Алхасов Станислав Сергеевич, Целых Александр Николаевич

# ОСНОВНЫЕ ЭЛЕМЕНТЫ БЛОКА ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ В ПРИКЛАДНЫХ ЗАДАЧАХ АНАЛИЗА ДАННЫХ

В настоящей работе представлены главные компоненты блока предварительной обработки данных, необходимой для корректности дальнейшего проведения статистического и интеллектуального анализа разнородных входных данных комплексной информационно-аналитической системы. Описаны основные методы предварительной обработки данных, применяемые для сравнения рядов данных, удаления выбросов, снижения размерности.

Адрес статьи: www.gramota.net/materials/1/2016/3/1.html



Источник

**Альманах современной науки и образования**

Тамбов: Грамота, 2016. № 3 (105). С. 11-13. ISSN 1993-5552.

Адрес журнала: www.gramota.net/editions/1.html

Содержание данного номера журнала: www.gramota.net/materials/1/2016/3/

УДК 504.064.36
**Технические науки**

*В настоящей работе представлены главные компоненты блока предварительной обработки данных, необходимой для корректности дальнейшего проведения статистического и интеллектуального анализа разнородных входных данных комплексной информационно-аналитической системы. Описаны основные методы предварительной обработки данных, применяемые для сравнения рядов данных, удаления выбросов, снижения размерности.*

*Ключевые слова и фразы:* предварительный анализ данных; информационно-аналитическая система; язык программирования *R*; разнородные входные данные; статистический анализ; интеллектуальный анализ.

**Алхасов Станислав Сергеевич**
**Целых Александр Николаевич**, д.т.н., профессор
*Южный федеральный университет*
*alkhasov@sfedu.ru; ant@sfedu.ru*

## ОСНОВНЫЕ ЭЛЕМЕНТЫ БЛОКА ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ В ПРИКЛАДНЫХ ЗАДАЧАХ АНАЛИЗА ДАННЫХ

**Введение.** Предварительная обработка входных данных является исключительно важным этапом анализа данных. Эффективное применение методик интеллектуального анализа данных (data mining) возможно в большинстве случаев лишь после обработки первоначальных данных. В настоящей статье представлены основные подходы к предварительной обработке входных данных на примере результатов измерений в задачах экологического мониторинга водных сред.

Аналитические мониторинговые системы работают не только с разнородными данными, но и в режиме неравномерности приема этих данных [1; 2]. Поэтому без предварительной обработки измерительных данных невозможна верная интерпретация экологической информации. Для этих целей перспективно использовать язык программирования *R* и интегрированную среду разработки *RStudio*. Важными достоинствами языка *R*, среды *RStudio* и дополнительных подключаемых библиотек являются свободный характер их распространения и использования, а также кроссплатформенность среды [5].

Комплексная система экологического мониторинга в режиме реального времени оперирует различными данными как количественного, так и качественного характера. Следовательно, такие данные соответствуют номинальным, порядковым и абсолютным шкалам.

Далее представлены основные задачи предварительной обработки данных комплексной системы импактного экологического мониторинга, а также обозначены основные подходы к их практической реализации в среде *RStudio* на языке программирования *R* [1].

**Сравнение рядов данных.** Сравнение рядов данных (оценка достоверности различий) используется для выявления постов экологического мониторинга, которые избыточно дублируют друг друга. Особенно важным является сравнение рядов данных для оптимальной работы мобильных постов экологического мониторинга, маршрут которых и места забора проб статистически определяются на основании архивных данных. Оценка достоверности различий рядов важна и при оценке погрешностей измерений, при исключении выбросов в полученных данных. К примеру, в точке А известны большие выбросы в измеренных значениях некоторых переменных, поэтому требуется не менее семи повторных измерений, тогда как в точке В достаточно трех измерений.

Для этой цели используют критерий Уилкоксона. Соответственно в языке *R* существует функция *wilcox.test*. Результат выполнения этой функции состоит из семи листов (List of 7), основным элементом которого является *p.value*. Если значение $p < 0{,}05$, то делается вывод о различии между двумя рядами данных [5]. После выполнения последней команды в консоль *RStudio* выводится результат, характеризующий близость данных между собой.

**Удаление выбросов.** Данные часто содержат выбросы, результаты измерений, выделяющиеся из общей выборки. Наличие выбросов сказывается на эффективности многих статистических методов. Поэтому исходные данные перед анализом нужно проверить на наличие выбросов [3]. Обычно выбросами считаются значения, не попадающие в диапазон $\left[q_1 - 1{,}5(q_3 - q_1);\quad q_3 + 1{,}5(q_3 - q_1)\right]$, где $q_1$ и $q_3$ – 1-й и 3-й квартили, а $(q_3 - q_1)$ – межквартильный интервал.

Функция *quantile* сохраняет квартильные значения для 0%, 25%, 50%, 75%, 100%. Соответственно 2-е и 4-е значения содержат $q_1$ и $q_3$. Аргумент *type* определяет способ расчета квартилей. По умолчанию используется *type=7*. Для вывода числового значения, например, $q_1$, нужно записать строку *q1 <- print(q[[2]])*.

