Санкт-Петербургский государственный университет

О.А. Игнатьева

# КОЛИЧЕСТВЕННЫЕ МЕТОДЫ АНАЛИЗА ПУБЛИЧНОЙ ПОЛИТИКИ

Учебно-методическое пособие

Санкт-Петербург
Астерион
2021

Рекомендовано к изданию Учебно-методической комиссией по УГСН 41.00.00 «Политические науки и регионоведение» СПбГУ для обучения студентов по направлению 5.5.2. «Политические институты, процессы, технологии».

**И26 Игнатьева О.А.** Количественные методы анализа публичной политики : учебно-методическое пособие / О.А. Игнатьева. – СПб. : Астерион, 2021. – 56 с.

ISBN 978-5-00188-122-3

Учебно-методическое пособие «Количественные методы анализа публичной политики» представляют собой основу методического сопровождения базового курса «Количественные методы анализа публичной политики», преподаваемого в рамках магистерской программы «Политическое управление и публичная политика» (направление обучения 41.04.04, шифр специальности 5.5.2 «Политические институты, процессы и технологии»). Освоение дисциплины направлено на ознакомление студентов с принципами организации научного исследования в области анализа публичной политики; изучение основных подходов к проведению количественных исследований публичной политики; изучение количественных методов и программного обеспечения для анализа публичной политики; формирование у студентов навыков подготовки аналитических записок по проблемам публичного управления.

**ISBN 978-5-00188-122-3**



## СОДЕРЖАНИЕ

Введение ..... 4

Глава 1. Теоретико-методологические основания работы с данными ..... 5

Глава 2. Статистика с использованием языка программирования R ..... 23

Глава 3. Сетевой анализ с использованием Pajek ..... 41

Список использованной литературы ..... 54

## ВВЕДЕНИЕ

Учебно-методическое пособие «Количественные методы анализа публичной политики» представляют собой основу методического сопровождения базового курса «Количественные методы анализа публичной политики», преподаваемого в рамках магистерской программы «Политическое управление и публичная политика» (направление обучения 41.04.04, шифр специальности 5.5.2 «Политические институты, процессы и технологии»). Освоение дисциплины направлено на ознакомление студентов с принципами организации научного исследования в области анализа публичной политики; изучение основных подходов к проведению количественных исследований публичной политики; изучение количественных методов и программного обеспечения для анализа публичной политики; формирование у студентов навыков подготовки аналитических записок по проблемам публичного управления. Для изучения данного курса студенты должны знать материал следующих учебных дисциплин: «Информационные технологии в политической науке и образовании», «Государственная политика и управление», «Политический анализ и прогнозирование», «Сетевой подход в государственной политике и управлении». Учебно-методическое пособие содержит основы разработки дизайна исследования, работы с языком программирования R при решении статистических задач и программным обеспечением Pajek.



## ГЛАВА 1. Теоретико-методологические основания работы с данными

### Данные как объект изучения

Статистика была первой научной отраслью, которая занималась обработкой социальных данных. Данные для нее собирались в основном путем анкетирования. Разрабатывались разные способы формирования репрезентативной выборки, которая бы корректно отражала весь массив данных.

Составление репрезентативной выборки – это отдельный раздел в методологии научного исследования [1, с. 29–35]. Только вероятностные выборки, составленные с учетом требований теории вероятности, могут служить основанием для экстраполяции результатов исследования на всю генеральную совокупность [2, p. 210–228]. Составление вероятностных выборок требует наличие списка элементов, входящих в генеральную совокупность, что не всегда возможно. В связи с этим мы начнем с описания выборок, которые сформированы без учета требований теории вероятности и поговорим об их ограничениях.

Первым типом формирования выборки без учета требований теории вероятности является выборка по случаю. Например, мы можем сформировать такую выборку, останавливая прохожих у входа на рынок в обеденное время. Такая выборка оправдана только в том случае, если мы хотим выяснить характеристики людей, посещающих рынок в данный момент времени, или если у нас нет возможности составить другую более репрезентативную выборку. Обычно к такому способу формирования выборки прибегают преподаватели, которые опрашивают студентов, посещающих поточные лекции. Такая выборка может позволить протестировать инструментарий, но такой метод формирования выборки нельзя использовать с целью обобщения выводов на генеральную совокупность.

Вторым типом выборки, не основанной на теории вероятности, служит целевая (оценочная) выборка. Элементы для такой выборки отбираются на основании суждения о том, какие из них более соответствуют целям исследования. Данная выборка может быть использована, когда мы хотим изучить небольшое подмножество элементов более крупной совокупности, которые легко идентифицировать, но сложно перечислить. Иногда целевая выборка может быть использована для анализа девиантных кейсов, чтобы лучше понять типичные кейсы. Данная вы-

борка подходит для тестирования инструментария или разведывательного исследования. Экстраполяцию выводов на генеральную совокупность лучше не делать.

Третьим типом «невероятностной» выборки является выборка по типу «снежного кома». Выборка по типу «снежного кома» может быть использована для разведывательного исследования, когда нет возможности перечислить элементы, входящие в генеральную совокупность, и сложно получить доступ к представителям данной социальной группы, например, при исследовании проблемы бездомности [3, с. 48–51].

Четвертым типом выборки, в основе которой нет положений теории вероятностей, является квотная выборка. Как и в случае с вероятностной выборкой, квотная выборка отвечает требованиям репрезентативности. Для формирования данной выборки используется таблица, в каждой из ячеек которой прописана пропорция населения, соответствующая определенной возрастной группе, полу или классу. Квотная выборка формируется в соответствии с требованиями репрезентативности, и повторяет тоже процентное соотношение внутри выборки, как и в генеральной совокупности. Однако при использовании квотной выборки могут возникнуть такие проблемы как своевременное обновление данных в ячейках, что может быть трудозатратным, и могут быть отклонения в характеристиках респондентов, связанные с поведением интервьюеров. Впервые квотная выборка была применена институтом исследования общественного мнения Гэллапа в США и была значительным шагом на пути к более точному предсказанию результатов выборов Президента. Так, благодаря данному типу выборки, Гэллап смогут точно предсказать результаты выборов Президента США в 1941 г.

Научно обоснованные выборки, результаты анализа которых можно распространять на генеральную совокупность, должны исходить из положений теории вероятностей, которая позволяет придать репрезентативный характер исследуемой выборки [4, с. 41–46]. Всего существует четыре типа вероятностных выборок: простая случайная выборка, систематическая выборка, стратифицированная выборка и многоступенчатая кластерная выборка [5, с. 197–220]. В основе этих выборок лежит необходимость использования нумерованного списка для элементов генеральной совокупности. Так, для простой случайной выборки нам необходимо использовать таблицу случайных чисел. Например, у нас есть генеральная совокупность размером одна тысяча учащихся определенного вуза. Нам нужно сделать выборку, состоящую из ста человек.



Мы нумеруем все элементы генеральной совокупности и затем на основе таблицы случайных чисел отбираем сто элементов для выборки.

В случае с систематической выборкой мы отбираем каждую k-ую единицу в списке и начинаем прохождение этого списка со случайного элемента. Например, если наша генеральная совокупность составляет одну тысячу человек, а выборка должна быть сто человек, то выборочный интервал составит десять. Мы будем должны отбирать из списка каждый десятый элемент. Данный дизайн менее трудоемкий, чем использование простой случайно выборки и дает схожие результаты. Но при использовании систематической выборки мы может столкнуться с проблемой цикличности данных, если данные в списке организованы в табличном виде, например, перечисление военнослужащих может быть сформировано в виде отрядов по десять человек, начиная с сержантов. Тогда у нас в выборке могут оказаться одни сержанты, и выборка будет чрезмерно направленной. Таким образом, нужно внимательно смотреть за дизайном организации данных в списке.

Стратифицированная выборка может расцениваться как модификация для трех других вероятностных выборок. Она позволяет достигнуть большей репрезентативности данных. Стратифицированная выборка – это группировка единиц, составляющих генеральную совокупность в однородные группы (страты) перед осуществлением отбора данных. Есть два способа осуществления стратифицированного отбора данных. Во-первых, мы можем распределить данные на две группы, например, по полу студентов и затем на основе таблицы случайных чисел отобрать нужные элементы из этих двух групп. Во-вторых, мы можем распределить данные по группам, записать эти группы друг под другом, и затем сделать из этих групп систематическую выборку.

И, наконец, мы можем сформировать многоступенчатую кластерную выборку, которая представляет собой движение от больших групп к индивидам, но без использования большого списка элементов. Это имеет смысл, когда нам нужно проанализировать жителей какого-нибудь города. Для этого, например, мы составляем список жилых домов, затем случайным образом отбираем из них выборку из ста домов. После этого мы составляем список домохозяйств по фамилиям хозяев, отбираем сто домохозяйств, составляем список членов домохозяйств. После этого мы из списка случайным образом отбираем членов семей для опроса [6, с. 23–24].

Таким образом, проведение социального исследования должно подчиняться строгим критериям научности, как на этапе формирова-

ния программы исследования и отбора дизайна для выборки данных из генеральной совокупности, так и на этапе проведения исследования с последующей интерпретацией данных на основе выбранной в начале исследования теоретической рамки.

**Методы работы с данными**

При работе с данными обычно выделяют методы сбора и методы анализа данных. Данное учебно-методическое пособие посвящено рассмотрению обоих видов. Эмпирическое исследование в противоположность теоретическому представляет собой совокупность последовательных методологических, методических и организационных процедур, направленных на получение достоверных данных об изучаемом явлении для их последующего практического применения.

Первой классификацией методов исследования можно назвать их деление на теоретические и эмпирические. К теоретическим методам исследования относятся сравнительно-исторический, структурно-функциональный, герменевтический и системный. К эмпирическим методам исследования относят методы наблюдения, анализа документов, методы анкетирования и интервьюирования. Эмпирические методы в свою очередь подлежат классификации. Иногда их делят на опросные и неопросные методы. При этом к опросным методам относятся метод анкетирования, метод интервьюирования, метод фокус-групп, а к неопросным методам: методы анализа документов, наблюдение. Данная классификация относится к традиционным методом сбора данных с их последующей обработкой статистическими методами. Также эмпирические методы исследования могут быть количественными и качественными. К количественным методам в социологии относятся метод контент-анализа документов, парсинг, метод опроса и прямого наблюдения. К качественным методам сбора информации относят глубинные интервью, фокус-группы и кейс-стади.

Использование методов наблюдения позволяет получить факты (свидетельства) о состоянии объективной реальности. Метод наблюдения является одним из самых эффективных методов сбора информации, поскольку предполагает вовлечение ученого в сбор информации. В свою очередь методы наблюдения можно подразделить на формализованное и неформализованное, включенное и невключенное виды наблюдения.

Необходимо отметить, что наблюдение является методом сбора информации посредством непосредственной регистрации исследователем событий и условий в поле. В отличие от обыденного наблюдения научное наблюдение является целенаправленным восприятием для достижения научных целей и задач. Объективность в плане возможности контроля путем повторного наблюдения или с использованием других методов исследования является отличительной чертой научного наблюдения.



Научные наблюдения бывают следующих видов: формализованное, и неформализованное, включенное и невключенное.

Формализованное наблюдение структурируется жесткой программой и предполагает разработку и использование инструментария (карточек и протоколов). Неформализованное наблюдение может ограничится общим планом. Необходимо отметить, что неформализованное наблюдение не позволяет собрать данные о тенденциях и закономерностях изучаемых процессов. Это всегда основа для проведения более тщательного формализованного наблюдения или эксперимента.

Включенность наблюдателя в процесс наблюдения предполагает выделение двух типов наблюдения: включенного и невключенного. При проведении невключенного наблюдения исследователь отмечает факты пассивно, наблюдая за явлением со стороны. Включенное наблюдение предполагает погружение в ситуацию и часто называется наблюдением «в маске», так как исследователь скрывает не только сам факт проведения исследования, но и свою роль.

Основным источником информации для исследователя служит текст, который размещается как на электронных носителях, так и на бумаге, и в интернете. Документ является материальным носителем, содержащим информацию в установленной форме и по установленным правилам, необходимую для научно-исследовательских и практических целей.

Для логического метода анализа документов характерная некоторая субъективность восприятия данных. Для того, чтобы сделать качественный анализ объективным, необходимо использовать логические приемы (индукция, дедукция, классификация, систематизация, анализ, синтез, сравнительный анализ, аналогия).

