

УДК 004.62, 311.2

Использование языка программирования R в вопросах пожарной безопасности: обработка и визуализация данных

Using the R programming language for fire safety issues: data wrangling and data visualization

Е.Н. Матеров
канд. физ.-мат. наук

ФГБОУ ВО Сибирская
пожарно-спасательная
академия ГПС МЧС России

E.N. Materov
**Ph.D. of Physico-mathematical
Sciences**

FSBEE HE Siberian Fire
and Rescue Academy
EMERCOM of Russia

Аннотация:

В статье сделан обзор языка программирования R, его возможностей для обработки и визуализации табличных данных на примерах, связанных с вопросами пожарной безопасности. Кроме того, показаны основные принципы построения графики средствами R, приведены примеры анализа данных, даны общие рекомендации по перспективным методам компьютерной обработки информации в среде R, показаны преимущества по сравнению с другими средствами обработки данных.

Ключевые слова: язык программирования R, анализ данных, статистика пожаров

Abstract:

The article provides an overview of the R programming language, its capabilities for processing and visualization of tabular data on examples related to fire safety issues. Besides, the basic principles of graphics by means of R are shown, as well as examples of exploratory data analysis, general recommendations on promising methods of computer information processing in the R environment, and advantages in comparison with other means of data processing.

Keywords: R programming language, data analysis, fire statistics.

Анализ обстоятельств возникновения чрезвычайных ситуаций (ЧС), прогноз возможного развития явлений природного и техногенного характера, а также принятие решений по снижению рисков, управлению силами и средствами невозможен без правильного подхода к исследованию данных. В настоящее время, для решения сложных аналитических задач, одной из самых перспективных областей знаний является наука о данных (Data Science) – междисциплинарная область, изучающая проблемы обобщения, анализа, алгоритмизации и представления данных в численной и визуальной формах. Наука о данных опирается на инструменты эмпирических наук, статистики, отчетности, анализа, визуализации, бизнес-аналитики, экспертных систем, машинного обучения, баз данных, хранения данных, интеллектуального анализа данных и больших данных.

Одним из наиболее эффективных современных инструментов Data Science является язык программирования R – активно развивающаяся бесплатная среда для обработки и визуализации данных. Язык R принадлежит к классу объектно-ориентированных высокоуровневых языков программирования. Будучи разработанным в начале 90-х годов, на сегодняшний день R является одним из наиболее используемых средств

для обработки статистических и научных данных. В данной статье дается краткое введение в R, а также обзор возможностей для обработки табличных данных и возможности визуализации данных на примерах, связанных с вопросами пожарной безопасности. Статья имеет обзорный характер и не претендует на глубокий анализ всевозможных методов исследования данных, а имеет целью дать синопсис некоторых методов Data Science с использованием специального программного обеспечения. Вопросы машинного обучения, работы с временными рядами, геоданными и прочие вопросы будут рассмотрены в последующих статьях. Отметим, что ближайшими конкурентами R являются SAS (лицензия на который стоит несколько тысяч долларов) и Python. Что предпочесть в той или иной ситуации: R или Python зависит от каждой конкретной задачи и реализации библиотек.

Кратко отметим следующие отличительные особенности, делающие язык R привлекательным инструментом и одним из лидеров в области свободно распространяемых систем статистического анализа данных.

- R изначально создан специально для статистической обработки данных и их визуализации. Это дает богатые возможности составления форм отчетностей, простоту в реализации отображения данных, высокое качество графики и гибкость, не доступные аналогичным программам.
- R обладает огромным количеством библиотек, которые реализуют большинство известных статистических методов, т.е. обладает огромными возможностями расширяемости. В настоящий момент количество известных пакетов, дополняющих R, только по данным основного репозитория R CRAN <https://cran.r-project.org/web/packages/> насчитывается уже более 13 000.
- R является лидером направления. Какой бы современный статистический метод ни был рассмотрен (алгоритмы, основанные на методах Deep Learning и т.п.), можно быть уверенным – нужный метод либо уже реализован, либо будет очень скоро реализован в R.
- R хорошо документирован – количество изданий по обработке данных и Data Science с использованием R растет с каждым годом. Сейчас издано уже несколько десятков книг (в том числе имеются издания на русском языке), где R является центральным инструментом анализа: от теории временных рядов до биоинформатики, от анализа результатов выборов в политике до исследования молекулярных данных, языко-

ведения и эконометрики. В России по R вышло несколько книг [3, 4, 5], см. также [2].

