Информационные технологии и управление в области безопасности жизнедеятельности

УДК 519.2, 004.94

Использование языка программирования R в вопросах пожарной безопасности: анализ статистики количества пожаров на основе теории временных рядов

Using the R programming language for fire safety issues: analysis of statistics on the number of fires on the basis of time series theory

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 основных приемов работы с временными рядами в языке программирования **R** на примере данных по статистике пожаров в Красноярском крае: разведочный анализ данных, разложение временных рядов на компоненты, интегрированная модель авторегрессии-скользящего среднего и некоторые другие методы.

Ключевые слова: язык программирования **R**, временные ряды, статистика пожаров

Abstract:

The article provides an overview of the main methods of working with time series in the **R** programming language on the example of data on fire statistics in the Krasnoyarsk region: exploratory data analysis, time series decomposition into components, integrated autoregression-moving average model and some other methods.

Key words: R programming language, time series, fire statistics

Одной из основных задач действия органов управления МЧС России в различных режимах функционирования является сбор данных обстановки, ее анализ и оценка, и, соответственно, прогнозирование обстановки на основе статистических данных. При этом часто измерения тех или иных показателей производятся не один раз, а повторяются через некоторые интервалы времени (например, месяц, год) – количество пожаров, количество травмированных людей на пожарах, прямой материальный ущерб от пожаров, и так далее.

Анализируемые данные по пожарам и их последствиям (например дата возникновения пожара, степень огнестойкости, количество погибших и т.д.) содержатся в электронной карточке учета пожара, которые затем заносятся в электронную базу. В данной статье были проанализированы данные по пожарам, собранные Главным управлением МЧС

России по Красноярскому краю. Напомним, что в настоящее время учет пожаров происходит в соответствии с Приказом МЧС России от 26 декабря 2014 года №727 «О совершенствовании деятельности по формированию электронных баз данных учета пожаров и их последствий». Кроме того, непосредственно учет пожаров и их последствий также ведется в соответствии с Приказом МЧС России от 21 ноября 2008 года №714 (ред. от 17 января 2012 года) «Об утверждении порядка учета пожаров и их последствий» с учетом изменений, установленных Приказом МЧС России от 08 октября 2018 года № 431.

Поскольку статистические данные электронной базы были измерены через упорядоченные промежутки времени, имеет смысл изучать статистический материал с помощью такого математического понятия как временной ряд. Вообще говоря, под временным рядом (динамическим рядом или временной последовательностью) понимается ряд значений, соответствующих измерениям в различные моменты (обычно через равные промежутки) времени.

Основная идея анализа данных по учету пожаров состоит в реализации следующих целей:

1. Общий анализ структуры данных, выполнение манипуляций над данными для получения отчетов по количеству пожаров по требуемым категориям и показателям, а также визуализация данных (разведочный анализ данных).
2. Построение математической модели временного ряда, учитывающего количество пожаров и прогнозирование.

В такой же последовательности изложен материала в статье. Статья имеет обзорный характер и не претендует на глубокий анализ всех методов исследования данных в применении к теории временных рядов, а имеет целью дать дайджест некоторых методов с использованием специального программного обеспечения. Вопросы применения методов машинного обучения, в частности, использования нейронных сетей и глубокого обучения будут рассмотрены в последующей статье. Использование временных рядов в применении к статисике пожаров рассматривалось ранее, например, в статьях [1-2].

Для реализации целей анализа структуры данных и построения моделей временных рядов был использован один из лидеров в области бесплатного программного обеспечения – язык программирования **R**. Язык **R** является одним из наиболее активных современных средств обработки и визуализации данных. Обзор языка **R** в применении к методам науки о данных (Data Science) в вопросах пожарной безопасности был дан в статье [3], см. также [4]. Кратко отметим некоторые преимущества **R** по сравнению с другими средствами обработки данных: огромные возможности по реализации современных статистических методов, полиграфическое качество графики, обработка больших групп данных и т.д. В настоящий момент **R** стал классическим инструментом для обработки временных рядов, уже было издано несколько монографий по теории временных рядов, в которых язык **R** играет ключевую роль [5, 6].