Визуализировать выбросы возможно посредством построения диаграммы размаха («ящик с усами», *boxplot*) с использованием библиотеки *ggplot2*.

**Стандартизация данных.** Двумя наиболее используемыми подходами стандартизации данных, т.е. приведения разнородных данных к некоторому конкретному диапазону значений, являются *z*- и *minmax*-нормализация. В программном коде *R* они могут быть реализованы как в традиционной форме, так и в усовершенствованной (замена среднего арифметического на медиану и т.п.).

*z*-нормализацию можно провести с помощью команды

*data2 <- scale(data1,center=TRUE,scale=TRUE)*

*minmax*-нормализация может быть проведена посредством следующей команды:

*data2 <- (data1 - min(data1,na.rm=TRUE))/(max(data1,na.rm=TRUE)-min(data1,na.rm=TRUE))*

**Работа с пропущенными значениями.** Сложным информационным системам, получающим данные из множества источников разной природы, как правило, приходится оперировать с пропущенными значениями. Во-первых, может не совпадать частотность проведения различных анализов, что связано либо с неполной автоматизацией сбора данных, либо с человеческим фактором. Во-вторых, пропущенные значения могут возникать в результате сбоев в работе некоторых мультисенсорных систем. Таким образом, требуется сформулировать алгоритм работы с пропущенными значениями во входных данных. Удаление переменных с пропущенными значениями является крайне нерациональным подходом. Удаление наблюдений, в которых присутствуют пропущенные значения, также не всегда рационально. Поэтому простейшим решением в таких ситуациях является замена пропущенных значений средними арифметическими по каждой конкретной переменной. В зависимости от структуры данных вместо средних могут быть взяты медианы [5].

Данный подход может быть усовершенствован так, чтобы средние или медианы рассчитывались не для всех непропущенных значений переменных, а лишь для *K* значений до и после пропущенного значения. Также могут быть использованы более сложные способы замены пропущенных значений, основанные на кластеризации и регрессии [7].

**Метод главных компонент.** Полученные данные могут содержать избыточное количество переменных, коррелированных между собой, что может негативно сказываться на эффективности многих методов статистического и интеллектуального анализа данных. Чтобы снизить размерность данных, применяют метод главных компонент [8]:

*data.pca <- prcomp(data[,2:ncol(data)],center=TRUE,scale.=TRUE)*

Получив массив главных компонент, можно вывести график зависимости дисперсии от числа главных компонент [4] (Рис. 1):

*plot(data.pca, type = "l")*

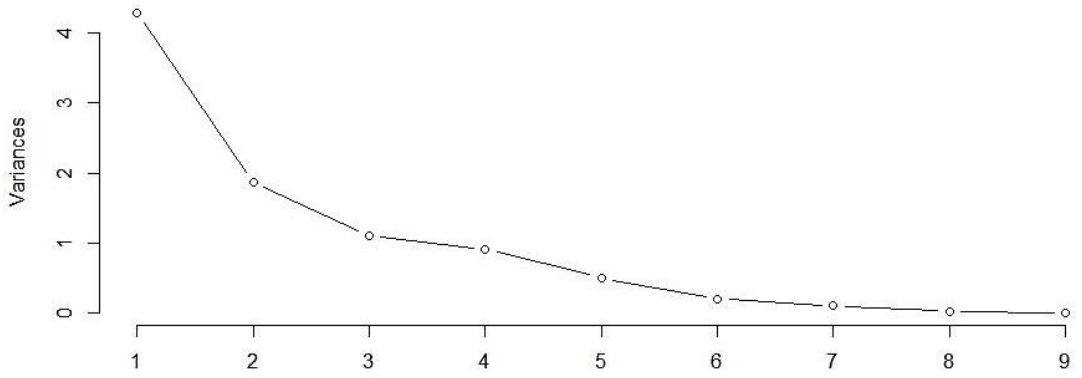


**Рис. 1.** *Зависимость дисперсии от числа главных компонент*

Сделать вывод о числе далее используемых переменных можно после выполнения команды *summary(data.pca)*

**Отправка *email*-уведомлений.** Возможность в реальном времени управлять как автоматизированным отбором проб мультисенсорными системами, так и проведением лабораторных физико-химических анализов важна для построения комплексной системы импактного экологического мониторинга. Выполнение отбора проб мультисенсорными системами и проведение лабораторных анализов определяются командами, передаваемыми с центрального компьютера посредством *email*-уведомлений. Для этих и подобных целей в языке существует несколько пакетов, в частности *mailR* [6]. Такой подход позволяет эффективно перераспределять усилия лабораторного персонала в нужных направлениях и в нужное время, избегая при этом избыточного использования рабочей силы, перерасхода химических реактивов и повышенного износа аналитического лабораторного оборудования. В работе [2] представлен вариант реализации подобного модуля отправки *email*-уведомлений в системе *MATLAB*. Аналогичным образом он реализуется функцией *send.mail* из библиотеки *mailR* в среде *RStudio*.

