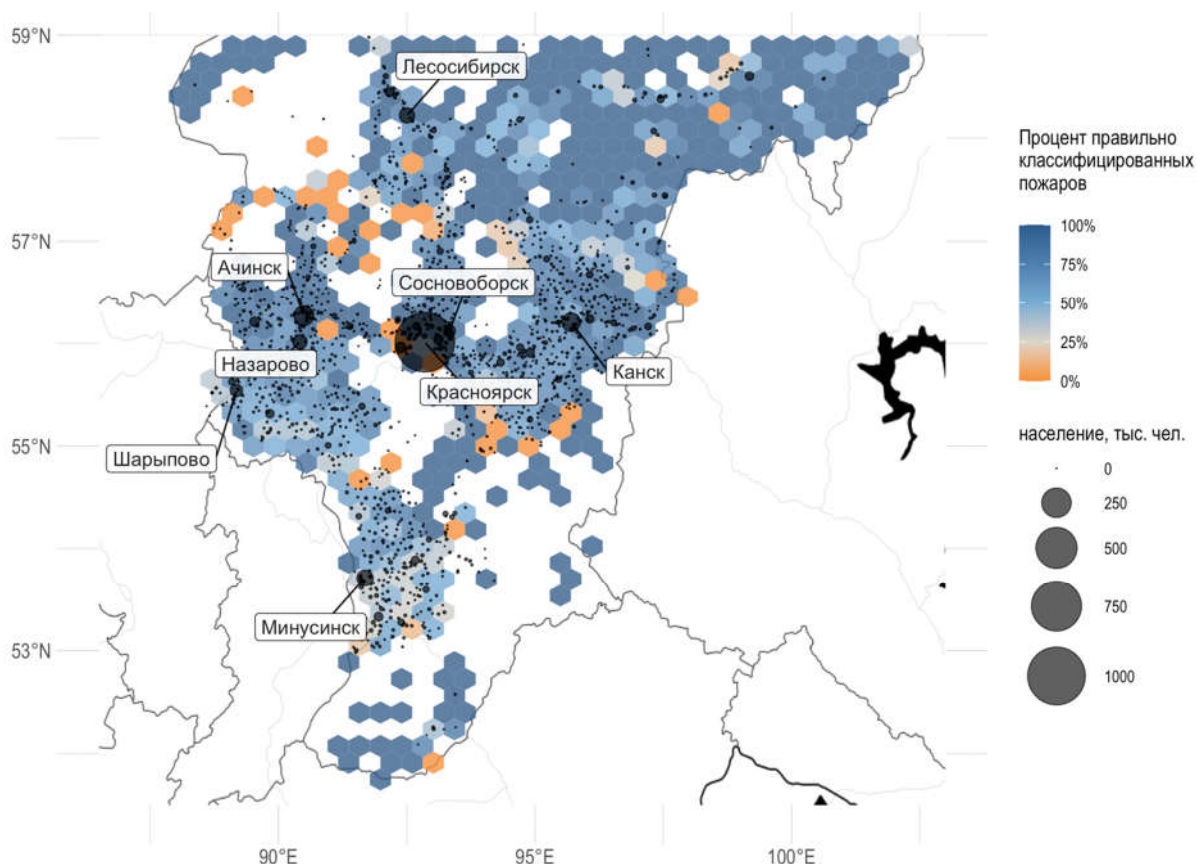


Рис. 8. Гексагональная карта, показывающая процент правильно классифицированных пожаров по принадлежности к критической зоне (разделение цвета сделано на 25%), на карту также нанесены населенные пункты

Точки на карте, соответствующие населенным пунктам, в основном лежат в диапазоне 75%–100% правильной классификации, территории около крупных населенных пунктов классифицируются хорошо.

Отметим, что для дальнейшего взаимодействия с моделью в R можно легко развернуть соответствующий API и опубликовать его с помощью библиотек *vetiver* и *plumber*, а также осуществить дальнейший мониторинг с помощью библиотеки *pins*.

