Информационные технологии и управление в области безопасности жизнедеятельности

УДК 519.2, 004.94

Использование языка программирования R в вопросах пожарной безопасности: анализ главных компонент

Using the R programming language for fire safety issues: Principal component analysis

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 рассмотрен метод главных компонент в применении к анализу данных по пожарам в Красноярском крае, а также рассмотрены возможности реализации данного метода на языке программирования R. Этапы реализации проиллюстрированы на конкретном примере данных пожаров на автотранспорте.

Ключевые слова: язык программирования R, анализ главных компонент, статистика пожаров.

Abstract:

The article describes the Principal Component Analysis (PCA) applied to the analysis of data on fires in the Krasnoyarsk Territory, as well as the possibility of implementing this method in the R programming language. The stages of implementation are illustrated with a specific example of these fires in motor vehicles.

Key words: R programming language, principal component analysis, fire statistics.

Информационная поддержка анализа чрезвычайных ситуаций и их последствий невозможна без эффективных современных инструментов обработки статистических данных. Одним из лидеров в области бесплатных средств обработки и визуализации данных является язык программирования R. Некоторые аспекты науки о данных (Data Science) и теории временных рядов с использованием R в применении к анализу пожаров в Красноярском крае были рассмотрены в статьях [1], [2] и в книге [3]. Данная статья является продолжением предыдущих работ и рассматривает практическое применение метода главных компонент в анализе пожаров и его реализацию на языке R.

Метод главных компонент (или PCA – Principal Components Analysis) относится к одним из основных статистических методов сжатия большого массива информации без большой потери информативности. Данный метод показывает себя эффективно, когда количество различных признаков у исследуемых объектов очень велико и необходимо выде-

лить признаки, имеющие наибольшую смысловую нагрузку. Основными предпосылками для введения главных компонент являются: дублирование информации, доставляемой сильно взаимосвязанными (коррелированными) признаками, мало меняющимся при переходе от одного объекта к другому, а также необходимостью наглядного представления данных путем проецирования данных на оси (обычно на двумерной плоскости), определяемыми новыми переменными, которые максимизируют выборочную дисперсию проекции данных.

Идея данного метода заключается в замене параметров модели на линейную комбинацию некоррелированных между собой новых компонентов, таких, что каждая очередная компонента дает наибольший возможный вклад в суммарную дисперсию параметров. В итоге такого преобразования каждое из наблюдений представляется в виде вспомогательных показателей с существенно меньшим числом компонент. Математическая реализация метода главных компонент изложена, например, в [4, Глава 13], а с точки зрения программирования на R узнать об основах интерпретации главных компонент можно их книги [5].

В качестве анализируемых данных рассмотрим данные по пожарам на автотранспорте по данным Главного управления МЧС России по Красноярскому краю.

Предобработка данных

Предобработка данных включает в себя следующие этапы (см. [1]).

1. Данные таблицы импортируются в R (например, из Excel-файла командой read_xlsx).
2. Коды значений полей переводятся в факторные переменные, названия которых соответствуют определяющему их приказу.
3. Из всей совокупности переменных выбираются только те, которые имеют наибольшую смысловую нагрузку (например, причины пожаров, материал горения, место возникновения пожара (загорания), тип горения, дата возникновения пожара, и т.д.).

Далее для манипуляций над данными используются библиотеки tidyr и dplyr из группы библиотек tidyverse. В частности,

1. Проводится фильтрация данных (например, можно выбрать данные только за определенный промежуток времени, тип горения, рассмотреть только нужный тип населенного пункта, например, г. Красноярск, и т.д.), поиск и заполнение пропущенных данных.
2. Составляется сводная таблица данных, в которой:

- строки отвечают категорным наблюдениям (например, району, месяцу года, дню недели, и т.д.);
- столбцы соответствуют категорным показателям (это могут быть: причины пожаров, материал горения, место возникновения пожара, и т.д.);
- элементы таблицы – количественные переменные (количество пожаров, количество погибших, и т.д.).

Рассмотрим пример на основе анализа пожаров на транспорте в Красноярском крае в 2009-2013 годах. Выделим две переменные, отвечающие, соответственно, за причины и объекты пожаров. Пример интересующих нас данных показан в Таблице 1 ниже, полные данные по причинам и объектам пожаров содержат 96 строк.

Таблица 1. Данные по объектам и причинам пожаров на транспорте в Красноярском крае с количеством пожаров более 100 в 2009-2013 гг.