Разведочный анализ данных по пожарам в Красноярском крае методами R

Основу современного языка программирования **R** в применении к структурной обработке и визуализации данных составляет группа библиотек `tidyverse`, объединенных общим принципом построения табличных данных – «`tidy data`»: столбцы данных отвечают переменным, строки – наблюдениям, а на их пересечении находятся искомые значения (см. [7]). Схематично этапы обработки данных и соответствующие библиотеки показаны на рис. 1.

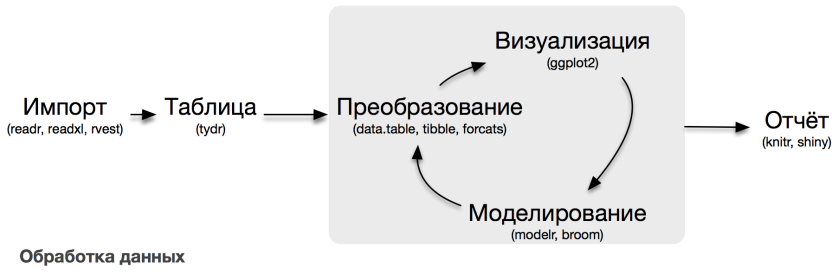

Рис. 1. Схема обработки и визуализации данных

В применении к временным рядам принцип tidy data реализуют, например, библиотеки **R**: `tidyquant`, `tsible`, `fable`, `fasster`.

Разведочный анализ данных по пожарам в Красноярском крае

Подготовительный этап. После загрузки необходимых библиотек загружаются данные, на-

пример, используя библиотеку `readxl` (для Excel-файлов). Пусть исходные данные представляют собой таблицу, у которой верхняя строка отвечает за имена переменных (столбцов), согласованные с карточкой учета пожаров (например, F6 отвечает за дату возникновения пожара, **F2** – район, **F9** – количество погибших, и т.д.). В исходной Excel-таблице каждая ячейка в столбце представляет собой код, значение которого указано в соответствующем приказе. Команда `factor` позволяет поставить коду в соответствие его полное наименование (например, в столбце, отвечающем причинам пожаров, коду «`10`» соответствует «`Нарушение правил монтажа электрооборудования`» и т.д.). Затем из всего набора переменных (например, для файла по 2016 году это около 200 переменных) выбираются только значимые для анализа: дата возникновения пожара, район, количество погибших, степень огнестойкости, место возникновения пожара, причина пожара, код конструкции и т.д. Пример такой «предобработанной» таблицы показан на рис. 2.

	F2	F6	F9	F20	F26	F27
1	Рыбинский район	2016-01-02	0	Площадка для мусора, мусор на территории жил...	Не указана	На территории домовладения
2	Рыбинский район	2016-01-04	0	Площадка для мусора, мусор на территории жил...	Не указана	На территории домовладения
3	Рыбинский район	2016-01-09	0	Полосы отчуждения и обочины дорог, луга, пуст...	Не указана	На пустыре
4	Рыбинский район	2016-01-09	0	Площадка для мусора, мусор на территории жил...	Не указана	На территории домовладения
5	Рыбинский район	2016-01-12	0	Легковой автомобиль	IV	Отсек двигателя
6	Рыбинский район	2016-01-13	0	Площадка для мусора, мусор на территории жил...	Не указана	На территории домовладения
7	Рыбинский район	2016-01-13	0	Площадка для мусора, мусор на территории жил...	Не указана	На территории домовладения
8	Рыбинский район	2016-01-15	0	Многоквартирный жилой дом	V	Веранда, терраса, тамбур
9	Рыбинский район	2016-01-15	0	Площадка для мусора, мусор на территории жил...	Не указана	На контейнерной площадке
10	Курагинский район	2016-01-02	1	Одноквартирный жилой дом	V	Кухня
11	Курагинский район	2016-01-03	0	Специальная техника	Отсутствует	Отсек двигателя
12	Курагинский район	2016-01-04	0	Одноквартирный жилой дом	V	Подсобные и вспомогательные производственны...
13	Курагинский район	2016-01-05	2	Одноквартирный жилой дом	V	Кухня
14	Курагинский район	2016-01-06	0	Надворная постройка	V	Подсобное помещение
15	Курагинский район	2016-01-07	0	Полосы отчуждения и обочины дорог, луга, пуст...	Не указана	На обочине дороги
16	Курагинский район	2016-01-09	0	Одноквартирный жилой дом	V	Чердачное помещение
17	Курагинский район	2016-01-09	0	Гараж, тент-укрытие и т.д.	V	Помещение для хранения и ремонта транспорта ...