Этапы анализа документа:

1. Выбор документа в соответствии с целями и задачами исследования.
2. Анализ контекста создания документа.
3. Логический анализ документа

При интуитивном анализе документов используются следующие логические процедуры анализ документов:

- индукция – метод рассуждения от частного к общему;
- дедукция – метод рассуждения от общего к частному;
- классификация – дифференциация идей текста по релевантному признаку;
- систематизация – объединение идей текста вокруг ключевой идеи по определенному принципу;
- анализ (разложение, расчленение, разборка) – метод исследования, характеризующийся выделением и изучением отдельных частей объектов исследования;
- синтез – представляет собой процесс соединения текстовых данных в единое целое и обычно используется вместе с анализом.

Контент-анализ является количественном методом анализа документов, осуществляющем перевод текстовых данных в количественные показатели с последующей их статистической обработкой и интерпретацией. Сущность контент-анализа состоит в подсчете интересующих нас ключевых слов с выяснением скрытого смысла текста.

Тексты в контент-анализе – это книги, статьи, рекламные тексты, речи выступлений, тексты на веб-страницах.

Необходимо отметить, что контент-анализ используется только как дополнительный метод при проведении значительного исследования, которое базируется на использовании программы эмпирического исследования.

Контент-анализ предполагает, что исследователь определяет объект, категории и единицы анализа, выбирает статистический метод и начинает сбор данных. Контент-анализ позволяет увидеть в документе его скрытый смысл.

Этапы контент-анализа:

1. выделяют единицы анализа и переводят их в числовой формат;

2. проводят подсчет частотных распределений методами статистического анализа;

3. исследователь проводит интерпретацию полученных результатов.

Объект контент-анализа – это тексты, публикуемые в книгах, газетах, на веб-страницах, также это ответы на открытые вопросы анкет. Единицы анализа – это темы, образы, метафоры, примеры, проблемы, аналогии, каламбуры, надписи на стенах, фамилии политиков, названия партий и т.д.. Выбор единицы анализа зависит от исследовательской программы, объекта, предмета исследования и цели. Главная задача контент-анализа заключается в установлении единиц анализа. При этом единицы анализа должны быть подчиняться требованиям единообразия, т. е. легко и однозначно идентифицироваться в тексте. Слова, выбранные для подсчета, должны быть интересны для последующей политологической, социологической, социальной интерпретации.



Метод анкетирования является еще одним распространенным количественным методом сбора данных. При составлении анкеты действуют не только логические, но психологические и эстетические законы. Составление анкеты – это искусство, которое опирается на определенные правила, но полностью не подчинено им. Стержень анкеты – это композиция расположения вопросов. Композиция – это «составление», «связывание». Структура – это совокупность устойчивых связей объекта, обеспечивающих его целостность и тождественность самому себе.

Анкета уникальна, что объясняется рядом причин:

1) создается под конкретное исследование;

2) отражает индивидуальное мировоззрение исследователя;

3) адресована для изучения именно этого объекта исследования;

4) при составлении новой анкеты возможна преемственность отдельных вопросов, блоков вопросов, но не анкеты в целом.

Структура анкеты состоит из следующих частей:

Введение (цель исследования, правила заполнения анкеты, личная выгода для респондента от участия в опросе, обещание анонимности или конфиденциальности).

Реквизитная часть (название анкеты, дата, время и место проведения опроса, фамилия интервьюера).

Информативная часть (содержательные вопросы с нарастанием степени сложности).

Классификационная часть (паспортичка).

Заключительная часть (благодарность за участие).

Части 1, 2, 5 – обслуживающий, вспомогательный аппарат анкеты, части 3 и 4 – основная, базисная часть анкеты. «Эффект эха» означает, что вопросы расположены в анкете так, что ответ на один вопрос предполагает схожие ответы на другие вопросы. При составлении анкеты нужно соблюдать следующие требования: термины должны быть понятны респонденту; базисная часть (информативная часть) должна содержать 20–30 вопросов; формулировка вопросов должна соответствовать исследовательской задаче; анкета должна соответствовать способностям респондента, не унижать его достоинства.

Программная логика вопроса не должна совпадать с логикой построения анкеты. Например, в соответствии с программной логикой выделяем и опрашиваем заинтересованные лица, затем незаинтере-

сованные лица в данном товаре. В логике построения анкеты должно быть сначала так: опрашиваем всех, затем только участников конфликта, затем незаинтересованные лица, а в заключении всех респондентов. Для получения достоверной информации и избегания «эффекта эха» сначала задаются вопросы о частных сторонах явления, а затем обобщающий вопрос. Вопросы должны идти от простых к сложному. Сначала задается общий вопрос для завязки диалога. Уточняется мнение респондента по каким-то вопросам. Далее идут более сложные вопросы, относящиеся к событиям, фактам. 2–3 самых сложных вопроса, требующих размышления и работы памяти, задаются «на пике». Затем все заканчивается «паспортичкой» (социально-демографической информацией). Вопросы должны идти от общего к конкретному (туннельный подход). Есть еще секционный подход (анкеты-омнибусы). Переключение между разными темами путем специальных вопросов.

Источниками информации для разработки анкеты служат чужие анкеты, литература, интуиция, предыдущий опыт и СМИ. Переменная исследования: объект, предмет (процесс, явление), интересующие исследователя. Переменная анкеты: вопрос анкеты (инструмент операционализации). Анкетный вопрос представляет совокупность знаний, включающую то, что нам известно и то, что мы хотим узнать. Выделяют три функции анкетных вопросов. Индикаторная функция направлена на получение искомой информации. Коммуникативная функция означает связь между исследователем и респондентом. Инструментальная функция характеризует измерительные возможности вопроса.

В данном учебно-методическом пособии пойдет речь о сборе текстовых данных посредством парсинга, работе с большими данными, а также обработки данных с использованием языка программирования R и сетевого анализа посредством Pajek.

### Уровни и виды социальных исследований

Когда мы планируем исследование. мы должны находится на позициях определенной научной парадигмы и выбрать в ее рамках соответствующую концепцию. Парадигма включает набор подходов и методов изучения реальности, которые определяют идеи о предмете исследования, а также устанавливает особые правила взаимодействия ученых друг с другом в рамках этих идей. Понятие парадигмы, как специфической формы научной работы, было предложен впервые Т. Куном в работе «Структура научных революций» [7, с. 120–132]. Необходимо



отметить, что научное сообщество делится на группы, которые работают в рамках определенной парадигмы. Мы можем говорить о научных парадигмах в рамках определенных наук. Так в рамках социологии можно выделить позитивизм, структурализм, структурный функционализм и социальный конструктивизм. В политической науке ключевыми парадигмами является институционализм, новый институционализм, структурный функционализм и аксиологических подход. Выбор той или иной парадигмы задает ракурс анализа и интерпретации данных, получаемых в ходе исследования. Однако и на уровне парадигмы существуют разные теории, и для того, чтобы корректно провести исследование и интерпретировать результаты, необходимо корректно определить не только парадигму, но и теорию. Дело в том, что каждая парадигма состоит из набора теорий и концепций. Например, неоинституциональная парадигма в политологии представлена теорией общественного выбора, теорией игр и сетевым анализом. Таким образом, когда мы готовимся проводить исследование нам необходимо определить не только парадигму, но также определить теорию или теоретическую рамку для концептуализации проблемы исследования и ее дальнейшей интерпретации.

Научное познание может осуществляться как на уровне логики с последующим подтверждением гипотезы эмпирическим материалом (дедуктивное обоснование), так и на эмпирическом уровне, когда мы начинаем со сбора эмпирических фактов и заканчиваем значимым выводом. Примером такой индуктивной теории является «обоснованная теория» Б. Глейзера и А. Страуса [8, с. 143–153]. Деление теорий на микро-, мезо- и макроуровни можно продемонстрировать на примере лестницы абстракций Дж. Сартори [9, с. 66–77]. Суть данной концепции заключается в том, что, находясь на нижнем уровне этой лестницы мы получаем знание, богатое деталями, но нам сложно сделать обобщение таких предметов на нижнем уровне абстракции, на среднем уровне мы можем уже распределить эти предметы по классам, на верхнем уровне у нас получаются генерализации, когда мы уже можем обобщить классы одним понятием. Движение по лестницы абстракций происходит с уменьшением детализации от нижнего к верхнему уровню знания. Но важно помнить, что не все знание мы получаем из эмпирики, есть знание, которое получается на верхнем уровне из логических рассуждений. Очень важно, чтобы при движении наверх контекстуальность знания уменьшалась, иначе мы можем получить обобщение детализированных описаний, что приведет к путанице.

Ученые, проводящие исследования, в рамках социальных наук могут быть классифицированы на две группы. Первые занимаются приращением знания ради знания, и мы можем назвать их фундаментальными учеными. Вторая группа состоит из ученых-прикладников, аналитиков, в чью задачу входит выявление проблемных сфер, их изучение с последующей формулировкой вариантов решения.

В данном учебном пособие мы будем говорить как об особенностях проведения социального исследования, так и об особенностях его разновидности – анализа публичной политики [10, с. 255–277]. Но необходимо отметить, что социальное исследование представляет собой объяснение социальных фактов через политические, психологические и экономические факты. Если строгое социологическое исследования зиждется на требовании свободы от оценочных суждений с точки зрения «плохо» или «хорошо», то некоторые виды социальных исследований в рамках политологии предполагают необходимость оценочного суждения.

Системный подход в политологии представляет совокупность методологических процедур, предложенных американским политологом В. Данном [11, p. 1–22]. В качестве таких методологических процедур выделяют мониторинг с совокупностью методов, используемых для проведения обычного социального исследования (опрос, наблюдение, кейс-стади, статистический анализ). Затем оценивание политики и политических программ, которое позволяет нам сделать вывод о степени достижения проводимой политикой намеченных в программе целей. Далее это прогнозирование, которое дает возможность сформировать представление об ожидаемых результатах политики. И, наконец, это рекомендации, которые позволяют нам сделать вывод касательно предпочитаемых вариантов проведения политики [11, p. 6–10] Необходимо отметить, что мониторинг и оценивание политики относятся к ретроспективному анализу (ex post). Они проводятся социологами и политологами, и касаются результатов реализации политических программ, поэтому данные процедуры не позволяют исправить то, что уже создано. Прогнозирование и рекомендации относятся к перспективному анализу (ex ante). Его проводят экономисты и специалисты в области теории принятия решений. Данный вид анализа политики проводится для того, чтобы предугадать варианты ее реализации и по возможности избежать ошибок при решении той или иной политической проблемы.

Мониторинг является центральной методологической процедурой в анализе публичной политики. Существуют четыре методологических подхода при анализе публичной политики: анализ социальных систем, социальный эксперимент, социальный аудит, синтез исследования и практики. Все они требуют разного типа контроля и информации [11, p. 276–302].



*Таблица 1*

**Основные ограничения четырех методологических подходов в мониторинге политики (Источник: W. Dunn, 2004. p. 284)**

| Подход | Тип контроля | Тип информации |
|---|---|---|
| Анализ социальных систем | Количественный | Доступная или новая |
| Социальный эксперимент | Прямые манипуляции и количественный тип | Новая |
| Социальный аудит | Количественный и / или качественный | Новая |
| Синтез исследования и практики | Количественный и / или качественный | Доступная |

Анализ социальных систем основан на использовании социальных показателей. Социальные показатели представляют собой статистику, которая позволяет измерять социальные условия и их изменения за время реализации политической программы. Социальные показатели являются выражением смысла переменных, которые характеризуют индивида, событие, объект, имея разные числовые значения.

Существует два способа определения переменных: конструктивное определение и операционное определение. Конструктивное определение представляет собой словесное определение из словаря с использованием синонимов. Например, мы можем определить «образовательные возможности», как свободу в определении образовательной среды в соответствии с чьими-либо способностями. Конструктивное определение переменной указывает на приблизительное отношение данной переменной к реальности.

Однако мы можем понять политические действия и их результаты только косвенно, используя операционные определения и показатели переменных. Операционные определения придают значение переменной посредством уточнения операций, необходимых для ее измерения. Например, с точки зрения операционного определения мы можем определить «образовательные возможности» как количество детей из семей с доходом менее трехсот тысяч рублей в год, посещающих колледжи и университеты согласно отчетам.