4. Заключение

В статье были рассмотрены возможности языка программирования R в применении к задачам, возникающим при обработке, визуализации и моделировании геопространственных параметров природных пожаров в Красноярском крае. В частности, анализ базовых данных и нанесение данных на различные карты; анализ динамики количественных показателей пожаров по годам; был приведен пример классификационной модели для определения критических термоточек, использующей методы машинного обучения.

Рассмотренные в статье инструменты и модели имеют как теоретическое, так и практическое значение, например, для определения рисков нелокализованных пожаров, приближающихся к критической 5-километровой зоне на основе оперативных данных вокруг населенного пункта или объекта инфраструктуры. Используемые здесь подходы могут быть аналогичным образом распространены на другие регионы Российской Федерации. Возможно продолжить работу над построением моделей, учитывающих погодные условия, глобальные

географические и временные закономерности пожарной активности, отражающие взаимодействие климата, растительности и факторов воспламенения.

Список источников

1. ЕМИСС. Официальные статистические показатели [Электронный ресурс]. – Режим доступа: <https://fedstat.ru/>
2. Пономарев Е.И., Харук В.И., Якимов Н.Д. Результаты и перспективы спутникового мониторинга природных пожаров // Сибирский лесной журнал. 2017. № 5. С. 25–36.
3. Волокитина А.В., Софронова Т.М., Корец М.А. Региональные шкалы оценки пожарной опасности в лесу: усовершенствованная методика составления // Сибирский лесной журнал. 2017. № 2. С. 52–61.
4. ИСДМ-Рослесхоз. Информационная система дистанционного мониторинга Федерального агентства лесного хозяйства [Электронный ресурс]. – Режим доступа: https://nffc.aviales.ru/main_pages/index.shtml
5. The R Project for Statistical Computing [Электронный ресурс]. – Режим доступа: <https://www.r-project.org/>
6. Robin Lovelace, Jakub Nowosad, Jannes Muenchow. Geocomputation with R. Chapman; Hall/CRC, 2019.
7. Самсонов Т.Е. Визуализация и анализ географических данных на языке R. М.: Географический факультет МГУ, 2022.
8. Sergio J. Rey, Dani Arribas-Bel, Levi John Wolf. Geographic Data Science with Python. Chapman; Hall/CRC, 2023.
9. Tidyverse. The R packages for data science [Электронный ресурс]. – Режим доступа: <https://www.tidyverse.org/>
10. Tidymodels: the collection of packages for modeling and machine learning using tidyverse principles [Электронный ресурс]. – Режим доступа: <https://www.tidymodels.org/>
11. Бабёнышев С.В. и др. Анализ пожарной и техносферной безопасности городской среды г. Красноярска с применением геопространственных инструментов // Сибирский пожарно-спасательный вестник. 2021. № 4. С. 128–136.
12. Бабёнышев С.В. и др. Применение геоинформационных инструментов для работы с большими данными при анализе пространственного распределения пожаров // Сибирский пожарно-спасательный вестник. 2021. № 2. С. 70–77.
13. Wickham H. и др. R for Data Science, 2nd edition. 2023.
14. Kuhn M., Silge J. Tidy Modeling with R: A Framework for Modeling in the Tidyverse. O'Reilly Media, 2022.
15. Об утверждении Правил тушения лесных пожаров: приказ Министерства природных ресурсов и экологии РФ от 1 апреля 2022 г. № 244 [Электронный ресурс]. 2022. – Режим доступа: <https://www.garant.ru/products/ipo/prime/doc/405045151/>
16. Spatialsample: a library to provide building blocks for creating and analyzing resamples of a spatial data [Электронный ресурс]. – Режим доступа: <https://spatialsample.tidymodels.org/>
17. Parsnip: Search parsnip models [Электронный ресурс]. – Режим доступа: <https://www.tidymodels.org/find/parsnip/>
18. Шитиков В.К., Мاستицкий С.Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 2017.