Рис. 2. Пример предобработанной карточки учета пожаров

Работа с данными. После того, как данные предобработаны, можно легко выделить статистические сводки по выбранным критериям. Отметим, что на каждом этапе будет соблюдаться принцип «`tidy data`», упомянутый выше.

Проиллюстрируем основную идею компьютерного анализа на данных по пожарам за 2016 год. Язык R позволяет работать со специальными типами объектов, представляющих собой дискретные категориальные (факторные) данные, что в дополнении с функциями `group_by`, `summarise`, `mutate` и `filter` из библиотеки `dplyr` позволяет *сортировать*, *фильтровать* данные, *создавать новые переменные* и рассматривать *различные виды описательных статистик*. Например, для того, чтобы узнать количество пожаров по видам населенных пунктов, можно воспользоваться командой `count` и упорядочить данные по убыванию командой `arrange` с использованием опции `desc`. Отметим, что команды **R** удобно группируются в последовательные цепочки командой «`pipe`» `%>%` из библиотеки `magrittr`. Здесь таблица данных обозначена переменной `df`, а населенным пунктам соответствует переменная (столбец) **F11**.

```
count(df, F11) %>% arrange(., desc(n))
## # A tibble: 7 x 2
##                                F11     n
##                              <fctr> <int>
## 1                             Город  8640
## 2         Сельский населенный пункт  4057
## 3  Населенный пункт городского типа   815
## 4 Вне территории населенного пункта   765
## 5                           Станция     7
## 6                  Разъезд, перегон     1
## 7                  Вахтовый поселок     1
```

Если есть необходимость рассмотреть, например, те же данные, но только по пожарам в многоквартирных жилых домах, то это возможно используя команду filter (здесь показатель **F201** равный 1 соответствует пожарам, **F20** – отвечает за объект пожара).

```
df %>%
 filter(F201 == 1) %>%
 filter(F20 %in% c("Многоквартирный жилой дом")) %>%
 count(., F11) %>% arrange(., desc(n))
## # A tibble: 7 x 2
##                                F11     n
##                              <fctr> <int>
## 1                             Город  796
## 2         Сельский населенный пункт  208
## 3  Населенный пункт городского типа   42
```

Такого рода анализ данных можно продолжить, выделяя различные критерии для группировки,

фильтрации и сводных статистик. В дальнейшем эти данные передаются для визуализации и последующего анализа временных рядов.

Визуализация данных. Обработанные данные легко визуализировать используя библиотеку ggplot2. Сначала данные записываются в новую таблицу, содержащую два столбца: Дата и Сумма (количество пожаров на эту дату).

```
data_time_series <- df %>%
  rename("Дата" = "F6") %>%
  group_by(Дата) %>%
  summarise(Количество = n())
```

Теперь легко построить график количества зарегистрированных в сутки пожаров в Красноярском крае в 2016 году (см. рис. 3).

```
data_time_series %>%
  ggplot(aes(x = Дата,
             y = Количество)) +
  geom_line()
```

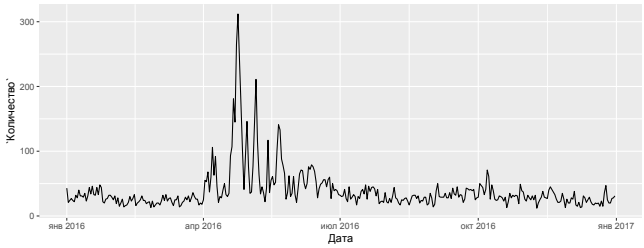

Рис. 3. Количество пожаров в Красноярском крае в 2016 году

Отметим, что согласно исследованию, наибольшие всплески по количеству пожаров каждый год всегда происходят в конце апреля.

