О ПРОБЛЕМАХ ВИЗУАЛИЗАЦИИ РЕЗУЛЬТАТОВ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ АНКЕТНЫХ ДАННЫХ

М.С. Ознобихина, А.Ю. Тимофеева
Новосибирский государственный технический университет,
Россия, г.Новосибирск, пр. К.Маркса, 20, 630073
E-mail: oznobikhina.ms@yandex.ru, a.timofeeva@corp.nstu.ru

PROBLEMS EMERGING IN VISUALISING OF THE RESULTS OF STATISTICAL PROCESSING OF SURVEY DATA

M.S. Oznobikhina, A.Yu. Timofeeva
Novosibirsk State Technical University, Russia, Novosibirsk, Prospekt K. Marksa, 20, 630073
E-mail: oznobikhina.ms@yandex.ru, a.timofeeva@corp.nstu.ru

***Abstract**. Visualization provides an accurate data transferring by means of charts, graphs, or tables. Visualization of categorical data and its concordance comes with lots of problems. We analyze different ways to visualize such types of data using diagram in parallel coordinates, Venn diagram, bubble chart, mosaic plot.*

При визуализации данных анкетных опросов в исходном «сыром» виде, как правило, возникают сложности, и требуется их предварительная, в том числе статистическая, обработка. В первую очередь необходимо удалить пропуски. Во-вторых, часто имеет смысл упорядочить значения переменных. Если значения числовые, они естественным образом ранжируются вдоль числовой оси. Для текстовых ответов чаще всего применяется алфавитный порядок. Однако он никак не отражает особенности данных, поэтому более корректна сортировка вариантов по убыванию частоты. Тем не менее, не всегда это приемлемо. Во многих случаях существует естественный порядок уровней, например, если используется шкала согласия, варианты ответа представляют упорядоченные категории (например, уровень образования). В этом случае порядок должен быть установлен пользователем (ответы кодируются в нужном порядке), в остальных случаях возможна автоматическая сортировка значений переменных.

В-третьих, часто приходится прибегать к удалению значений: исключению резко выделяющихся наблюдений в числовых данных, а также ненужных уровней в категориальных данных. Методы отбраковки выбросов детально разработаны в статистике. Что касается ненужных уровней, то это, как правило, варианты «нет ответа», «затрудняюсь ответить», а также варианты с очень малыми частотами. Такое удаление можно выполнить автоматически, не прибегая к помощи пользователя. Наконец, в-четвертых, для большей наглядности часто необходима группировка значений. Для числовых данных она сводится к выбору числа и ширины интервалов группирования, это процесс может быть сведен к автоматической процедуре, использующей только исходный массив данных. С текстовыми данными дело обстоит сложнее. Здесь уровни могут быть сгруппированы только по смыслу, для чего, естественно, необходимо привлекать эксперта. Однако, возможно, автоматическое выделение группы «прочее», содержащей уровни с малыми частотами.

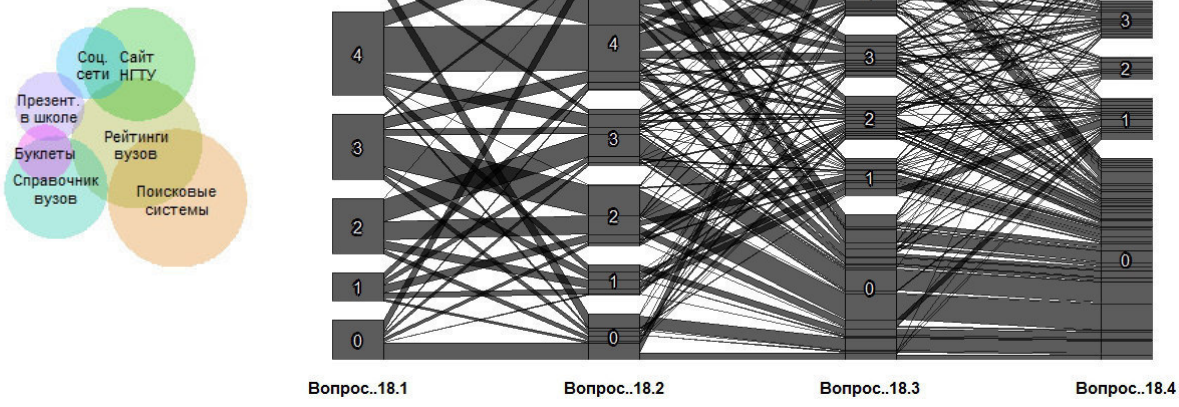
Типы диаграмм, которые можно построить по предварительно обработанным данным, зависят от типа этих данных. Наиболее широкий спектр способов визуализации применим для представления числовых данных. Отметим, однако, что бывает полезно рассматривать числовые данные с малым числом

вариантов (например, число лет обучения) как категориальные, это позволяет более наглядно их отобразить. Поэтому предлагается для определения типа данных использовать число уникальных значений переменной. Если оно невелико, то переменную лучше рассматривать как категориальную.