- R умеет работать с пропущенными данными, геоданными; функции в R векторизованы, что упрощает вычисления.
- Среда R и GUI для использования R является бесплатной, в одинаковой мере поддерживающей все ведущие платформы: Windows, Linux и macOS.

Также следует отметить следующие преимущества и возможности R, делающие его притягательным для анализа данных:

- возможность создания динамических html-отчетов, написанных на RMarkdown или в виде книг на bookdown, а также работа с Git-проектами;
- работа с интерактивной графикой в библиотеках Shiny и Plotly;
- современные возможности машинного обучения (от XGboost и Random Forest до суперсовременных методов глубокого обучения Keras).

Установка R

Установить R можно, перейдя по адресу <https://cran.r-project.org/> и загрузив R для соответствующей платформы. Интерфейс R представляет собой командную строку, поэтому рекомендуется в качестве оболочки R использовать RStudio (<https://www.rstudio.com/>), которая позволяет в удобной форме сохранять R-скрипты, работать с R Markdown для представления отчетов, иметь под рукой открытые библиотеки, получать помощь, просматривать загруженные данные, и многое другое.

Обзор библиотек Tidyverse

Основу современного языка R в применении к Data Science составляет группа библиотек tidyverse, объединенных единой философией, грамматикой и структурой данных (<https://www.tidyverse.org/>). Идеальным вдохновителем tidyverse является новозеландец Хадли Викхэм (Hadley Wickham), книга которого «R for Data Science» (<http://r4ds.had.co.nz/>) является введением в tidyverse. Код, написанный с применением данных библиотек, обладает рядом преимуществ по сравнению с базовым R: он более читаем, легче в отладке и работает на порядок быстрее.

Основные библиотеки tidyverse включают в себя: dplyr для манипуляции с данными, tidyr для изменения формы данных, forcats для работы с категориальными переменными, ggplot2 для визуализации данных.

Визуализация данных средствами ggplot2

Для двумерной визуализации табличных данных служит библиотека ggplot2. Отличительной особенностью ggplot2 является то, что все манипуляции с графиками возможно делать послойно, последовательно выполняя действия с визуализацией, полученной на предыдущем этапе, а также возможности тонких настроек. Графики, полученные с применением ggplot2 обладают высоким полиграфическим качеством с возможностью публикации в научных изданиях. Отметим, что выбор масштаба и легенды происходит автоматически. Графическое представление информации происходит, как правило, как отображение табличных данных на соответствующие оси с использованием геометрии, которая определяет тип графика. Кратко, идеологию соответствия большинства канонических графиков входным типам данных и геометрии (командам ggplot2) можно представить в таблице. Напомним, что здесь категориальные переменные – это те (качественные) переменные, представление которых может быть разделено на группы: номинальные (без естественного упорядочивания) и порядковые данные.

Таблица. Двумерное представление типов данных, базовые виды диаграмм и соответствующие команды библиотеки ggplot2

переменная (x)	отклик (y)	тип графика	ggplot2 geom
численная		гистограмма	geom_histogram()
		график плотности	geom_density()
категорная		столбиковая диаграмма	geom_bar()
численная	численный	диаграмма рассеяния	geom_point()
категорная	численный	диаграмма размаха	geom_boxplot()
категорная	категорный	мозаичная диаграмма	geom_mosaic()

Рассмотрим несколько примеров диаграмм из таблицы выше.

Столбиковая диаграмма. На рис. 1 дан пример столбиковой диаграммы, показывающей количество спасенных на пожарах за 12 месяцев 2017 года по Сибирскому федеральному округу с разбиением по субъектам.