При этом показатели переменных – это непосредственно наблюдаемые характеристики, которые заменяют косвенно обозреваемые или не обозреваемые характеристики переменных. Например, количество

подростков, ежегодно поступающих в университет на бюджетные места, количество наркозависимых или количество выбросов серы в воздух. Количество делает восприятие переменной обозреваемой, в то время как качество мы не можем измерить (удовлетворенность работой, качеством жизни, экономическим прогрессом). Необходимо отметить, что для определения одной и той же переменной мы можем использовать разные показатели, что порождает проблему их интерпретации.

Поскольку отношение между переменными и показателями являются неоднозначными, желательно использовать множественные показатели (индексы) для характеристики действий и результатов политики. Индекс представляет собой комбинацию двух и более индикаторов, которые позволяют лучше измерить продукты реализации политики. Например, существует индекс загрязнения воздуха, индекс потребления электроэнергии, индекс качества жизни, индекс покупательной способности, индекс потребительских цен. Индекс позволяет отслеживать изменения в результатах реализации политики во времени, показывая, насколько изменились их показатели по отношению к базовому периоду.

Индексы могут быть простыми и составными. Простые индексы состоят только из одного показателя. Например, количество преступлений на сто тысяч жителей региона. Составные индексы включают несколько показателей. Например, индекс потребительский цен строится на основе показателей стоимости четырехсот товаров и услуг в США. Индекс конструируется либо посредством агрегирования, либо посредством усреднения. Агрегированные индексы строятся путем суммирования значений показателей (потребительских цен) за определенные период. Усреднение требует вычисление изменения средних значений показателей за определенный период времени. Например, индекс потребительской способности является агрегированным индексом. Он измеряет реальный заработок в соответствующие периоды, исходя из уровня индекса потребительских цен.

Использование индексов для характеристики изучаемых переменных часто связано с проблемой неточности информации, зафиксированной в этих показателях, сложностью формирования выборки для расчета индексов, которая бы однозначно отражала интересы всех слоев общества. Также индексы не всегда характеризуют качественные изменения в их значениях.

Таким образом, анализ социальных систем представляет собой статистический анализ показателей и индексов, характеризующих макросостояние социальных общностей на уровне государства, региона



или муниципалитета. Социальные показатели позволяют оценить воздействие осуществляемой политики на целевые группы. Однако этот метод часто называется случайно инновацией, так как мы можем оценить качество реализации политики только постфактум, что является не лучшем способом избежать дорогостоящих ошибок, заложенных в политической программе.

Таким образом, как видно из приведенного материала, анализ публичной политики является более комплексным, чем простое социальное исследование. Системный подход предполагает выделение таких методологических процедур как мониторинг, оценивание политики, прогнозирование и рекомендации. В свою очередь, мониторинг ближе всего находится к традиционному социальному исследованию, включая такие методологические подходы, как анализ социальных систем, социальных эксперимент, социальный аудит и синтез исследований и практики.

Проведение исследования предполагает выполнение определенной последовательности шагов, ведущих от общей идеи исследования (исследовательского вопроса) к эффективно сконструированным переменным, которые можно измерить на практике. Измерение означает аккуратное наблюдение реального мира с целью описания объектов и событий в терминах атрибутов (характеристик), которые составляют переменные. Необходимо отметить, что большинство переменных не существуют в реальности. Они создаются исследователями, поэтому они редко имеют однозначное, непротиворечивое значение. Например, переменная религиозность может быть измерена количеством посещения церкви в неделю или принадлежностью к соответствующей конфессии.

Существуют следующие критерии для измерения качества переменных: точность, надежность и валидность. Надежность означает, что при использовании данного метода по отношению к тому же объекту, мы получим похожие результаты исследования. Валидность означает степень, до которой эмпирическое измерение переменной адекватно отражает реальное значение изучаемого концепта. Переменные могут быть номинальными, т.е. обозначаемые словами, порядковыми с возможностью ранжирования на большее или меньшее, метрическими, т.е. предполагающими изменение переменных на равных интервалах.

### Разработка программы эмпирического исследования

Проведение научного исследования предполагает не только определение проблемы и постановку исследовательского вопроса, но и

разработку полноценной программы исследования. Однако до разработки данного документа необходимо прежде всего четко определиться с проблемой исследования, определить теоретическую рамку и создать набор переменных, которые будут измеряться на практике или оцениваться на примере других теоретических источников, например, как при использовании кейс-стади [8, с. 132–172].

Определение теоретической рамки тесно связано с выбором научной парадигмы. В социальном исследовании существуют несколько наиболее распространенных парадигм, которые определяют основные подходы к структурированию и интерпретации проблемы, а также правила работы ученый, работающих в данной парадигме. Мы поговорим в рамках нашего учебного пособия о позитивизме, конфликтной парадигме (марксизме), структурном функционализме и социальном конструктивизме.

Представителем раннего позитивизма является О. Конт, который первым заявил и обосновал, что общество нужно изучать научными методами. До Конта именно религиозные парадигмы объясняли различия между обществами. Конт первым пришел к выводу, что общественное устройство должно изучаться логически и рационально. Таким образом, позитивизм представляет собой научное направление в методологии, которое полагает в качестве основного источника знания об обществе те знания, которые получены на основе эмпирического исследования, а не путем философских рассуждений.

Марксизм («конфликтная парадигма» в западной научной традиции – *прим. автора*) рассматривает социальное поведение как конфликтный процесс, заключающийся в стремлении одних доминировать над другими. Формации сменяют друг друга, но конфликт между основными классами каждой формации сохраняется. Процесс общественного развития объясняется на основе действия трех законов диалектики: переход количества в качество, единства и борьбы противоположностей, а также отрицания отрицания.

Парадигма структурного функционализма берет свое начало в работах Г. Спенсера, Б. Малиновского и А. Редклиффа-Брауна, однако наибольшего расцвета она достигает в работах Т. Парсонса и его ученика Р. Мертона. Данная парадигма представляет макросоциальных подход к анализу социальных и политических проблем. Общество в нем представляет собой систему, состоящую из политической, экономической, социальной и культурной подсистем. Каждая подсистема выполняет определенную функцию, необходимую для слаженного функционирования данной системы. В свою очередь выявленные подсистемы также состоят из институтов, выполняющих определенную роль, позволяющую сохранять устойчивость системы. Для Т. Парсонса конфликт является чуждым явлением для общества и является признаком его болезни. Однако Р. Мертон, продолживший разработку данной парадигмы, выявляет возможность возникновения дисфункций социальной системы, таких как бюрократизм и коррупция.

Парадигма социального конструктивизма является микросоциальным подходом к анализу социальных явлений. Ведущими представителями данной парадигмы являются П. Бергер и Т. Лукман. Суть данной парадигмы заключается в том, что социальные отношения начинают формироваться на уровне взаимодействия индивидов. Эти самые первичные взаимодействия, например, при формировании новой семьи становятся привычными практиками, которые для второго и третьего поколения передаются как традиция и не подвергаются сомнению. Таким образом, устойчивые социальные институты становятся таковыми на основе первичного взаимодействия индивидов, т.е. взаимодействие является источником возникновения институтов.

Далее нам необходимо определиться с проблемой и провести концептуализацию и операционализацию выбранных для ее описания научных понятий. Концептуализация понятия связывает дизайн исследования или результаты исследования с теорией, выбранной для обоснования данного исследования. Операционализация выбранного понятия предполагает создание системы переменных и индикаторов для описания и измерения концепта.

Традиционная модель науки предполагает прохождение следующих этапов при проведении исследования [12, с. 72–117]. Так, исследователь сначала должен определиться с теоретической рамкой. При этом возможно использовать несколько теоретических рамок в зависимости от количества переменных, которые будут использованы при построении модели исследования. На основании теории исследователь формулирует гипотезу, которая устанавливает каузальные отношения между двумя переменными и более. Например, гипотезу можно сформулировать следующим образом, студенты, изучающие экономику, склонны уделять больше времени математике, чем студенты, изучающие политологию. Это утверждение нужно проверить. Для того, чтобы протестировать гипотезу, нам необходимо замерить все переменные, которые в нее входят. Заключительным этапом в традиционной модели науки служит наблюдение того, как ведут себя индикаторы перемен-

ных в реальности. Мы не просто тестируем гипотезу путем ее верификации, но и должны ее фальсифицировать [13, с. 78].

Исследовательский дизайн (программа исследования) позволяет нам упорядочить наши усилия в ходе проведения научного исследования. С точки зрения Добренькова В.И. и Кравченко А.И. [14, с. 151], именно разработка программы исследования, а также анализ данных и их интерпретация занимают наибольшее количество времени у исследователя. В российской традиции программа исследования состоит из двух частей. Первая часть – теоретико-методологическая, вторая часть методическая. Первая часть с необходимостью включает в себя проблему исследования, объект, предмет, цель и задачи исследования. Также в этой части описывается концептуализация и операционализация базовых понятий исследования. Цель исследования означает модель ожидаемого результата (способ решения проблемы), который можно достичь при проведении данного исследования. Проблема исследование – это отражение проблемной ситуации, которая, в свою очередь, содержит противоречие между тем, что есть и тем, что должно быть, или между знанием и незнанием. Задачи – это этапы на пути достижения цели, поэтому они не могут быть шире или одного объема с целью по содержанию. Объект исследования – это та часть объективной реальности, на которую направлен фокус внимания ученого. Например, объектом исследования могут выступать избирательная кампания 2020 г. в США. Предмет исследования – это одна из сторон (один из аспектов) объекта исследования, например, результаты избирательной кампании 2020 г. в США.

Методическая часть программы исследования включает описание выборки и правил ее формирования, совокупность методов сбора и анализа данных, обоснование логики их использования. Необходимо отметить, что в качественных и количественных исследованиях выборки формируются по-разному, о чем речь пойдет далее. В качестве методов сбора информации можно указать количественные методы сбора информации (опрос, наблюдение, социометрический опрос) и качественные методы (кейс-стади, интервью, фокус-группы) [15, с. 48–55]. К методам обработки информации можно отнести статистические методы анализа, метод конденсации смыслов [16, с. 224–225] метод обоснованной теории и т.д. [17, с. 29–38].

Необходимо также отметить, что программа исследования должна сопровождаться графиком выполнения работ, так как задержка в выполнении того или иного этапа проведения исследования чревата



несвоевременным выполнением работы. Сам процесс исследования предполагает реализацию следующих этапов: определение проблемы исследования, создание программы (дизайна) исследования, сбор данных (полевая стадия), анализ данных и написание отчета, научно-исследовательской работы или научной статьи.

### Основные ошибки при сборе данных

Познавая мир, мы ориентируемся, с одной стороны, на наш опыт, а с другой стороны, на знание о нем других, которые передаются из поколения в поколение. Знание, которое мы получаем от других, преобладает над тем, которое мы можем получить из собственного опыта, так как мы не можем самостоятельно охватить все многообразие окружающего мира самостоятельно. Следовательно, мы опираемся на информацию, которая содержится в традиции семьи и социума, а также на экспертные знания. Часто таким экспертом может выступать медийное лицо, известный журналист, актер, политик. Но на сколько мы можем доверять знанию такого лица, например, в области самолетостроения? Соответственно мы можем говорить о повседневном знании, которое не отвечает строгим критериям научности, и, собственно, о научном знании, которое обладает системностью, объективностью и точностью.

В философии существует отрасль под названием эпистемология, которая занимается теорией познания, устанавливает его принципы. В свою очередь методология представляет собой подраздел эпистемологии, направленный на поиск способов получения истинного, научного знания, разработку процедур научного исследования.

Какие ошибки в ходе проведения научного исследования уводят нас от получения научно-достоверного знания? Американский социолог E. Babbie выделяет четыре обычные ошибки при сборе данных – это неаккуратное наблюдение, чрезмерное обобщение, выборочное наблюдение и нелогичное обоснование [2, p. 6–8].