№	Объект пожара	Причина пожара	Количество
1	легковой автомобиль	Неисправность систем, механизмов и узлов ТС	546
2	легковой автомобиль	Поджог	490
3	легковой автомобиль	Неосторожное обращение с огнём неустановленных лиц	454
4	грузовой автомобиль	Неисправность систем, механизмов и узлов ТС	112

Метод главных компонент

Для работы с методом главных компонент (далее PCA) в R существует множество библиотек, например, auto.pca, FactoMineR, ade4, amap, factoextra, помимо базовых функций R. Отметим, что в методе главных компонент данные, как правило, масштабируются, в большинстве указанных библиотек R это делается автоматически.

Напомним, что цель PCA состоит в том, чтобы определить направления (или основные компоненты), вдоль которых изменение данных является максимальным. Другими словами, PCA уменьшает размерность многомерных данных до нескольких (как правило двух) главных компонентов, которые могут быть визуализированы графически, с минимальными потерями информации. Первая размерность – главная компонента (PC1) выбирается так, чтобы изменчивость облака рассеяния вдоль этой компоненты была максимальной, затем выбирается следующая компонента (PC2) и так далее. При этом величина дисперсии (меры разброса), кодируемая

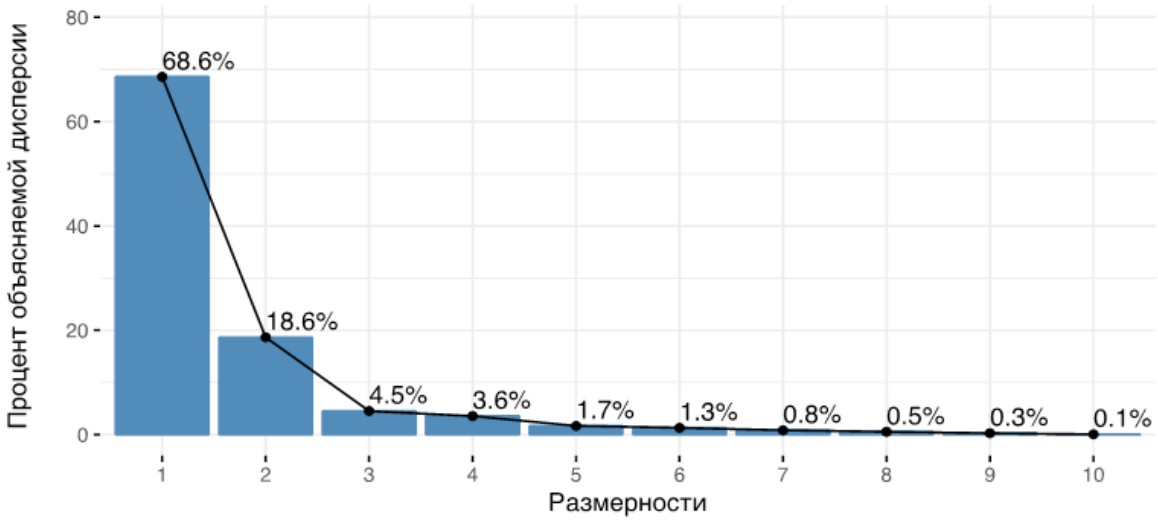

Рис. 1. График собственных значений

каждой главной компонентой, измеряется так называемым собственным значением.

В примере по причинам пожаров на транспорте в качестве показателей выберем причины пожаров, тогда соотношение между объектами и причинами пожаров можно представить таблицей смежности размерности 14x26, часть которой показана в Таблице 2.

Таблица 2. Соотношения между объектами и причинами пожаров на транспорте в Красноярском крае в 2009-2013 гг.
(показана часть данных)

	Взрывы	Дорожно-трансп. происшествие	Наруш. пожароб. при экспл. печей	Наруш. монтажа тепло-генер. устр.	Наруш. пожаробезоп. при огнев. работах
автобус	0	0	0	0	1
автомобильная цистерна	1	0	0	0	0
грузовой автомобиль	0	0	2	0	2
легковой автомобиль	3	1	0	2	14

Используя функцию fviz_eig из библиотеки factoextra можно визуализировать график (Scree plot) собственных значений (см. Рис. 1).

Из графика следует, что 68,6% дисперсии объясняется первым собственным значением, 18,6% – вторым собственным значением и так далее. В нашем примере уже первые две компоненты объясняют 87,2% процента дисперсии, что является весьма хорошим показателем.

Далее, покажем как визуализировать переменные и сделать на основе графиков выводы об их взаимосвязи. Затем выделяются переменные в соответствии с их качеством представления на карте факторов или их вкладом в основные компоненты.

Круг корреляций

Здесь корреляция между переменной и главным компонентом (PC) используется в качестве координат переменной на PC. Представление переменных отличается от графика наблюдений: наблюдения представлены их проекциями, а переменные – их корреляциями.