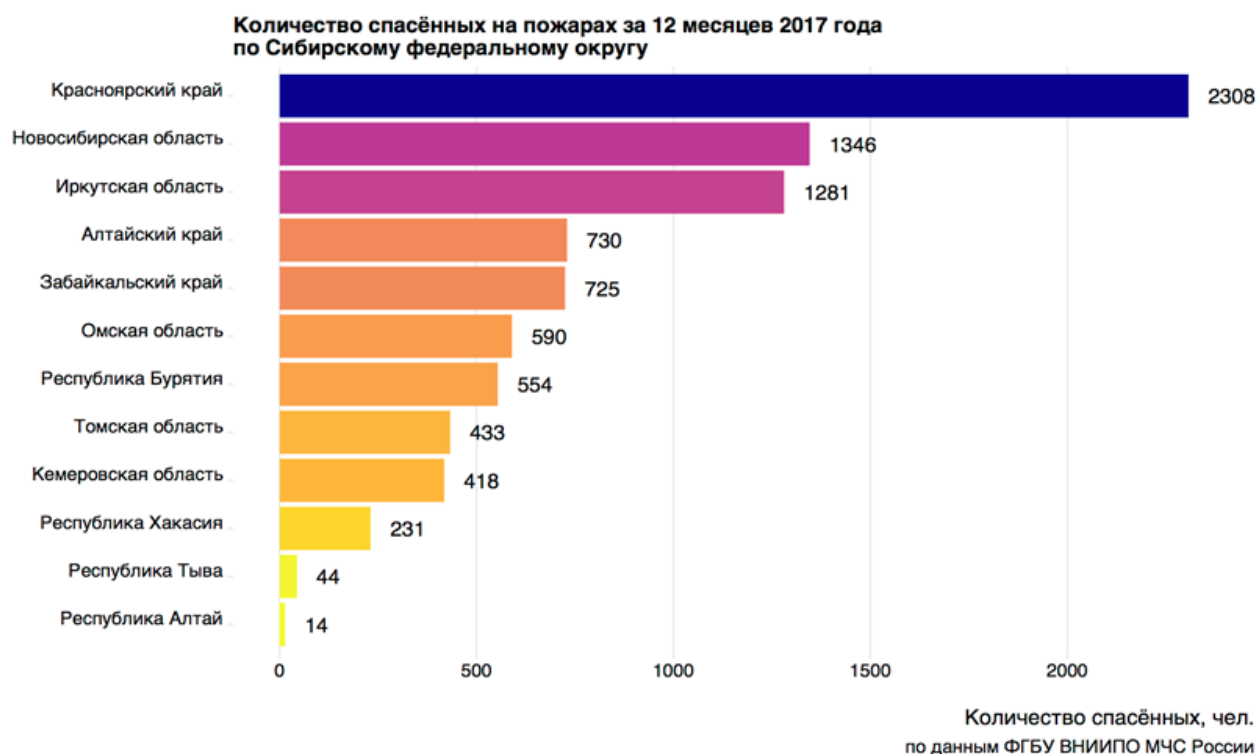


Рис. 1. Пример столбиковой диаграммы

Отметим, что в данном примере для наглядности был использован градиентный цвет столбцов, который позволяет удобно визуальнo кластеризовать данные по смежным значениям.

В случае большого количества категорий рекомендуется использовать так называемые диаграммы

Кливленда, введенные в статье [6], у которых данные группируются и сортируются в соответствии с уровнями, задаваемыми переменной. Часто группы выделяются своим цветом, а если внутри маркеров проставляются значения, то такие диаграммы называют леденцовыми как на рис. 2.