С неаккуратным наблюдением мы сталкиваемся тогда, когда пытаемся вспомнить о каком-то событии постфактум, когда мы не используем программу наблюдения с указанием объектов наблюдения, единиц наблюдения, временных промежутков проведения этого наблюдения, т.е. тогда, когда вместо научного наблюдения, мы ограничиваемся повседневным. Например, при анализе отношения к вашему кандидату, которого вы продвигали в рамках политической кампании, вы пытаетесь вспомнить, что говорили представители гражданского населения

на встрече с ним. Вы не сможете сделать это точно, так как на момент встречи у вас не было четко поставленной цели.

Следующая ошибка – это чрезмерное обобщение. Данная ошибка имеет место в наших исследованиях тогда, когда мы анализируем небольшую совокупность единиц наблюдения и затем пытаемся распространить полученные знания на всю генеральную совокупность или на весь объект наблюдения. Такая ошибка часто встречается тогда, когда мы проводим кейс-стади. Например, перед нами стоит задача проанализировать процесс приватизации на примере приватизации одного из вагоноремонтных заводов в Великобритании в эпоху М. Тэтчер. После того как мы проанализировали этот кейс, мы пытаемся сделать значимые выводы обо всем процессе приватизации и не только в Британии, но и в других странах. Это и есть ошибка чрезмерного обобщения.

Также мы можем столкнуться с ошибкой селективного (выборочного) наблюдения. Например, когда-то мы проанализировали большое массив данных и пришли к выводу, что в США наибольшее количество преступлений совершается людьми из низкого страта. Если последующее исследование покажет, что убийства совершаются также и представителями средних классов, мы будем просто игнорировать эту информацию для того, чтобы она не помешала нашей прекрасно разработанной и подтвержденной в рамках одного исследования гипотезы.

И, наконец, речь пойдет об ошибке нелогичного обоснования. Она встречается, например, в широко известной пословице «Исключение подтверждает правило». Однако данное утверждение само по себе нелогично, поскольку это лишь умозаключение не имеет под собой эмпирического обоснования. Еще одним примером нелогичного обоснования является «ошибка игрока». Так, например, игрок в покер предпочитает оставаться в игре до последнего в надежде, что за неудачей последует удача.

Таким образом, нам необходимо очень осторожно следить за возможностью появления таких ошибок в ходе проведения исследования, так как они могут стать причиной получения недостоверного знания, в результате которого наша статья или более значимая работа может быть отклонена эпистемическим сообществом.

### Блок самопроверки

Составить программу эмпирического исследования, определив научную парадигму и используя процедуры концептуализации и операционализации.



## ГЛАВА 2. Статистика с использованием языка программирования R

### Введение в программирование на языке R

Использование статистических методов стало проще для ученых и студентов с появлением статистического пакета SPSS [18]. Однако несмотря на дружелюбный интерфейс данного программного обеспечения, оно имеет один существенный минус, который ограничивает возможности его широкомасштабного использования. Это стоимость SPSS. Только крупные учебные заведения и исследовательские организации могут позволить себе установить лицензионную версию данного пакета.

Язык программирования R позволяет решить данный существенный недостаток SPSS, так его использование не предполагает значительных издержек для исследователя. Язык программирования R представляет собой целую вселенную, позволяющую обрабатывать количественные данные как в статистике, так и в сетевом анализе и парсинге. Однако в отличие от SPSS для того, чтобы провести статистический анализ данных, недостаточно просто нажать на «кнопку», и машина все сама посчитает. Нет, здесь необходимо владеть знаниями функций на языке R для осуществления статистических расчетов. Язык программирования необходимо учить также, как и иностранный язык посредством повседневной практики и освоения все новых и новых разделов статистики. Необходимо отметить, что вокруг языка программирования R сложились сообщества его разработчиков и пользователей, которые позволяют решать возникающие проблемы при использовании тех или иных функций для статистического анализа на языке R. Речь идет о сообществах www.habr.com, Community Stack Overflow, RStudio Community [19].

Для начала работы на языке программирования R необходимо установить две программы: R и RStudio. Для этого нужно воспользоваться ссылками www.cran.r-project.org и www.rstudio.com/products/rstudio/download, которые позволяет загрузить лицензионные программные продукты бесплатно. Статистические вычисления производятся в программе RStudio, которая согласовано работает с программой R. При открытии скрипта в RStudio перед нами появляется окно, состоящее из четырех разделов: 1) окно для скрипта; 2) окно консоли; 3) объекты, созданные в среде «История операций»; 4) файлы, рисунки, установочные пакеты и помощь [20].

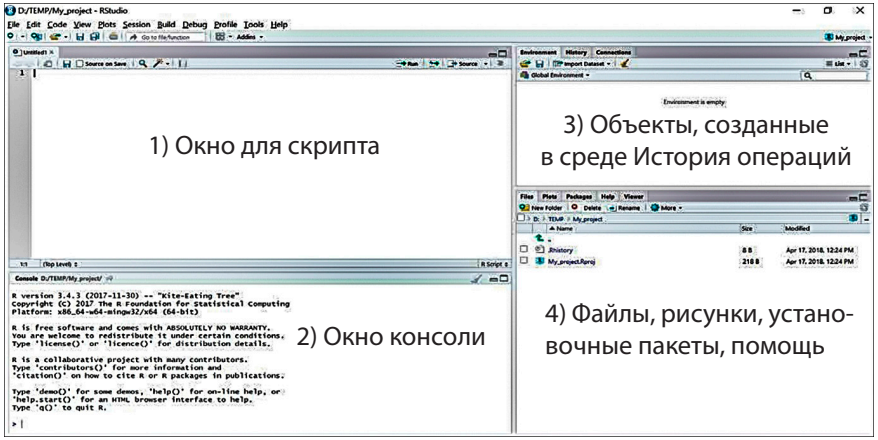


*Рис. 1. Окно для работы в RStudio*

Для написания функций, позволяющих проводить статистические расчеты, необходимо использовать окно для скрипта. Результаты вычислений отражаются в окне консоли [21, с. 35–43]. Документы, созданные в RStudio, сохраняются с расширением .RData. Однако в RStudio можно также работать и с документами (базами данных), имеющих расширение scv, xls и sav. Для того, чтобы открыть документ в формате .RData достаточно использовать команду File -> Open File. Для работы с файлами в другом формате нужно прописать директорию. Важно учесть, что на современных компьютерах есть только системный диск C, куда сохранять файлы с базами данных нежелательно. Для этого лучше использовать флеш карту, которая будем идентифицироваться компьютером, как диск d. Однако выбор диска для написания директории – это еще полпроблемы. Основной вопрос: где брать базы данных для обучения или проведения самостоятельных научных исследований? Некоторые базы есть в специализированных пакетах R, например, в пакете MASS [22]. Также данные можно брать на портале открытых данных, например, на сайте портала открытых данных правительства г. Санкт-Петербург: http://data.gov.spb.ru/. На порталах открытых данных базы данных в основном доступны в csv и xls формате.

Доступ к файлам, указанным в пособии обеспечивается на информационном портале Центра технологий электронного правительства Института дизайна и урбанистики Университета ИТМО (регистрация обучающихся – https://news.egov.itmo.ru/for-students).

Для открытия базы данных в формате xls, необходимо активировать пакет readxl. Для активации пакета мы используем функцию library() и



укажем в скобках данных пакет. Таким образом, активация пакета выглядит так: library (readxl). После чего нам необходимо запустить данную функцию используя комбинацию клавиш CTRL+ENTER. Сохраним на диске d с портала открытых данных правительства г. Санкт-Петербург файл Post offices_SPb.xls. Вложим содержимое директории в переменную post при помощи знака присвоения «<–». Получим переменную post<–read_xls('e:/Post offices_SPb.xls',sheet=1). Проверим результат при помощи функции чтения верхних строк базы данных head(), т.е. head(post, n=10).

```
library(readxl)
post<-read_xls('e:/Post offices_SPb.xls',sheet=1)
head(post, n=10)
## # A tibble: 10 x 8
## name district address nearest_subway_~ mode closed_for_lunch
telephone
## <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 Санк~ Адмирал~ "г. Са~ ст.м.Невский пр~ "кру~ <NA> 088; 314~
## 2 Санк~ Адмирал~ "г. Са~ ст.м.Технологич~ "пон~ 13.00–14.00 251–00–14
## 3 Санк~ Адмирал~ "г. Са~ <NA> "пон~ 13.00–14.00 714–59–14
## 4 Санк~ Адмирал~ "г. Са~ ст.м.Технологич~ "пон~ 13.00–14.00 316–14–20
## 5 Санк~ Адмирал~ "г. Са~ ст.м. Балтийска~ "пон~ 13.00–14.00 251–21–7~
## 6 Санк~ Адмирал~ "г. Са~ ст.м.Сенная пло~ "пон~ 14.00–15.00 310–72–37
## 7 Санк~ Адмирал~ "г. Са~ ст.м.Сенная пло~ "пон~ 14.00–15.00 314–47–80
## 8 Санк~ Адмирал~ "г. Са~ ст.м.Невский пр~ "пон~ 13.00–14.00 314–93–27
## 9 Санк~ Адмирал~ "г. Са~ ст.м. Балтийская "пон~ 14.00–15.00 251–18–03
## 10 Санк~ Адмирал~ "г. Са~ ст.м.Невский пр~ "пон~ 12.00–13.00 570–34–71
## # ... with 1 more variable: class <chr>
```

Теперь откроем файл в формате csv. Это можно сделать двумя способами. Во-первых, используя функцию read.table(). И тогда данная операция будет выглядеть следующим образом: read.table('e:/Districts_SPb2.csv', sep=",", header=TRUE).

```
read.table('e:/Districts_SPb2.csv',sep=",",header=TRUE)
##        name                okato code_     district
## 1   Центральный            40298                31
## 2   Адмиралтейский         40262                32
## 3   Выборгский             40265                36
## 4   Красносельский         40279                16
## 5   Василеостровский       40263                 6
## 6   Фрунзенский            40296                13
```

```
## 7   Московский          40284          14
## 8   Петродворцовый      40290          21
## 9   Калининский         40273          10
## 10  Пушкинский          40294          35
## 11  Кировский           40276          15
## 12  Курортный           40281          38
## 13  Красногвардейский   40278          11
## 14  Колпинский          40277          37
## 15  Невский             40285          12
## 16  Кронштадтский       40280          50
## 17  Приморский          40270          34
## 18  Петроградский       40288           7
```

Во-вторых, мы можем открыть данный тип файла, используя функцию read.csv. Тогда последовательность операций выглядит следующим образом: districts<-read.csv('e:/Districts_SPb2.csv', header=TRUE). Проверим данные при помощи функции head(districts, n=3).

```
districts<-read.csv('e:/Districts_SPb2.csv',header=TRUE)
head(districts, n=3)
##          name    okato code_district
## 1 Центральный 40298           31
## 2 Адмиралтейский 40262        32
## 3 Выборгский  40265           36
```

Формат csv расшифровывается как «comma separated values», т.е., другими словами, это формат, в котором данные разделены запятыми. Этот тип данных относится к «опрятным данным». «Опрятные данные» означают, что в таких данных нет пропусков значений или объединения ячеек. Обычно они также пишутся в Excel, но их значения разделены запятыми. Такие данные наиболее удобны для чтения при использовании многих языков программирования.