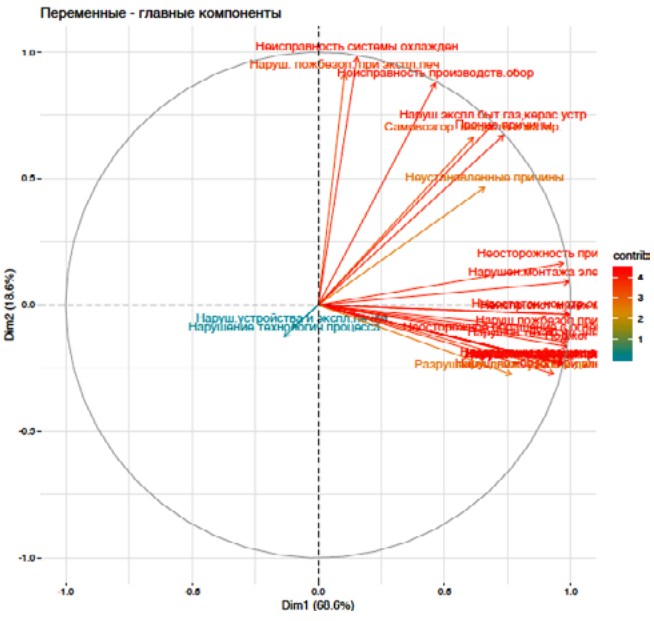

Рис. 2. Круг корреляций

На Рис. 2 показан круг корреляций (реализовано командой fviz_pca_var из библиотеки factoextra), проекция на две главные размерности. Чем ближе

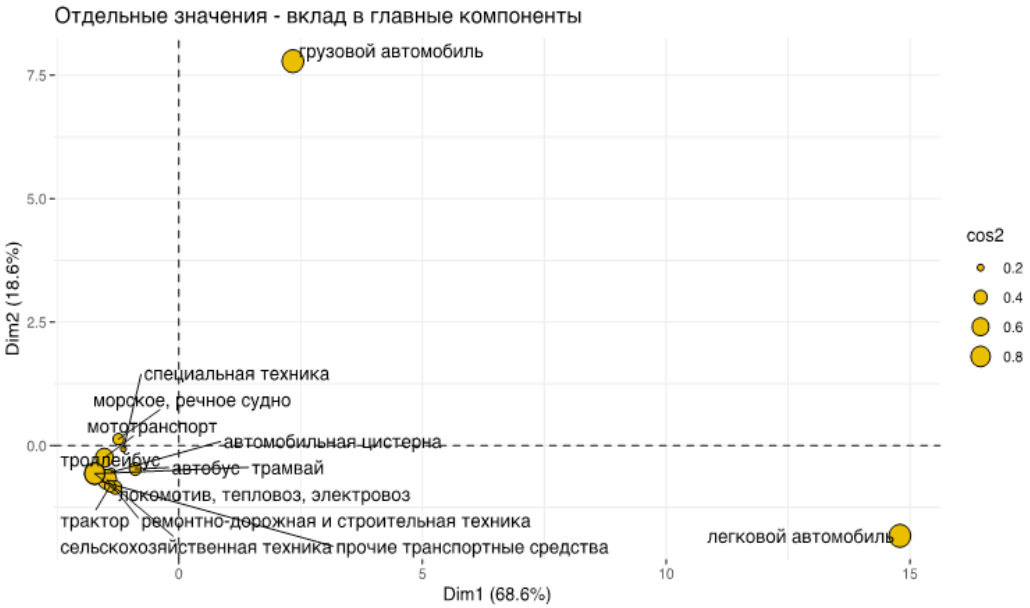

Рис. 3. Диаграмма отдельных значений (наблюдений)