Разложение временных рядов на компоненты

Как было отмечено во введении, данные по пожарам представляют собой временной ряд и могут быть исследованы соответствующим образом. Для исследования временных рядов в **R** имеется довольно большое количество как встроенных средств так и подключаемых библиотек (например, `astsa`, `forecast`, `sweep`, `timetk`, `trend`, `TSA`, `TSTutorial`, `tseries`, `xts`, `zoo`, и т.д.). В базовой версии **R** используется специальный встроенный класс объектов `ts` для работы с данными, представляющими временные ряды. Список библиотек **R**, работающих с временными рядами, можно посмотреть по адресу: https://cran.r-project.org/web/views/TimeSeries.html

Основная цель анализа временных рядов заключается в разработке математических моделей, которые обеспечивают правдоподобное описание для выборки данных, а именно:

- описание основных характеристик временного ряда;
- подбор статистической модели, описывающей временной ряд;
- прогнозирование будущих значений на основе имеющихся наблюдений, относящихся к прошлому.

Для моделирования временного ряда традиционно используют либо мультипликативную, либо аддитивную модель. Например, пусть y_t описывает временной ряд в момент времени t. Тогда аддитивная модель использует следующую формульную зависимость:

$$y_t = f(T_t,S_t,Z_t) = T_t + S_t + Z_t$$

со следующими составляющими:

T_t = *тренд* или *основная детерменированная тенденция временного ряда* – плавное долгосрочное изменение показателей временного ряда.

S_t = *сезонная компонента* – циклические изменения уровня ряда с постоянным периодом.

Z_t = *случайная компонента* (или *остаток*) – несистематическая компонента, вызванная помехами и шумами измерений.

Для разложения временного ряда на компоненты можно использовать разложение STL (Seasonal and Trend decomposition using Loess) – от англ. «сезонное и трендовое разложение с использованием метода Кливленда» [5, Section 6.5]. Метод Кливленда Loess введен в работе [8].

Ниже на рис. 4 приведен пример такого разложения для аддитивной модели ряда, построенной с помощью функции `stl` в **R**, и соответствующей временному ряду, отображающему количество пожаров в Красноярском крае в период с 2009 по 2012 года. Исходные данные (**data**) на рисунке являются суммой графиков ниже: остатков (**remainder**) + сезонной компоненты (**seasonal**) + тренда (**trend**). Тренд показывает, что общее количество пожаров в рассматриваемый период увеличилось.

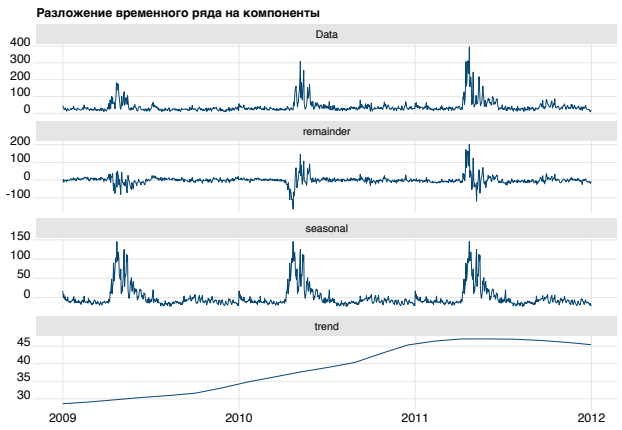

Рис 4. Пример разложения временного ряда на компоненты

ARIMA-модель временного ряда

Для прогнозирования количества пожаров можно использовать множество различных моделей. Рассмотрим как работает в **R** простейшая модель – интегрированная модель авторегрессии-скользящего среднего Бокса-Дженкинса. Напомним, что модель ARIMA*(p, d, q)* для временного ряда имеет следующий вид

$$\phi(B)\,(1-B)^d\,y_t = \delta + \theta(B)\,w_t,$$

где

B – разностный *оператор обратного сдвига*, работающий по правилу $By_t = y_{(t-1)}$;

$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ – *оператор авторегрессии*;

$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$ – *оператор скользящего среднего*;

δ – так называемый «*снос*» (drift);

w_t – набор некоррелированных случайных величин с распределением $iid(0; 1)$ (*гауссов белый шум*).

Для построения ARIMA-моделей в **R** очень хорошо работает функция `auto.arima` из библиотеки `forecast`. Компьютер перебирает все возможные ARIMA модели и ищет ту, которая соответствует минимуму по информационному критерию Акаике, применяемому для выбора из нескольких статистических моделей и его аналогов (`AIC`, `AICc`, `BIC`). Например, для временного ряда, соответствующего количеству пожаров в 2016 году, изображенном на рис. 3 оптимальной является модель ARIMA(2, 1, 1) и имеет вид:

$$\nabla y_t = 0.83\nabla y_{t-1} - 0.15\nabla y_{t-2} + w_t - 0.93\,w_{t-1}$$

где $\nabla y_t = (1-B)y_t = y_t - y_{t-1}$. Убедиться в том, что модель выбрана верно (остатки являются белым шумом), можно с помощью команды `checkresiduals` (см. рис. 5).