RStudio также может читать SPSS файлы с расширением .sav [23]. Для этого нужно активировать пакет R: foreign, т.е. мы должны указать library(foreign) и нажать комбинацию клавиш CTRL-ENTER. Создаем переменную birth_rate для того, чтобы прочитать данные об индексе рождаемости в разных странах, хранящиеся в файле SPSS. Для того, чтобы прочитать данный файл в R, нам потребуется функция read.spss(). Запишем: birth_rate<–read.spss('e:/birth_rate.sav', to.data.frame=TRUE). Получаем следующий результат, который можно посмотреть при помощи функции head() или tail(), позволяющих просматривать первые и последние строки.



```
library(foreign)
birth_rate<-read.spss('d:/birth_rate.sav',to.data.frame=TRUE)
head(birth_rate)
##       Country
## 1       UAE
## 2       DRC (Congo)
## 3       Czech Republic
## 4       Germany
## 5       Finland
## 6       Egypt
##   Maternalmortalityrateper100000livebirths
## 1       6
## 2       693
## 3       4
## 4       6
## 5       3
## 6       33
##   Numberofdeathsofnewbornsbeforetheageof28days
## 1       415
## 2       97832
## 3       175
## 4       1604
## 5       104
## 6       31796
##   Adolescentbirthratenumberofbirthsper1000womenaged1519years
## 1       28.0758
## 2       126.2696
## 3       10.4158
## 4       7.1462
## 5       7.0306
## 6       51.8828
```

Для извлечения данных и расчетов из программной среды RStudio мы будем использовать написание презентации на языке RMarkdown с сохранением ее в формате Word. Для этого выполним команду File>New File>RMarkdown, затем выберем из списка документ в формате Word. Написание презентаций на языке RMarkdown предполагает использование чанков, в которых прописывается и активируется код. Для вставки чанка нужно набрать комбинацию клавиш CTRL+ALT+I. Первый чанк должен остаться неизменным. Это настроечный код, который запускает функцию knitr. По умолчанию в нем прописан следующий текст:

```{r setup, include=FALSE}
knitr:: opts_chunk$set(echo=TRUE)
```

В остальных чанках прописываются коды, которые позволяют делать статистические расчеты и визуализировать данные. Для того, что содержимое в чанках отображалось и запускался код с соответствующей функцией необходимо указать после r «echo=TRUE, eval=TRUE». Между чанками прописывается текст, который отражает основное содержание исследования. После того как написание презентации закончено на панели выбираем «клубок» Kint. Теперь коды в наших чанках активируются, и мы получим документ в формате Word.

### Описательная статистика с использованием R

Обратимся к описательной статистике и рассчитаем медиану и квантили, а также среднее и стандартное отклонение для данных из датасета energy.xls. Для начала активируем пакет readxl при помощи функции library().

```
library(readxl)
energy_data<-read_xls('e:/energy.xls',sheet=1)
```

Далее рассчитаем медиану и квантили для переменной, содержащейся в данном датасете, т.е. для AlternativeNuclearEnergy. Построим график для медианы и квантилей в виде боксплота (ящика с усами).

```
energy<-energy_data$AlternativeNuclearEnergy
median(energy)
## [1] 15.11033
quantile(energy)
##         0%        25%        50%        75%       100%
##  0.9332368  5.9623467 15.1103274 27.8468201 43.2411504
library(ggplot2)
data_energy<-data.frame(y=energy)
ggplot(data=data_energy)+geom_boxplot(aes(x='Медиана \nи квантили',y=y))+labs(x='мощности',y='использование \nэнергии')
```

Далее рассчитаем медианной значение и значение квантилей для выбросов $CO_2$ с использованием соответствующих функций языка R.

```
emission<-energy_data$CO2Emissions
median(emission)
## [1] 7.088592
quantile(emission)
##        0%       25%       50%       75%      100%
##  3.896324  5.580589  7.088592  9.283754 15.388950
```



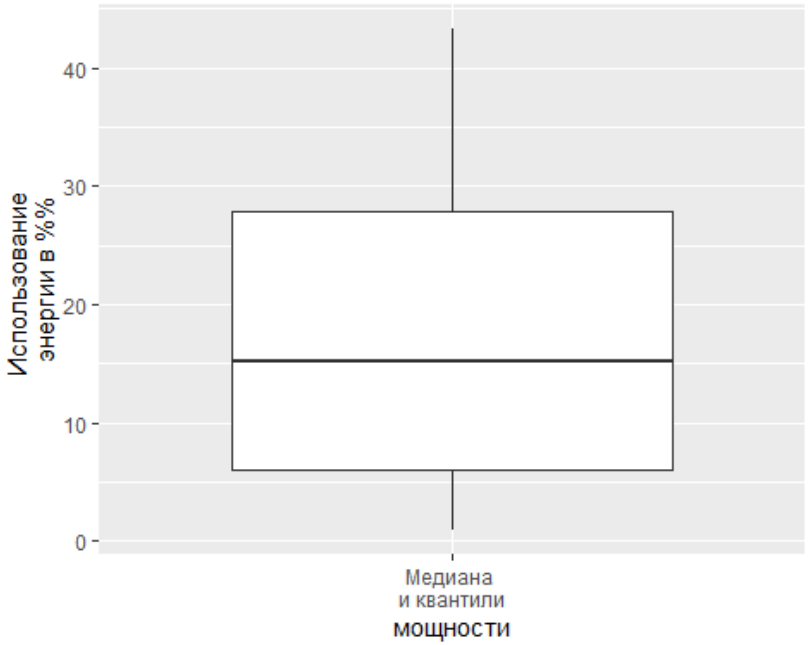


*Рис. 2. Визуализация медианного значения использования источников альтернативной энергии*

Рассчитаем среднее и стандартное отклонение для обоих переменных.

```
mean(energy)
## [1] 17.80113
sd(energy)
## [1] 13.44134
mean(emission)
## [1] 7.971809
sd(emission)
## [1] 3.450858
```

Далее отразим на графике переменную использования альтернативных источников энергии.

```
library(ggplot2)
data_energy<-data.frame(y=energy)
ggplot(data=data_energy)+geom_boxplot(aes(x='Медиана \nи квантили',y=y))+stat_summary(geom='pointrange',fun.data=mean_sdl,fun.args=list(mult=1),aes(x='Среднее\nи ст.отклонение',y=y))+labs(x='Использование \nэнергии',y='Процент \nиспользуемой энергии')
```

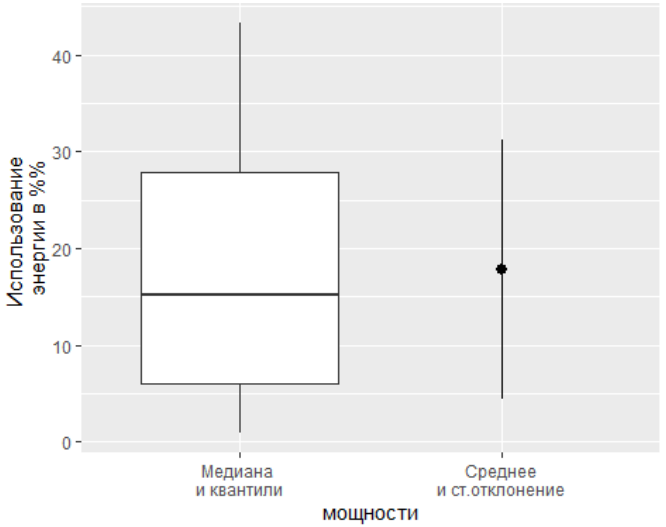


*Рис. 3. Сопоставление среднего и медианного значения количества используемой альтернативной энергии*

Теперь построим среднее и медианное значение для количества выбросов углеводорода, выраженного в процентах.

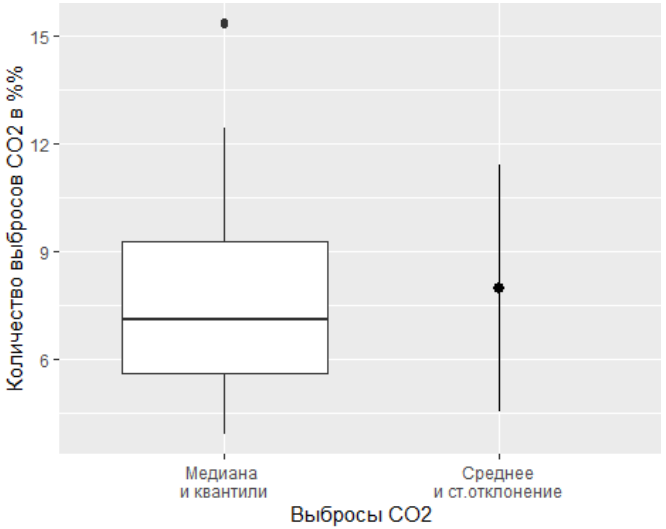


*Рис. 4. Сопоставление среднего и медианного значения количества выбросов $CO_2$*



Используя данный пример, также рассчитаем корреляцию Пирсона и ранговую корреляцию Спирмена для переменных «Альтернативные источники энергии» и «Выбросы $CO_2$».

```
cor(energy,emission,method="pearson")
## [1] -0.4402592
cor(energy,emission,method="spearman")
## [1] -0.477193
```

### Построение модели множественной регрессии на языке R

В качестве примера использования языка программирования R в статистическом анализе построим множественную линейную регрессионную модель с дискретными и непрерывными предикторами на основании данных о весе новорожденных у курящих и некурящих рожениц [24, С. 49–56]. Данный датасет расположен в пакете MASS, который мы должны активировать при помощи функции library(). Конструируемая модель носит ознакомительный характер с кодировками основных статистических функций в языке программирования R. С более сложными регрессионными моделями можно познакомиться в учебнике Дж. Фаравей [25].

```
library(MASS)
newborn<-birthwt # Переименуем датасет
#Проверим, из каких переменных состоит датасет
str(newborn)
## 'data.frame':    189 obs. of  10 variables:
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

Преобразуем значения и тип переменных. Напоминаем, что переменная Smoke – дискретная, т.е. она должна стать классифицирующим фактором.

```
newborn$smoke[newborn$smoke==1]<-'Smoke'
newborn$smoke[newborn$smoke==0]<-'Dont smoke'
newborn$smoke<-factor(newborn$smoke)
```

Используя точечные диаграммы Кливленда, проверим данные на наличие выбросов. Сначала проверим непрерывную переменную age (возраст матери).

```
library(ggplot2)
theme_set(theme_bw())
gg_point<-ggplot(newborn,aes(y=1:nrow(newborn)))+geom_point()
gg_point+aes(x=age)
```

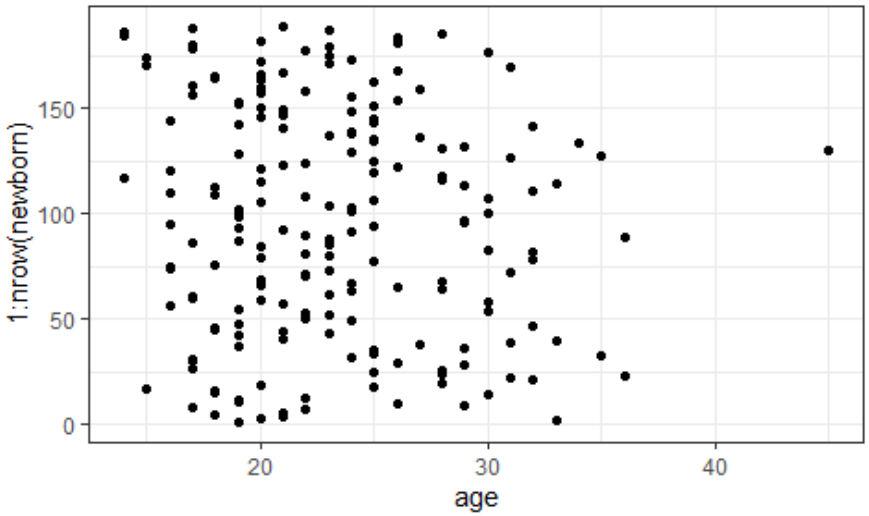


*Рис. 5. Диаграмма Кливленда для независимой переменной age*

Данные по переменной age более-менее однородны.

Теперь исследуем зависимую переменную bwt (вес ребенка при рождении).