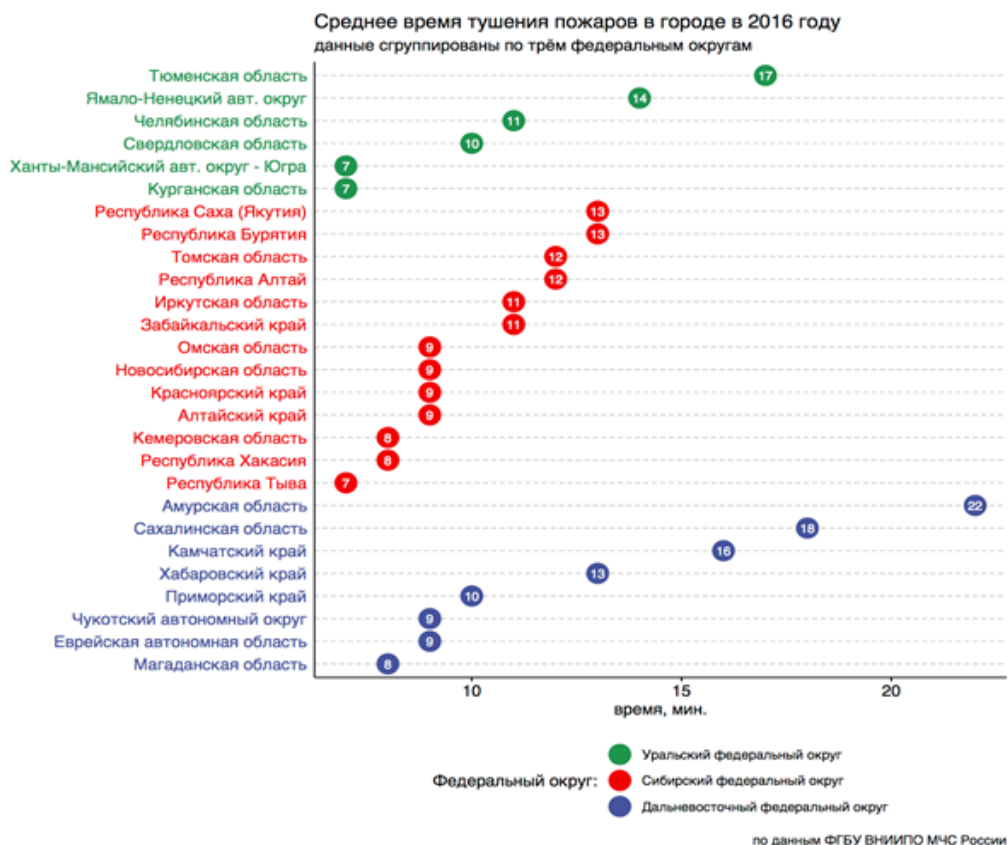


Рис. 2. Пример диаграммы Кливленда

Диаграмма размаха (или боксплот). Напомним, что диаграммы размаха используют статистику квартилей для представления групп числовых данных: на графике отмечаются каждые 25% данных (квартилям), прямоугольник содержит основную часть данных, от которого выходят так называемые усы, а значения, находящиеся далеко от основной массы значений, отображаются как точки – выбро-

сы. Линия внутри прямоугольника соответствует медиане. Пример диаграммы размаха приведен на рис. 3. Диаграммы размаха позволяют узнать: насколько плотно и симметрично сгруппированы данные, есть ли выбросы, каковы медианы внутри каждой категории, сравнить данные нескольких категорий.