**Заключение.** Язык программирования *R* является современным, многофункциональным и широкодоступным средством для сбора и передачи, предварительной обработки, статистического и интеллектуального анализа данных. В сети Интернет доступны многочисленные библиотеки, позволяющие расширить набор доступных операций с данными. Более того, возможно создание собственной библиотеки для решения специфических статистических и аналитических задач. В сравнении со средой *MATLAB*, язык *R* более гибок в возможности построения сложных программных конструкций, что важно, поскольку оба языка – интерпретируемые. Вышеперечисленные факты являются основной причиной для использования языка программирования *R* и среды *RStudio* в построении информационно-управляющего блока комплексной системы импактного экологического мониторинга. И в особенности язык *R* важен на этапах предварительной обработки и интеллектуального анализа данных.

*Список литературы*

1. **Алхасов С. С., Милешко Л. П., Целых А. А.** Определение концентраций ионов тяжёлых металлов посредством блока обработки данных мультисенсорной системы для мониторинга водных сред // Известия ЮФУ. Технические науки. 2013. № 4 (141). С. 161-168.
2. **Алхасов С. С., Милешко Л. П., Шестова Е. А.** Основы построения мультисенсорных систем для экологического мониторинга водных сред: учебное пособие. Ростов-на-Дону: Изд-во ЮФУ, 2014. 99 с.
3. **Алхасов С. С., Целых А. Н.** Основные подходы к построению информационной системы для моделирования оттока клиентов услуг связи // Известия ЮФУ. Технические науки. 2015. № 2 (163). С. 106-115.
4. **Мастицкий С. Э.** R: Анализ и визуализация данных [Электронный ресурс]. URL: http://r-analytics.blogspot.ru/2012/05/blog-post_20.html#.VZ1TpPntmfg (дата обращения: 21.02.2016).
5. **Шипунов А. Б., Балдин Е. М., Волкова П. А., Коробейников А. И., Назарова С. А., Петров С. В., Суфиянов В. Г.** Наглядная статистика. Используем R! М.: ДМК Пресс, 2014. 298 с.
6. **A Utility to Send Emails from the R Programming Environment** [Электронный ресурс]. URL: https://github.com/rpremraj/mailR (дата обращения: 20.02.2016).
7. **Bolker B. M.** Ecological Models and Data in R. Princeton: Princeton University Press, 2008. 408 p.
8. **Computing and Visualizing PCA in R** [Электронный ресурс]. URL: http://www.r-bloggers.com/computing-and-visualizing-pca-in-r/ (дата обращения: 21.02.2016).

**KEY ELEMENTS OF MEASUREMENTS RESULTS PRE-PROCESSING UNIT IN APPLIED PROBLEMS OF DATA ANALYSIS**

**Alkhasov Stanislav Sergeevich**
**Tselykh Aleksandr Nikolaevich**, Doctor in Technical Sciences, Professor
*Southern Federal University*
*alkhasov@sfedu.ru; ant@sfedu.ru*

The paper presents the key components of the unit of data pre-processing required for the correctness of the further statistical and intellectual analysis of diverse input data of the complex information and analytical system. The main methods of data pre-processing used to compare data series, to remove overshoots and to reduce dimensionality are described.

*Key words and phrases:* pre-analysis of data; information and analytical system; *R* programming language; diverse input data; statistical analysis; intellectual analysis.

---

УДК 030:614.2
**Филологические науки**

*В статье освещается результат совместной работы специалистов-медиков и филологов по созданию терминологического словаря-справочника по общественному здоровью и здравоохранению с английскими эквивалентами и примерами сочетаемости слов. Такой многофункциональный словарь не только впервые дает достаточно полное (более 3500 терминов и терминологических словосочетаний) описание понятий в сфере организации здравоохранения, но и, по мнению авторов, способствует повышению коммуникативной компетенции специалиста-медика.*

*Ключевые слова и фразы:* специальная лексика организации здравоохранения; многофункциональный терминологический словарь; дефиниция термина; безэквивалентная лексика; английский эквивалент; лексическая сочетаемость.

**Багметов Николай Петрович**, к. мед. н.
**Мульганова Татьяна Борисовна**, к. филол. н., доцент
*Волгоградский государственный медицинский университет*
*bagmetoff.nickolai@yandex.ru*

## ИЗ ОПЫТА СОЗДАНИЯ ТЕРМИНОЛОГИЧЕСКОГО МНОГОФУНКЦИОНАЛЬНОГО СЛОВАРЯ ПО ОРГАНИЗАЦИИ ЗДРАВООХРАНЕНИЯ

Профессиональная подготовка современного врача в России предполагает не только глубокое изучение медицины как «системы научных знаний и практической деятельности, направленной на лечение и предупреждение болезней человека» [1, с. 1], но и изучение сложной и многогранной структуры общественного здравоохранения, методов эффективного администрирования и управления им.

Аспект профессиональной подготовки врача в медицинских вузах России определяет содержание учебной дисциплины «Общественное здоровье и здравоохранение», а также программы повышения квалификации на факультете усовершенствования врачей (ФУВ) специалистов-медиков, проработавших в системе здравоохранения не менее 10-ти лет.