```
gg_point+aes(x=bwt)
```

Строим регрессионную модель в дискретными и непрерывными предикторами:

bwt=b0 + b1*age+ b2*Smoke + b3*age*Smoke + e

Строим модель с взаимодействием предикторов

```
Mod<-lm(bwt~age*smoke, data=newborn)
```

Далее проверим полученную модель на условия применимости модели линейной регрессии, среди которых можно выделить проверку на наличие линейной связи, нормальность распределения, мультиколлинеарность, на отсутствие гетероскедастичности остатков, на независимость наблюдений.



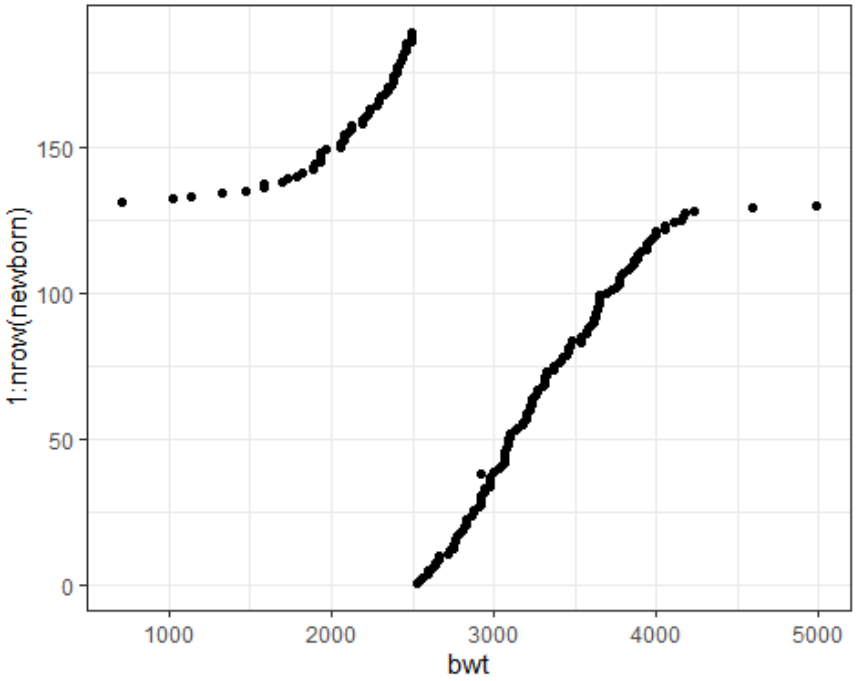


*Рис. 6. Диаграмма Кливленда для зависимой переменной bwt*

Проверим данные на нормальность распределения, используя квантильный график остатков. Для этого активируем пакет car.

```
library(car)
qqPlot(Mod,id=FALSE)
```

Данные подчинены нормальному распределению, так как значимых отклонений от линии на квантильном графике остатков нет.

Для проверки данных на коллинеарность построим регрессионную модель без взаимодействия предикторов, в которой предикторы будут соединены "+". Для того, чтобы проверить наличие мультиколлинеарности у предикторов будем использовать фактор инфляции дисперсии (Variance inflation factor, vif). Функция vif() расположена в пакете car.

```
Mod_vif<-lm(bwt~age+smoke,data=newborn)
library(car)
## Loading required package: carData
vif(Mod_vif)
##    age   smoke
## 1.00197 1.00197
```

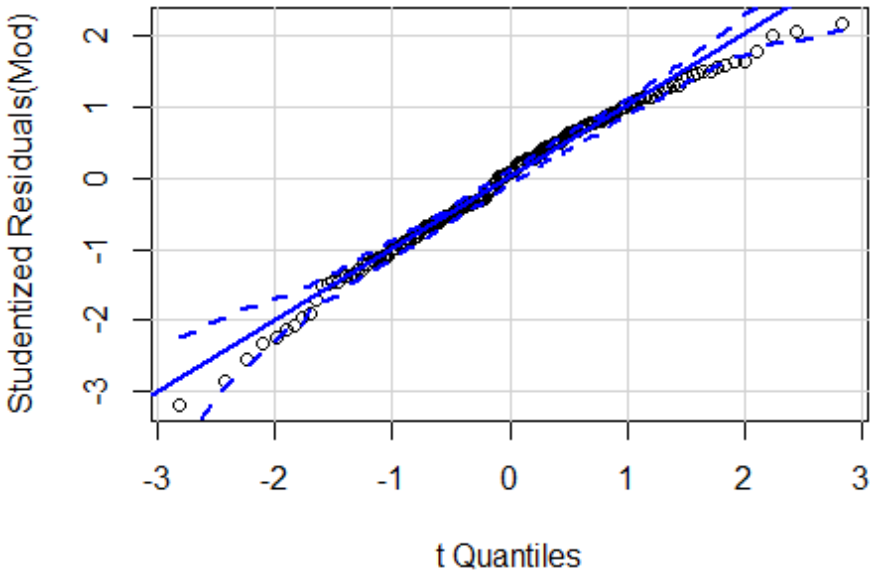


*Рис. 7. Квантильный график остатков*

Значения vif меньше двух, следовательно, коллинеарности предикторов нет.

Для анализа остатков рассчитаем дополнительный набор данных с использованием функции fortify() из пакета ggplot2. Пакет ggplot2 также используется для построения графиков в среде RStudio.

```
Mod_diag<-fortify(Mod)
head(Mod_diag,n=10)
## bwt age smoke .hat .sigma .cooksd .fitted .resid
## 85 2523 19 0 0.014443758 710.5267 1.242008e-03 2932.954 -409.954161
## 86 2551 33 0 0.035590176 708.8245 1.128088e-02 3321.193 -770.193456
## 87 2557 20 Smoker 0.018179969 710.8940 6.854098e-04 2827.422
-270.422121
## 88 2594 21 Smoker 0.015549613 710.9999 3.671583e-04 2808.582
-214.581585
## 89 2600 18 Smoker 0.026666860 710.9027 9.831399e-04 2865.103
-265.103194
## 91 2622 21 0 0.010422673 710.6600 7.101780e-04 2988.417 -366.416917
## 92 2637 22 0 0.009292381 710.6240 6.763800e-04 3016.148 -379.148295
## 93 2637 17 0 0.020812178 710.9529 6.239089e-04 2877.491 -240.491404
## 94 2663 29 Smoker 0.033220950 711.1785 4.671731e-07 2657.857 5.142708
```



```
## 95 2665 26 Smoker 0.018528745 711.1691 2.330828e-05 2714.379
-49.378902
## .stdresid
## 85 -0.582227688
## 86 -1.105776291
## 87 -0.384790670
## 88 -0.304925535
## 89 -0.378863239
## 91 -0.519336577
## 92 -0.537074606
## 93 -0.342661111
## 94 0.007374397
## 95 -0.070275007
```

Проверим данные на наличие влиятельных наблюдений при помощи графика расстояний Кука. У влиятельных наблюдений расстояние Кука будет больше единицы.

```
ggplot(Mod_diag,aes(x=1:nrow(Mod_diag),y=.cooksd))+geom_
bar(stat='identity')
```

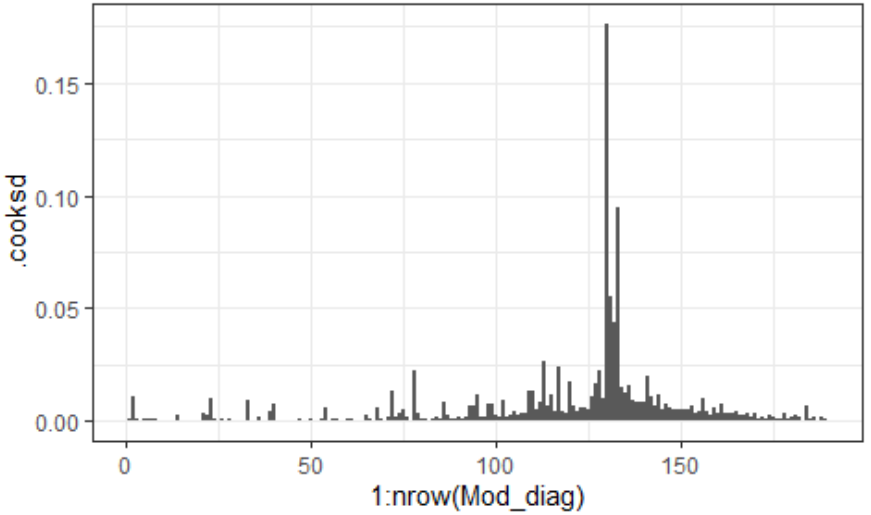


*Рис. 8. График расстояний Кука*

Расстояния Кука у всех наблюдений меньше единицы, поэтому в данном датасете влиятельных наблюдений нет.

Построим график остатков в зависимости от предсказанных значений.

```
gg_residue<-ggplot(data=Mod_diag,aes(x=.fitted,y=.
stdresid))+geom_point()+geom_hline(yintercept=0)+geom_smooth()
gg_residue
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

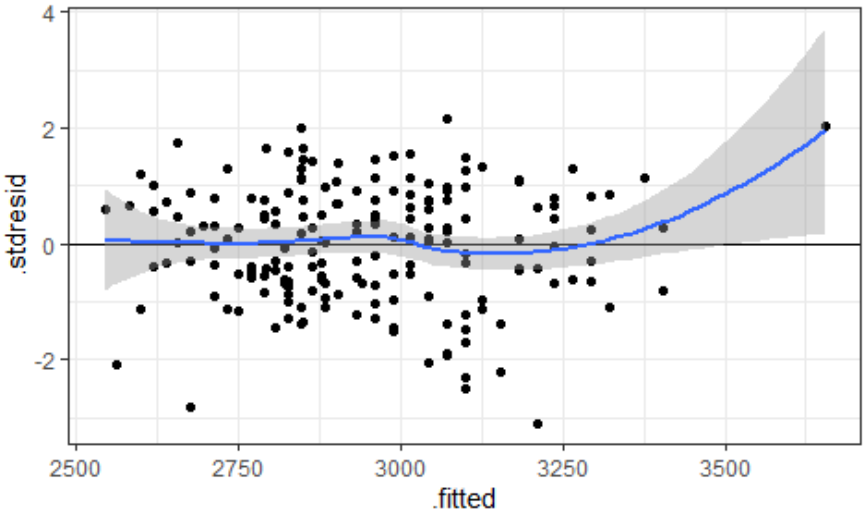


*Рис. 9. График остатков в зависимости от предсказанных значений зависимой переменной*

Предсказанные значения распределены равномерно, поэтому гетероскедастичность отсутствует.

Построим графики остатков в зависимости от предикторов, используемых в данной модели.

```
gg_residue+aes(x=age)
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'

ggplot(Mod_diag,aes(x=smoke,y=.stdresid))+geom_
boxplot()+geom_hline(yintercept=0)
```

В обоих случаях гетерогенность дисперсии не выявлена.

При помощи функции summary() посчитаем коэффициенты и их значимость, а также коэффициент детерминации.

```
summary(Mod)
##
## Call:
## lm(formula = bwt ~ age * smoke, data = newborn)
```



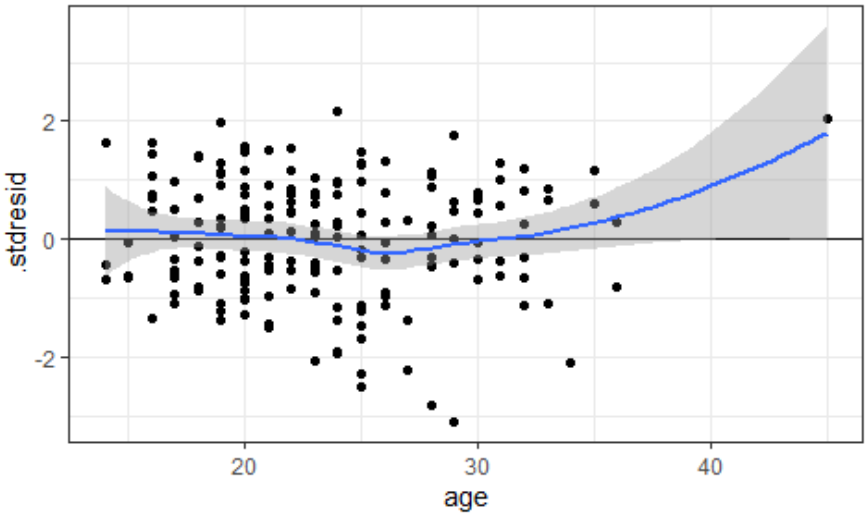


*Рис. 10. График остатков для непрерывного предиктора age*

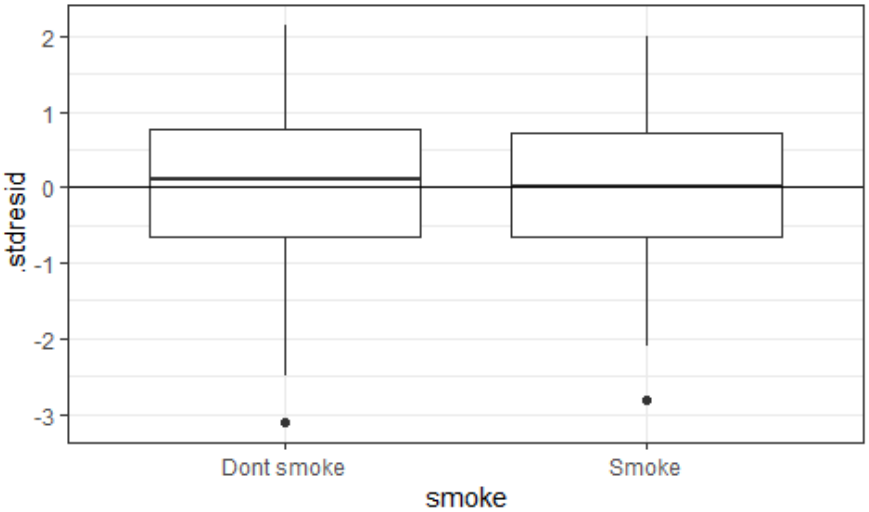


*Рис. 11. График остатков для дискретного предиктора smoke*

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2189.27  -458.46    51.46   527.26  1521.39
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2406.06     292.19   8.235 3.18e-14 ***
## age              27.73      12.15   2.283   0.0236 *
## smokeSmoker      798.17     484.34   1.648   0.1011
## age:smokeSmoker  -46.57      20.45  -2.278   0.0239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 709.3 on 185 degrees of freedom
## Multiple R-squared:  0.06909,    Adjusted R-squared:  0.054
## F-statistic: 4.577 on 3 and 185 DF,  p-value: 0.004068
```

Попробуем упростить регрессионную модель, избавившись от взаимодействия предикторов. Проверим при помощи функции drop1(), можно ли избавиться от взаимодействия.