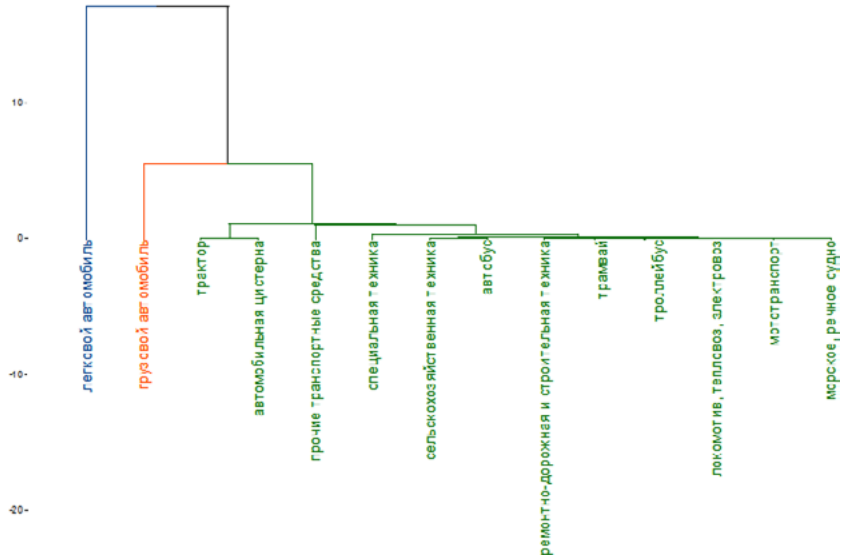

Рис. 4. Кластерная дендрограмма по причинам пожаров

переменная к границе круга корреляций, тем больше ее вклад (и тем важнее интерпретировать эти компоненты). В то же время, переменные, удаленные от центра графика менее важны для вклада в первые компоненты. Наиболее важные (вносящие наибольший вклад) компоненты также выделены градиентным цветом в соответствии с легендой графика. Кроме того, угол между векторами, соответствующими компонентам показывает корреляцию между ними – положительно коррелирующие переменные сгруппированы вместе, а отрицательно коррелирующие переменные располагаются по разные стороны относительно центра круга. Например, в данном примере наибольший вклад вносит переменная, отвечающая за неисправность систем, механизмов и узлов транспортных средств, а наименьший – за причины, связанные с эксплуатацией печей.

Качество представления и индивидуальные значения

Качество представления переменных на факторной карте определяется показателем cos2 (квадратный косинус). Для фиксированной переменной сумма cos2 на всех главных компонентах равна единице. Чем больше показатель cos2, тем лучше представление переменной на главном компоненте.

На Рис. 3 показана диаграмма отдельных значений (в данном случае это объекты пожара), где раз-

мер кружка соответствует cos2 для фиксированного объекта пожара, реализованная командой fviz_pca_ind из библиотеки factoextra. Показательным для данного представления является вклад в две первые компоненты – можно наглядно выделить три группы: легковые автомобили, грузовые автомобили и все остальные транспортные средства в зависимости от причин пожаров.

Иерархическая кластеризация

Близость различных причин пожаров в рассматриваемом примере можно также наглядно показать с помощью дендрограмм – вида диаграмм, представляющих собой дерево для иллюстрации иерархической зависимости.

Пример кластерной дендрограммы показан на Рис. 4. Чем ближе в иерархии изображены транспортные средства, тем более «похожи» соответствующие причины пожаров. Отметим, что Рис. 3 полностью соответствует представлению на Рис. 4, где выделяются три кластера точек, отвечающих за наблюдения.

Заключение

Отметим, что данный вид анализа применим ко многим практическим данным, например при анализе пожаров по электронной карточке учета пожаров.

В качестве наблюдений и определяющих переменных в таблице первичных данных могут выступать совершенно различные характеристики, участвующие в анализе. Например, можно определять «близость» субъектов РФ по причинам пожаров, коду конструкции и т.п.

В данной работе был рассмотрен лишь простейший вид анализа, основанный на рассмотрении количественных данных. Для номинальных (категориальных) переменных, а также смешанных переменных применяется анализ соответствий, факторный анализ смешанных данных и множественный факторный анализ, которые выходят за пределы статьи. Также, за рамки статьи выходят многие вопросы реализации метода главных компонент на R (см. [5]).

Автор благодарен старшему научному сотруднику Отдела прикладных исследований и инновационных технологий Научно-технического центра ФГБОУ ВО Сибирская пожарно-спасательная академия ГПС МЧС России Ворошилову Роману Феофановичу за предоставленные данные.

Выводы

В данной работе рассмотрено эффективное применение метода главных компонент к анализу обстановки с пожарами и их последствиями, реализованные на языке программирования R, который обладает большими возможностями для обработки и визуализации данных. Большими преимуществами R также являются простота использования и бесплатность реализации. Данная методика при ее применении к вопросам, связанными с пожарной безопасностью, позволит проводить более глубокий и визуально доступный анализ данных, связанный с приоритетными направлениями научной деятельности МЧС России, а также быть основой для разработки автоматизированных систем поддержки принятия решений.

Литература

1. 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, No 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.
2. Матеров, Е.Н. Использование языка программирования R в вопросах пожарной безопасности: анализ статистики количества пожаров на основе теории временных рядов / Матеров Е.Н. // Научно-аналитический журнал «Сибирский пожарно-спасательный вестник», 2019, №1.- С.52-57.- Режим доступа: http://vestnik.sibpsa.ru/wp-content/uploads/2019/v1/N12_52-57.pdf, свободный. – Загл. с экрана. – Яз. рус., англ.
3. 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.
4. Айвазян, С.А. Прикладная статистика. Классификация и снижение размерности [Текст]: Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин; Под ред. проф. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с., ил.
5. Kassambara, A. Practical Guide to Principal Component Methods in R [Текст]. / A. Kassambara. – Data Novia, 2017. – 170 P