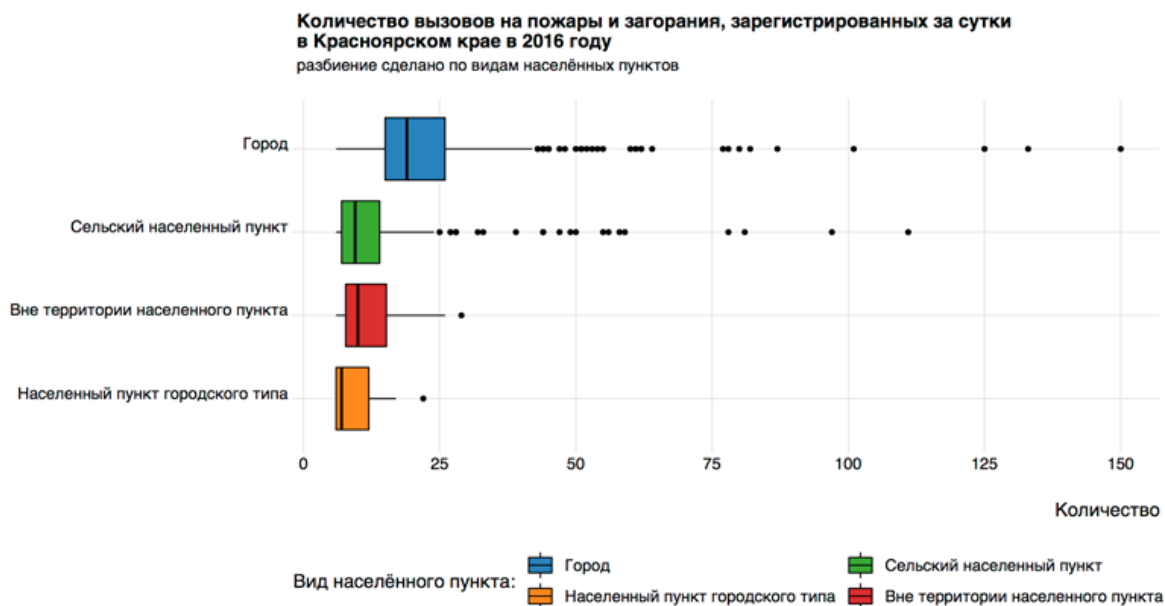


Рис. 3. Пример диаграммы размаха

Мозаичная диаграмма. Мозаичная диаграмма является способом визуализации двух и более категориальных переменных. Для изображения мозаичной диаграммы переменные упорядочиваются и им назначаются оси координат. Затем на диаграмме заполняются прямоугольники со сторонами, которые пропорциональны числу наблюдений в каждой категории. Следовательно, площадь каждого прямоугольника дает пропорцию количества наблюдений в выбранной категории. На рис. 4 показан пример

мозаичной диаграммы (построенной с использованием библиотеки ggmosaic), которая визуализирует долю срабатывания систем пожарной автоматики при пожаре в зависимости от количества пожаров по данным из [1]. Вдоль вертикальной оси отложено количество систем пожарной автоматики, задействованных во время пожаров, а вдоль горизонтальной оси – состояние пожарной автоматики во время пожара (цвет заполнения блоков).