```
drop1(Mod,test='F')
## Single term deletions
##
## Model:
## bwt ~ age * smoke
##         Df Sum of Sq      RSS    AIC F value  Pr(>F)
## <none>               93062605 2485.2
## age:smoke  1   2609683 95672288 2488.5  5.1878 0.02389 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Поскольку уровень значимости меньше 0,05, то при удалении взаимодействия из модели, модель значимо изменится. Следовательно, взаимодействие из данной модели удалять нельзя.

В результате у нас получается две регрессионные модели:

№ 1. Для базового уровня:

bwt = 2406.06 + 27.73*age (т.е. для некурящих рожениц)

№ 2. Для объектов, не относящихся к базовому уровню:

bwt = 2406.06 + 27.73*age + 798.17 + (–46.57)*age (т.е. для курящих рожениц)

Далее визуализируем полученную регрессионную модель. Для классификации данных по группам используем функцию group_by(). Для этого активируем пакет dplyr.

```
library(dplyr)
```



```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:car':
##
##     recode
## The following object is masked from 'package:MASS':
##
##     select
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
new_data<-newborn%>%group_by(smoke)%>%do(data.frame(age=
seq(min(.$age),max(.$age),length.out=100)))
Predictions<-predict(Mod,newdata=new_data,se.fit=TRUE) #Рассчи-
таем предсказанные значения
new_data$fit<-Predictions$fit
new_data$se<-Predictions$se.fit #Рассчитаем стандартные ошибки
t_crit<-qt(0.975,df=nrow(newborn)-length(coef(Mod))) #Рассчитаем
критические значения для 95% доверительного интервала
#Строим доверительный интервал
new_data$lwr<-new_data$fit-t_crit*new_data$se
new_data$upr<-new_data$fit+t_crit*new_data$se
```

Визуализируем данные на основе рассчитанных значений, а также добавляем данные наблюдений в виде точек при помощи геома (geom_point).

```
Plot_smoke<-ggplot(new_data,aes(x=age,y=fit))+geom_ribbon
(alpha=0.2,aes(ymin=lwr,ymax=upr,group=smoke))+geom_line
(aes(colour=smoke))
Plot_final<-Plot_smoke+geom_point(data=newborn,aes(x=age,y=b
wt,colour=smoke))+scale_colour_discrete('Пристрастие к курению',
labels=c('Не курит','Курит'))+labs(x='Возраст роженицы', y='Вес ребенка')
Plot_final
```

После визуализации модели необходимо интерпретировать полученные результаты. График на рис. 12 позволяет сделать вывод, что у некурящих рожениц с возрастом есть тенденция рожать детей с боль-

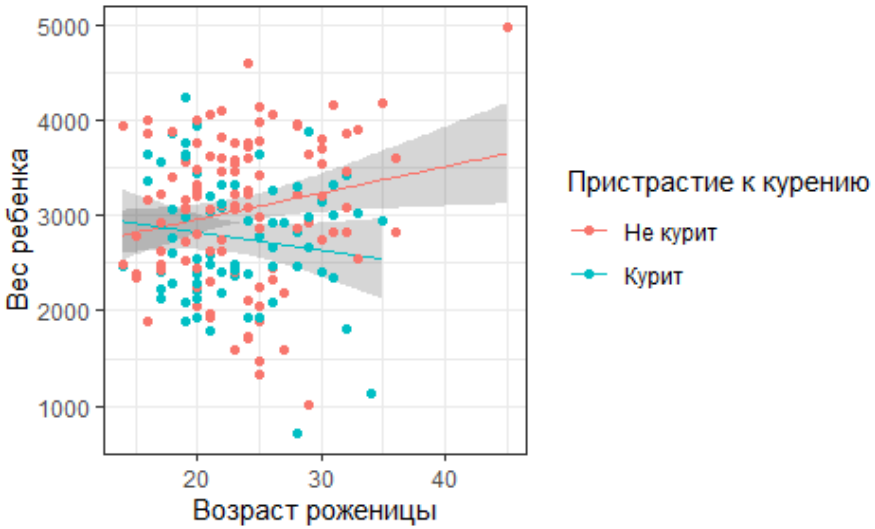


*Рис. 12. График предсказанных значений зависимой переменной с указанием исходных значений*

шим весом, у курящих рожениц тенденция обратная. Расчеты показали, что существует значимое взаимодействие между предикторами возраст и пристрастие к курению, поскольку уровень значимости был меньше 0,05. Данную тенденцию можно увидеть и на данном графике предсказанных значений.

Таким образом, R позволяет делать статистический анализ как в дескриптивной статистике, так и в более сложной, требующей построение регрессионных моделей как с линейной связью предикторов и отклика, так и с нелинейной, позволяет делать анализ временных рядов и работать с непараметрической статистикой, делать сложные визуализации для нескольких предикторов. Однако для того, чтобы выполнять все эти расчеты, нужно познакомиться с разными функциями и пакетами, созданными для программирования на языке R.

### Блок самопроверки

Построить модель линейной регрессии на основании датасета whiteside в пакете MASS.



## ГЛАВА 3. Сетевой анализ с использованием Pajek

### Построение сети с использованием программного обеспечения Pajek

В последнее десятилетие большую популярность в социальных исследованиях набирает методология сетевого анализа. В политической науке в рамках неоинституционализма зародился и бурно развивается сетевой подход [26, с. 14–32]. Конечно, этот подход имеет как ряд преимуществ, так и ряд недостатков. К сильным сторонам сетевого подхода относится возможность визуализации связей между разнообразными акторами, вычисление их роли в данном сетевом взаимодействии. В отличие от структурно-функционального анализа сетевой подход указывает на возможность построения горизонтальных связей между политическими акторами, например, [27, с. 68–86]. К слабым сторонам данного подхода относится сложность в сборе данных об акторах, по крайне мере, в России, отсутствие однозначной взаимосвязи между активностью в политической сети и изменением политической повестки дня. Кроме того, очень сложно доказать мобилизационный эффект социальных сетей, связь взаимодействия акторов онлайн и оффлайн [28], [29].

В социологии сетевая методология позволяет визуализировать отношения между акторами, которыми являются индивиды, а в политологии мы можем выстроить эти отношения между государственными учреждениями и организациями гражданского общества. Такая визуализация позволяет определить акторов (узлы), которые находятся ближе к центру или периферии сети, в связи с этим определить их роли, рассчитать центральности произвести ранжирование. Для автоматизации процесса построения сетей были разработаны специальные компьютерные программы: Pajek, UCINET, Gephi.

В нашей работе мы остановимся на программном обеспечении Pajek (от слова «паук»), разработанный словенскими социологами [30]. Pajek является бесплатным программным обеспечением, и его можно скачать по ссылке http://mrvar.fdv.uni-lj.si/pajek/. Для того, чтобы корректно скачать данное программное обеспечение нужно выбрать зеленую иконку 32 или 64 bit в зависимости от возможностей вашего компьютера. Данный вариант Pajek рассчитан на 1000 узлов (акторов) и позволяет визуализировать информацию при помощи специального инструмента (карандаш в правом углу командной строки Сеть/Network). Другие варианты данной программы – Pajek XXL, Pajek 3XL – работают с количеством узлов значительно больше тысячи узлов и не позволяют

визуализировать данную информацию, мы можем только производить расчеты по этим данным.

В социологии принято классифицировать социологические (социальные) исследования по трем типам: разведывательное, описательное и объяснительное [31, с. 43]. Эти типы исследования можно представить как движение по лестнице вверх: от самого простого вида исследования к самому сложному. Всегда желательно проводить исследование на уровне объяснительного, т.е. с использованием серьезной аналитики, но, к сожалению, это не всегда возможно. И проблема не в том, что у исследователей не хватает аналитических способностей и навыков проведения таких исследований, камнем преткновения служат возможности соответствующего дизайна исследования. Так сетевой анализ по определению является разведывательным (эксплораторным). Он не предполагает выдвижение конкретной гипотезы в момент подготовки программы исследования. Сначала мы строим сеть (визуализируем) информацию, а потом с ее использованием пытаемся выявить закономерности и сделать значимые выводы.

Сетевой анализ в Pajek позволяет не только визуализировать информацию, но и делать несложные математические расчеты. Например, мы можем рассчитать плотность сети, коэффициенты корреляции Пирсона и Спирмена, показатели центральности. На основании расчетов мы можем также выделить наиболее связанные участки сети (ядра, кластеры). Таким образом, Pajek дает возможность комплексного анализа построенной сети.

Для построения сети нам необходимо четко сформулировать проблему исследования, и, исходя из нее, идентифицировать акторов и наличие взаимодействий между ними. Все это можно сделать, используя ивент-анализ, контент-анализ и интервью по типу «снежного кома». К сожалению, использование интервью не всегда возможно, его возможности ограничены статусом опрашиваемого респондента. Если мы имеем дело с проблемой местного сообщества, то скорее всего встретиться и переговорить с депутатом муниципального совета для нас не составит труда. Гораздо сложнее, когда в качества респондентов выступает представители государственных органов федерального уровня [32, p. 47–50]. Здесь мы ощущаем эффект «стеклянного потолка», и поэтому лучше использовать контент-анализ (анализ интернет-медиа и традиционных СМИ) или ивент-анализ (анализ участников знаковых политических мероприятий). Когда нужно остановиться в сборе данных? Тогда, когда информация начнет повторяться.



В основе визуализации в сетевом анализе лежит теория графов, где граф – это набор вершин и линий между парами вершин. Сеть – это такой же вариант графов, но уже с подписанными вершинами. Сеть в сетевом анализе бывает двух видов: направленная и ненаправленная. Направленная сеть предполагает, что вершины связаны между собой направленными дугами, в то время как в ненаправленной сети вершины связаны двухсторонними дугами. В зависимости от типа сети мы можем использовать к ней те или иные аналитические процедуры, заложенные в программное обеспечение.

Обсуждение использования Pajek начнем с построения сети. Сеть в Pajek можно построить двумя способами: 1) при помощи написания матрицы сопряженности в программе Блокнот/Notepad; 2) используя функции диалогового окна Сеть/Network в Pajek.

Возьмем некоторую гипотетическую страну Хэппилэнд, в которой есть три уровня власти. Федеральный уровень власти представлен президентом, премьер-министром и тремя министрами. Региональный уровень представлен губернатором, вице-губернатором и тремя руководителями комитетов. Муниципальный уровень представлен главой муниципального образования, заместителем главы муниципального образования и тремя специалистам. Всего у нас в данной сети получается 15 акторов.

При написании мини-программы для построения сети с использованием программы Блокнот, мы должны отдельно прописать вершины и матрицу сопряженности.