Возможности для визуализации категориальных данных несколько ограничены. В основном используются столбчатые и круговые диаграммы, которые показывают распределение частот. Однако круговые диаграммы часто критикуют. По ним, во-первых, сложно сравнивать значения сегментов, когда их количество велико, – диаграмма теряет наглядность, так как различие между сегментами становится несущественным. Во вторых, круговую диаграмму нельзя использовать, когда сумма относительных частот превышает 100%, например, когда один респондент мог выбирать несколько вариантов ответа. Для анализа ответов на такие вопросы можно использовать диаграмму Венна. На рис. 1а представлен пример диаграммы, отражающей внешние источники информации, которые используют абитуриенты при выборе вуза. По рис. 1а можно понять, какая доля аудитории охватывается сразу несколькими источниками, что позволяет выбрать подходящие рекламные средства.

Для визуализации взаимосвязи между категориальными переменными рекомендуется использовать диаграмму в параллельных координатах [1], пузырьковую диаграмма [2], мозаичную диаграмму [3].

Рис. 1б изображает в параллельных координатах переменные, измеренные по шкале согласия (от 0 до 5 баллов). Студенты давали ответы на вопросы №18 «Я полагаю, что ...» с утверждениями: 1) выпускники НГТУ хорошо трудоустраиваются; 2) НГТУ - лидирующий вуз по качеству подготовки; 3) решающим при выборе вуза явилась доступность поступления; 4) стоит прислушаться к рекомендациям родителей по выбору этого вуза. Видно, например, что большая часть респондентов, согласных с первым вариантом, соглашаются и со вторым, то есть между ними есть некоторая ассоциативная связь.

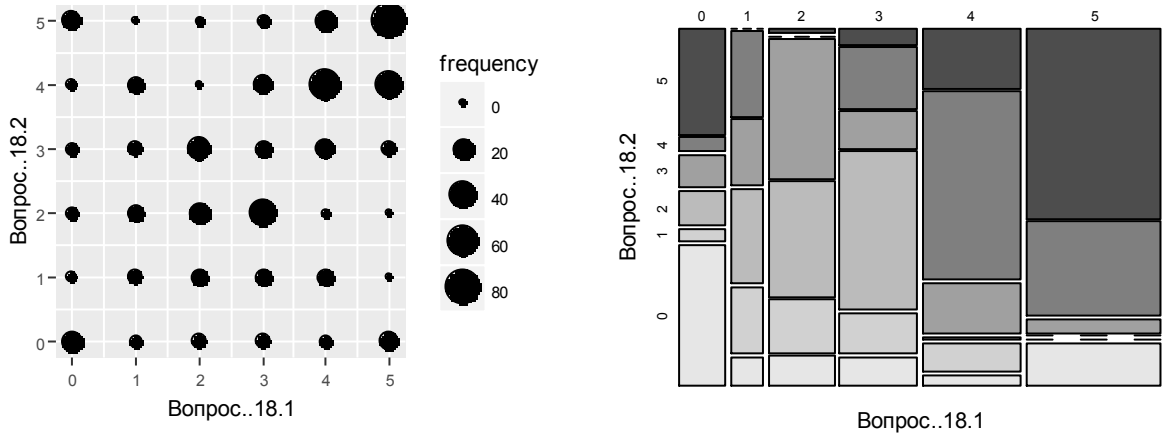


(а) *(б)*

Рис. 1. Диаграммы Венна (а) и в параллельных координатах (б)

Взаимосвязь между категориальными переменными наглядно представляется с помощью пузырьковой (рис. 2а) или мозаичной диаграммы (рис. 2б). Площади фигур на диаграммах пропорциональны частотам. Пузырьковая диаграмма не всегда обеспечивают требуемую наглядность и показывает взаимосвязь только двух переменных. Мозаичная диаграмма позволяет получить

графическое изображение таблицы сопряженности [4]. Такие диаграммы дают возможность отразить большой объем информации и могут использоваться для визуализации взаимосвязей как двух, так и большего числа переменных. С помощью цветов и оттенков в мозаичной диаграмме возможно отображение остатков от подобранной модели [3].



(а) *(б)*

Рис. 2. Визуализация взаимосвязи между двумя порядковыми категориальными переменными с помощью (а) пузырьковой диаграммы, (б) мозаичной диаграммы

Предварительная обработка данных и определение типов переменных позволяет реализовать автоматический выбор подходящего типа диаграммы для наиболее удобной, наглядной и простой визуализации данных. На этих принципах авторами реализована автоматизированная система визуализации анкетных данных с использованием интегрированной среды разработки RStudio.

Работа выполнена при поддержке Совета по грантам Президента РФ для государственной поддержки молодых российских ученых (проект МК-5385.2016.6).

СПИСОК ЛИТЕРАТУРЫ

1. Шипунов А.Б. Наглядная статистика. Используем R! [Электронный ресурс]. – Режим доступа: https://cran.r-project.org/doc/contrib/Shipunov-rbook.pdf. – 28.02.17.
2. Vidmar G., Rode N. Visualising concordance // Computational Statistics. – 2007. – Vol. 22. – No. 4. – P. 499-509.
3. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.
4. Pilhoefer A., Unwin A. New approaches in visualization of categorical data: R package extracat // Journal of Statistical Software. – 2013. – Vol. 53. – No. 7. – P. 1-25.