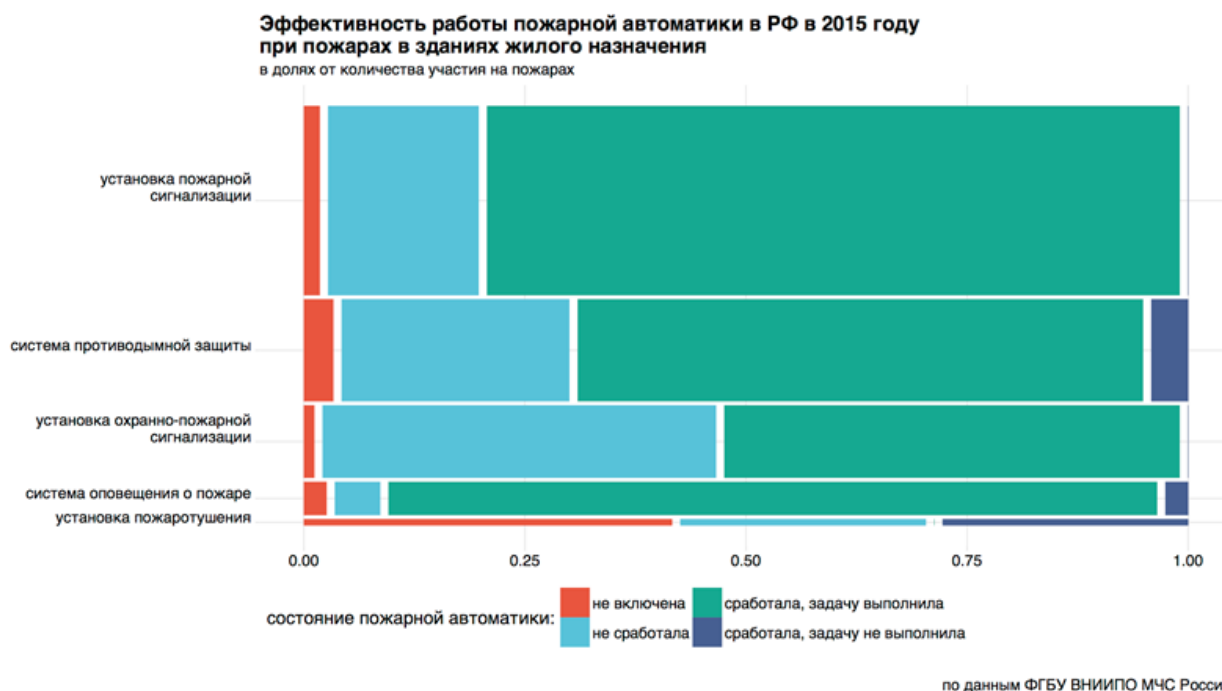


Рис. 4. Пример мозаичной диаграммы

Помимо упомянутых графиков выше, ggplot2 предлагает и множество других неклассических вариантов графиков (более подробно см. [2]).

Пример непосредственной работы в среде R

Общая идея обработки данных в Data Science с использованием R хорошо символизирует схема, предложенная Х. Викхэмом [7] на рис. 5, где:

- Импорт – импорт данных средствами библиотек readr, readxl, rvest;
- Таблица – упорядочивание табличных данных средствами tidy (при этом на каждом этапе обработки данных должен соблюдаться принцип организации данных tidy data, сформулированный в статье [8] (см. также [7]) – столбцы отвечают переменным, строки – наблюдениям, а на их пересечении находятся искомые значения);
- Преобразование – преобразование данных с помощью forcats для категориальных данных, lubridate для временных данных, а также tibble или data.table;

- Моделирование – моделирование реализовано в broom и modelr;
- Визуализация – происходит в ggplot2;
- Отчет – формирование отчетов средствами knitr, bookdown, shiny и т.д.

В качестве примера, рассмотрим обработку данных по карточкам учета пожаров в Красноярском крае. В настоящее время учет пожаров происходит в соответствии с Приказом МЧС России от 26.12.2014 года № 727 «О совершенствовании деятельности по формированию электронных баз данных учета пожаров (загораний) и их последствий». Кроме того, непосредственно учет пожаров и их последствий также ведется в соответствии с Приказом МЧС России от 21.11.2008 г. №714 (ред. от 17.01.2012 г.) «Об утверждении порядка учета пожаров и их последствий», который с 1.01.2009 г. вводит на территории РФ статистический порядок учета пожаров и их последствий.

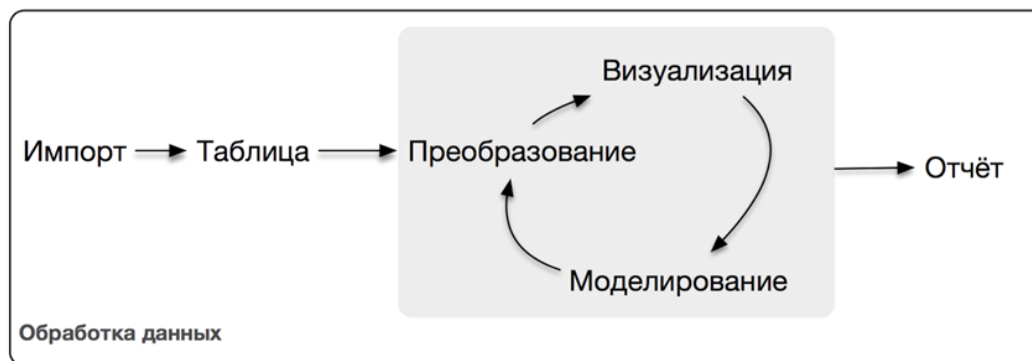


Рис. 5. Общая схема обработки данных

После загрузки необходимых библиотек загружаются данные, например, для Excel-таблиц, используя команду `readxl`. Пусть исходные данные представляют собой таблицу, у которой верхняя строка отвечает за имена переменных (столбцов), согласованные с карточкой учета пожаров (например, пусть F6 отвечает за дату возникновения по-

жара/загорания, F11 – за вид населенного пункта и т.д.).