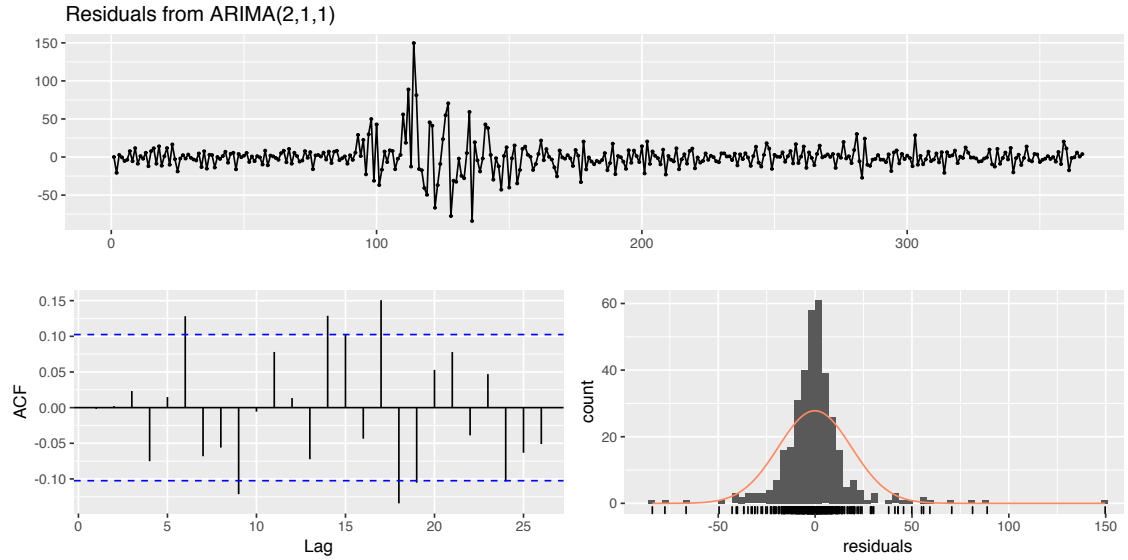

Рис 5. Остатки модели ARIMA(2, 1, 1)

Для прогнозирования количества пожаров на несколько дней вперед можно использовать, например, команду `forecast` из одноименной библиотеки, либо команду `sarima.for` билиотеки `astsa`, которая, помимо прогноза значений, изображает график прогноза (см. рис. 6).

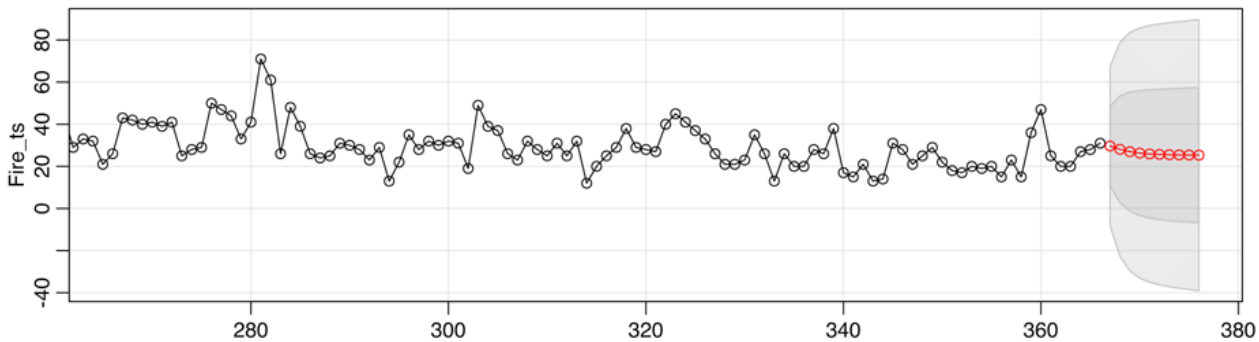

Рис 6. Прогноз количества пожаров на 2017 год (красный цвет)

Стационарность временного ряда можно проверить используя стандартные статистические тесты, например, аугментированный (расширенный) тест Дики-Фуллера командой `adf.test` из библиотеки `tseries`, в котором основной (нулевой) гипотезой является отсутствие стационарности, а альтернативной гипотезой является стационарность ряда или *Квятковский-Филлипс-Шмидт-Шинмест* (*Kwiatkowski–Phillips–Schmidt–Shin*) командой `kpss.test`.