List of sources

7. EMISS. Official statistical indicators [Electronic resource]. – Access mode: <https://fedstat.ru/>
8. Ponomarev E.I., Kharuk V.I., Yakimov N.D. Current Results and Perspectives of Wildfire Satellite Monitoring in Siberia // Sibirskij Lesnoj Zurnal (Sib. J. For. Sci.), No. 5, 25-36, 2017 (in Russian with English abstract).

9. Volokitina, A.V., Sofronova, T.M., Korets M.A. Regional Forest fire hazard scales: improved method // *Sibirskij Lesnoj Zurnal (Sib. J. For. Sci.)*, No. 2, 52–61, 2017 (in Russian with English abstract).
10. ISDM-Rosleskhoz. Wildfires Monitoring Information System of the Federal Forestry Agency [Electronic resource]. – Access mode: https://nffc.aviales.ru/main_pages/index.shtml
11. The R Project for Statistical Computing [Electronic resource]. – Access mode: <https://www.r-project.org/>
12. Robin Lovelace, Jakub Nowosad, Jannes Muenchow. Geocomputation with R. Chapman; Hall/CRC, 2019.
13. Samsonov T.E. Visualization and analysis of geographical data in the R language. Moscow, Russia: Faculty of Geography of Moscow State University, 2022.
14. Sergio J. Rey, Dani Arribas-Bel, Levi John Wolf. Geographic Data Science with Python. Chapman; Hall/CRC, 2023.
15. Tidyverse. The R packages for data science [Electronic resource]. – Access mode: <https://www.tidyverse.org/>
16. Tidymodels: the collection of packages for modeling and machine learning using tidyverse principles [Electronic resource]. – Access mode: <https://www.tidymodels.org/>
17. Babenyshev S.V., Malyutin O.S., Materov E.N., Elfimova M.V. Analysis of fire and technosphere safety of the urban environment of Krasnoyarsk using geospatial tools // *Siberian Fire and Rescue Bulletin*, No. 4, 128–136, 2021 (in Russian).
18. Babenyshev S.V., Boyko G.M., Malyutin O.S., Materov E.N. Application of geoinformation tools for working with big data in the analysis of the spatial distribution of fires // *Siberian Fire and Rescue Bulletin*, No. 2, 70–77, 2021 (in Russian).
19. Wickham H. и др. R for Data Science, 2nd edition. 2023.
20. Kuhn M., Silge J. Tidy Modeling with R: A Framework for Modeling in the Tidyverse. O'Reilly Media, 2022.
21. On the approval of the Rules for extinguishing forest fires: Order of the Ministry of Natural Resources and Ecology of the Russian Federation dated April 1, 2022 № 244 [Electronic resource]. 2022. – Access mode: <https://www.garant.ru/products/ipo/prime/doc/405045151/>
22. Spatialsample: a library to provide building blocks for creating and analyzing resamples of a spatial data [Electronic resource]. – Access mode: <https://spatialsample.tidymodels.org/>
23. Parsnip: Search parsnip models [Electronic resource]. – Access mode: <https://www.tidymodels.org/find/parsnip/>
24. Shitikov V.K., Mastitsky S.E. Classification, regression and other Data Mining algorithms using R. 2017.

Информация об авторах

С.В. Бабёнышев – кандидат физико-математических наук

Е.Н. Матеров – кандидат физико-математических наук

Т.А. Миловидова – кандидат физико-математических наук

Information about the author

S.V. Babenyshev – Ph.D. of Physical and Mathematical Sciences

E.N. Materov – Ph.D. of Physical and Mathematical Sciences

T.A. Milovidova – Ph.D. of Physical and Mathematical Sciences

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Статья поступила в редакция 27.11.2023; одобрена после рецензирования 10.12.2023; принята к публикации 15.12.2023.

The article was submitted 27.11.2023, approved after reviewing 10.12.2023, accepted for publication 15.12.2023