Далее, приведем необходимые для исследования колонки в таблице к необходимым типам данных в R и назовем их в соответствии с приказом, определяющим вид карточки пожаротушения. Пример такой таблицы показан на рис. 6.

	F2	F6	F9	F11	F20
1	Енисейский район	2017-01-04	0	Город	Обособленная свалка бытовых и промышленных отхо...
2	Ирбейский район	2017-01-02	0	Сельский населенный пункт	Надворная постройка
3	Сухобузимский район	2017-01-01	0	Сельский населенный пункт	Многоквартирный жилой дом
4	Сухобузимский район	2017-01-07	0	Сельский населенный пункт	Полосы отчуждения и обочины дорог, луга, пустыри (в...
5	Сухобузимский район	2017-01-10	0	Сельский населенный пункт	Надворная постройка
6	Сухобузимский район	2017-01-17	0	Сельский населенный пункт	Многоквартирный жилой дом
7	Рыбинский район	2017-01-08	0	Сельский населенный пункт	Многоквартирный жилой дом
8	Рыбинский район	2017-01-07	0	Город	Полосы отчуждения и обочины дорог, луга, пустыри (в...
9	Рыбинский район	2017-01-11	0	Город	Полосы отчуждения и обочины дорог, луга, пустыри (в...
10	Туруханский район	2017-01-07	0	Сельский населенный пункт	Легковой автомобиль
11	Туруханский район	2017-01-09	0	Сельский населенный пункт	Многоквартирный жилой дом
12	Туруханский район	2017-01-09	0	Сельский населенный пункт	Неэксплуатируемое здание (сооружение)
13	Туруханский район	2017-01-16	0	Город	Склад со смешанным или универсальным ассортимент...
14	Бирюлюсский район	2017-01-07	0	Сельский населенный пункт	Легковой автомобиль
15	Бирюлюсский район	2017-01-15	0	Сельский населенный пункт	Многоквартирный жилой дом
16	Большеулуйский район	2017-01-16	0	Сельский населенный пункт	Одноквартирный жилой дом
17	Большеулуйский район	2017-01-18	0	Сельский населенный пункт	Баня, сауна на территории домовладения
18	Краснотуранский район	2017-01-06	0	Сельский населенный пункт	Одноквартирный жилой дом

Рис. 6. Пример таблицы электронной карточки учета пожаров

Язык R позволяет работать со специальными типами объектов, представляющих собой дискретные категориальные (факторные) данные, что в дополнении с функциями `group_by` и `filter`, `arrange`, `summarize` из пакета `dplyr` позволяет сортировать, фильтровать данные и рассматривать различные виды статистик, исходя из любых определяющих критериев (период, вид показателей, усреднение и т.д.). Затем полученные данные визуализируют и анализируют. Пример визуализации показан на рис. 7, где использован такой прием как панелирование – (от

`faceting`, англ.), то есть разбивка исходного графика на таблицы диаграмм, в каждой ячейке которой размещена диаграмма для отдельной категориальной переменной. В примере панелирования представленном на рис. 7 каждый график соответствует разбиению по видам населенных пунктов. Затененная гистограмма серого цвета соответствует суммарному недельному количеству пожаров.