Для определения тренда, а также минимального/максимального количества пожаров по дням недели/месяцам (даже учитывая праздничные дни) можно использовать библиотеку `prophet` (см. пример на рис. 7).

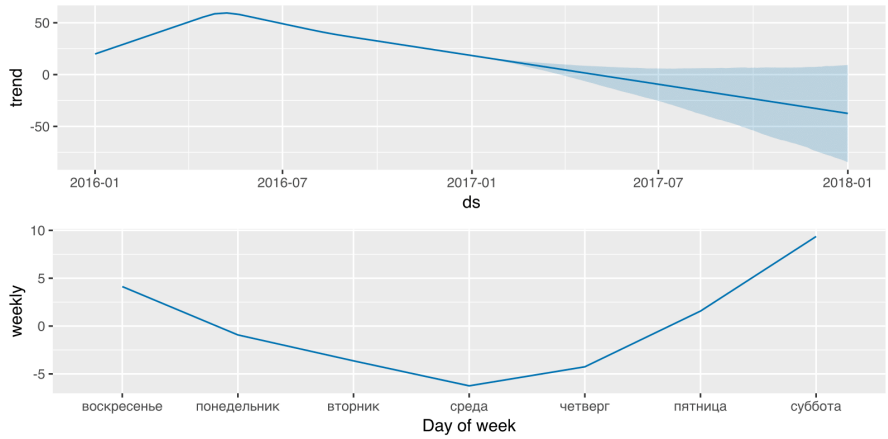

Рис 7. Пример недельного разложения и прогноза тренда временного ряда

Тема исследования временных рядов очень обширна. За рамки данной статьи выходят: исследование сезонности, аномалий, экспоненциального сглаживания Холта-Винтерса, исследование временных рядов с помощью нейронных сетей и т.д. Более подробно применение теории временных рядов к вопросам пожарной безопасности также изложено в [4].

Выводы. Использование языка программирования **R** для исследования статистики пожаров с точки зрения теории временных рядов позволяет более точно, чем существующие методики, моделировать оперативную обстановку по пожарной безопасности, существенно подняв уровень исследований в этой области на основе математически обоснованной теории и современного уровня программирования.

Литература

1. Батуро, А. Н. Прогнозирование количества пожаров в регионе на основе теории временных рядов. Технологии гражданской безопасности. 2013. Т. 10. № 3 (37). С. 84-88.
2. Батуро, А. Н. Среднесрочное прогнозирование количества пожаров с использованием автокорреляционных функций. Природные и техногенные риски (физико-математические и прикладные аспекты). 2014. № 3 (11). С. 28-36.
3. Матеров, Е.Н. Использование языка программирования R в вопросах пожарной безопасности: обработка и визуализация данных / Матеров Е.Н. // Научно-аналитический журнал «Сибирский пожарно-спасательный вестник», 2018, № 4. - С. 60-66. – Режим доступа: http://vestnik.sibpsa.ru/wp-content/uploads/2018/v4/N11_60-66.pdf, свободный. – Загл. с экрана. — Яз. рус., англ.
4. 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.
5. Athanasopoulos, G. Forecasting: principles and practice [Электронное издание]. / Athanasopoulos G., Hyndman R. – O.Text, 2014. Режим доступа: https://www.otexts.org/book/fpp, свободный.
6. Shumway, R. Time Series Analysis and Its Applications with R Examples [Текст]. / Shumway R., Stoffer D. – [4th edn.] – Springer Texts in Statistics, 2017. – 564 P.
7. Wickham, H. Tidy data. The Journal of Statistical Software, vol. 59, 2014.
8. Cleveland, R. STL: A Seasonal-Trend Decomposition Procedure Based on Loess [Текст]. / Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning // Journal of O cial Statistics, 1990. – Vol.6, No.1. PP. 3–73.