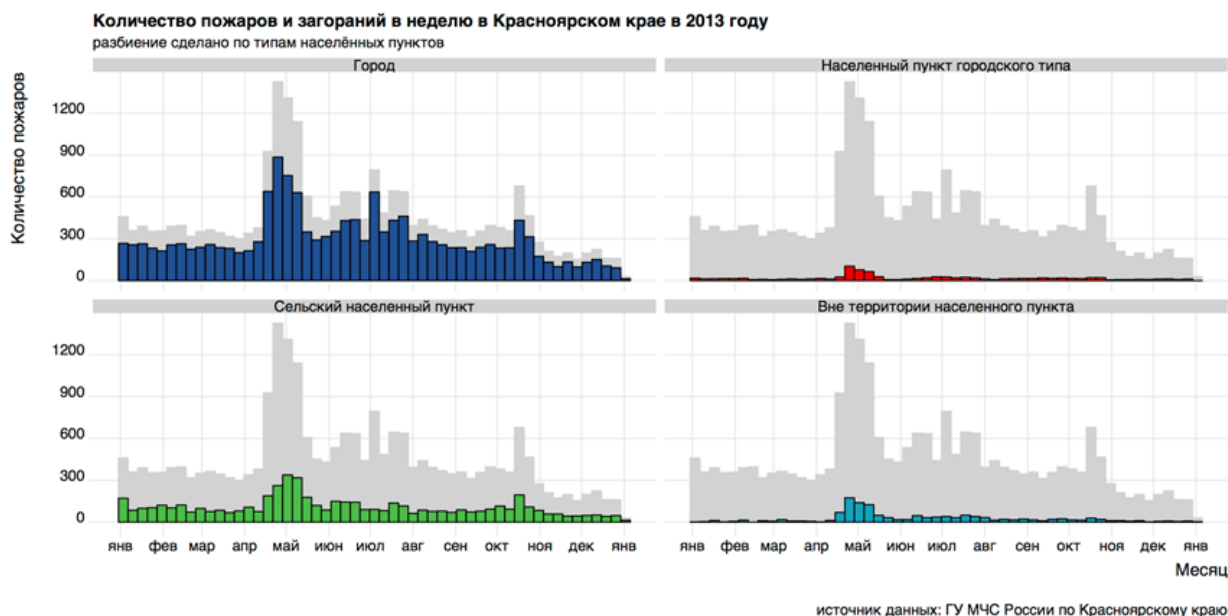


Рис. 7. Пример панелирования

Выводы. Язык программирования R является современным инструментом, который имеет лучшие в своем классе инструменты для визуализации, отчетности и интерактивности, стремительно набирающим популярность благодаря обилию прикладных библиотек, огромных возможностей для анализа данных и его бесплатности. Использование языка R в анализе пожарной статистики имеет большие перспективы. Следующим шагом, может стать моделирование и прогнозирование опасных процессов, что должно стать базисом для рекомендаций по снижению рисков.

Литература

1. Сайт «Статистика пожаров» [Электронный ресурс]. – Режим доступа: sites.google.com/site/statistikapozaro/, свободный.
2. 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.
3. Мастицкий, С. Э. Визуализация данных с помощью ggplot2 [Текст]. – М.: «ДМК-Пресс», 2017. – 222 с.
4. Мастицкий, С. Э. Статистический анализ и визуализация данных с помощью R [Текст]. / С. Э. Мастицкий, В. К. Шитиков. – М.: «ДМК-Пресс», 2015. – 496 с.
5. Шипунов, А. Б. Наглядная статистика. Используем R! [Текст]. / А. Шипунов, Е. Балдин, П. Волкова, А. Коробейников, С. Назарова, С. Петров, В. Суфиянов. – М.: «ДМК-Пресс», 2014. – 298 с.
6. Cleveland, W. Graphical Preception: Theory, Experimentation, and Application to the Development of Graphical Methods [Текст]. / William S. Cleveland, Robert McGill // Journal of the American Statistical Association, 1984. Vol. 79, No. 387, P. 531–534.
7. Wickham, H. R for Data Science [Текст]. / Wickham H., Grolemund G. – O'Reilly Media, 2016.
8. Wickham, H. Tidy data. The Journal of Statistical Software, vol. 59, 2014.