*Vertices 15
1 “President”
2 “Prime-minister”
3 “Minister 1”
4 “Minister 2”
5 “Minister 3”
6 “Governor”
7 “Vice-governor”
8 “Committee head 1”
9 “Committee head 2”
10 “Committee head 3”
11 “Municipal head”
12 “Minisipal deputy”
13 “Specialist 1”
14 «Specialist 2»
15 «SPecialist 3»

*Таблица 2*

**Матрица сопряженности в программе Блокнот/Notepad*Matrix**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 4 | 2 | 3 | 3 | 2 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 3 | 3 | 3 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 2 | 0 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 2 | 2 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 2 | 2 | 1 | 1 | 2 | 0 | 2 | 2 | 2 | 2 | 2 | 0 | 1 | 1 | 1 |
| 1 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 1 | 2 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 0 | 1 | 2 | 2 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 3 | 3 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 3 | 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 1 | 1 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 1 | 0 |

При написании программы мы начинаем со звездочки (*), далее с большой буквы мы указываем наименование Vertices и количество вершин «15». Далее на каждой строке через пробел мы прописываем номер и название вершины в кавычках. По умолчанию все вершины в Pajek круглые, но мы можем их сделать, например, квадратными, указав после названия вершины square. После перечисления вершин без лишней строки разделителя пишем звездочку (*) и Matrix (матрицу сопряженности). В матрице сопряженности по горизонтали мы указываем тех, кто выбирает; по вертикали тех, кого выбирают. Например, вершина «President» (первая строка) выбирает «Prime-minister» довольно часто (4 вертикаль). Цифры показывают значение линий или силу связи. Так «0» означает, что связи между этими акторами (вершинами) нет, «1» показывает, что связь есть, но она не очень сильная. Чем больше используемое число, тем больше связь. Между цифрами в матрице сопряженности должны быть пробелы, на диагонали должны быть нули, поскольку мы не можем выбирать сами себя. Данные в файле пишутся на английском, и сам файл мы сохраняем латинскими буквами. С русскими обозначениями Pajek не работает. После того, как программа написана, мы открываем ее через Pajek при помощи команды File>Network>Read. Если все правильно, то во вспомогательном окне будет указано, что



«прочитано такое-то количество линий». Далее мы можем визуализировать построенную сеть при помощи «карандаша» в правом углу диалогового окна Сеть/Network.

При визуализации открывается окно для рисования. Здесь для нас важными функциями являются подразделы меню окна рисования Layout, Options, Export. При помощи команды Kamada-Kawai в подразделе Layout мы можем видоизменять (энергенизировать) нашу сеть до тех пор, пока мы не сочтем визуализацию наиболее удачной. При помощи подраздела Options мы можем отображать вершины, обозначенные либо именами, либо цифрами; показывать значение связей на линиях; менять толщину линий и цвет фона. При помощи подраздела Export мы можем выкачать полученный рисунок сети в формате eps, jpg, что является важным условием при написании презентаций.

Давайте сохраним полученную сеть в файл HappyLand.net. Теперь мы можем создать проектный файл, добавив к нему файл с классификацией. Напомним, что у нас есть три уровня власти в данном государстве. Давайте сгруппируем акторов в три группы соответственно. Для этого используем файл Partition.clu.

Теперь можно сделать совместную визуализацию построенной сети и классификации (partition). В разделе сеть открываем нашу сеть HappyLand.net, в разделе классификации ставим наш файл Partition.clu, помечаем галочкой. Визуализация сети со сгруппированными вершинами представлена на рис. 13.

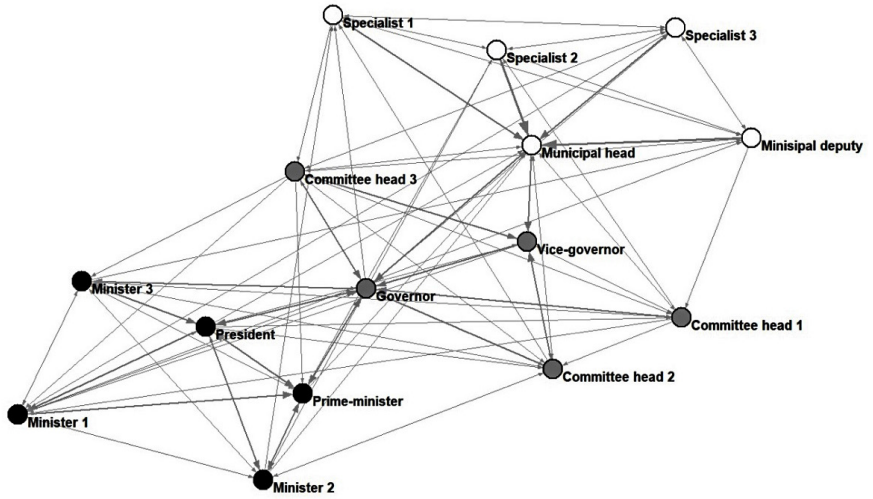


*Рис. 13. Визуализация уровней власти в стране Хэппилэнд*

Теперь можно создать проектный файл с расширением .paj. Для этого нужно указать в первом окне сети нашу сеть HappyLand.net, в первом окне классификации наш файл Partition.clu. Затем выполнить команду File>Pajek Project File>Save. Теперь у нас появился проектный файл HappyLand.paj. С ним можно работать далее. Теперь мы можем воспользоваться командой File>Pajek Project File>Read, и в окне сети, и в окне классификации одновременно откроются документы, сохраненные в проектном файле.

Второй способ построения сети позволяет нам использовать непосредственно функционал Pajek. Для начала необходимо создать пустую сеть при помощи команды Network>Create New Network>Empty Network. В диалоговом окне нужно напечатать то количество вершин, которые нам нужны. У нас появляется новая сеть в Network drop-down menu. Если визуализировать данную информацию, то пока мы увидим только набор вершин, расположенных в виде эллипса. Продолжаем рисовать, используя команду File>Network>View/Edit. Эта команда открывает диалоговое окно, которое позволяет выбрать вершину по ее порядковому номеру или названию. Появляется Editing Network screen. В нем мы двойным щелчком активируем Newline, что позволяет пользователю добавить линию в/из выбранной вершины. Например, вершина «1». Чтобы добавить двухстороннюю дугу, нужно просто напечатать номер другой вершины. Если ввести знак «+» перед новой вершиной, то добавляется дуга в выбранную вершину (дуга в вершину «1»). Если ввести знак «-», то будет добавлена дуга из выбранной вершины, т.е. из вершины «1» в новую вершину. Каждая новая линия или дуга отражается как линия в Editing Network Screen. Обозначение «4.1» – это дуга из вершины «4» в вершину «1». Обозначение «1.3» – это дуга из вершины «1» в вершину «3». И, наконец, «1–2» – это линия (двухсторонняя дуга) из вершины 1 в вершину 2. Линию можно удалить, кликнув по ней дважды.

Полученную сеть нужно сохранить с расширением «.net». Далее через программу Блокнот/Notepad мы можем редактировать названия вершин, так как Pajek обозначает их цифрами. В ручном режиме, открыв документ в Блокноте, мы сможем переименовать данные вершины. Вообще важно помнить, что Pajek не сохраняет сети и их модификации по умолчанию. Если нужно что-то сохранить, то желательно переименовать сеть в момент сохранения, чтобы не стереть предыдущие данные.

В Pajek мы можем работать с шестью разными диалоговыми окнами, которые соответственно открывают документы с разными расширениями. Это Сеть/Network, Классификации/Partition, Векторы/Vectors, Перестановки/Permutation, Кластеры/Clusters, Иерархии/Hierarchy. В случае классификаций документы имеют расширение «.clu», в случае векторов расширение «.vec». Также мы можем работать с проектными файлами с расширением «.paj», которые объединяют несколько разных типов документов в Pajek.



Для более подробного знакомства с программным обеспечением Pajek можно приобрести учебник W. de Nooy, A.Mrvar, V.Batagelj «Exploratory social network analysis with Pajek» с базами данных [de Nooy, Mrvar, Batageli, 2018. с. 36–222, 149–169]. Задачей же данного учебного пособия является общий обзор возможностей использования Pajek с объяснением сути наиболее часто используемых команд.

**Использование классификаций для упорядочивания данных**

Следующим диалоговым окном Pajek являются Классификации/Partitions. Они позволяют распределить вершины по принадлежности к определенным группам. Группы задаются исследователем произвольно в соответствии с дизайном исследования. Посмотреть принадлежность вершин к группам можно при помощи View/Edit на вкладке Partitions. Посмотреть распределение вершин по группам (частотный анализ) можно при помощи команды Info на вкладке Partitions. Классификации/Partitions позволяют не только распределять вершины (узлы) по группам, но и производить некоторые манипуляции с сетью. Допустим мы хотим посмотреть, как выглядят отношения между акторами на федеральном уровне. Для этого мы используем команду Operations>Network+Partition>Extract>SubNetwork Induced by Union of Selected Clusters. В появившемся окне мы набираем цифру «1», которой закодирован данный уровень власти. Результаты визуализации представлены на рис. 14.

Если мы хотим посмотреть отношения между уровнями власти, мы можем сжать совокупность актров на каждом уровне до одного узла. Для этого мы используем команду Operations>Network+Partition>Shrink Network. В диалоговом окне, в котором нас спрашивают о минимальном количестве связей, мы ставим «1». Названия уровне власти добавляем вручную через функцию File>Partition>View/Edit. По умолчанию каждый уровень власти будет обозначен названием должности, которая стоит первой в списке по алфавиту. Данная команда позволяет нам посмотреть характер связей между уровнями власти.

Если мы захотим проверить, как выглядит отношения между уровнями власти и внутри конкретной властной группы, мы можем использовать ту же команду, но указав во второй строке диалогового окна номер группы, который не нужно «вырезать», например, единицу. Тогда мы получим следующий рисунок сети (См. рис. 16).

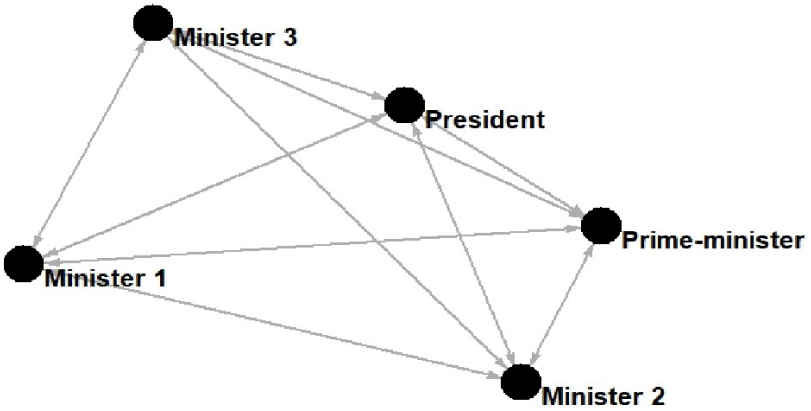


*Рис. 14. Отношения между акторами на федеральном уровне власти в стране Хэппилэнд*

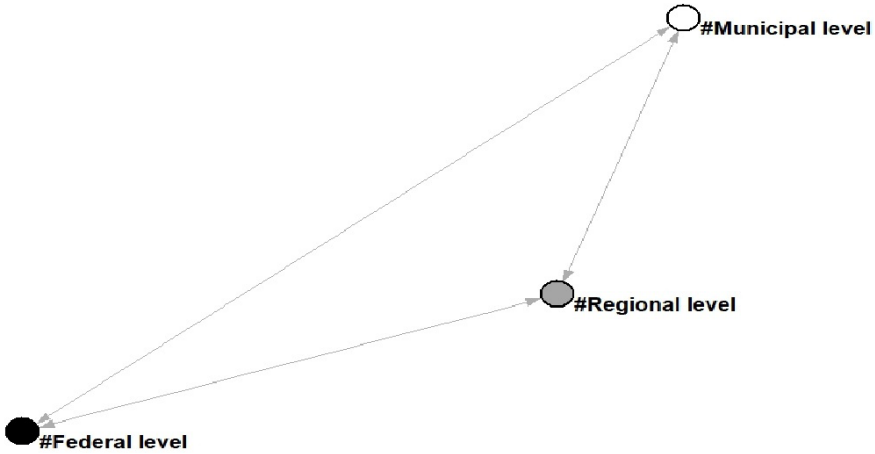


*Рис. 15. Отношения между уровнями власти в стране Хэппилэнд*

Информация, касающаяся частотного распределения акторов, расчеты показателей сети может являться важной, но отображается на вспомогательной вкладке Pajek. Для того, чтобы вызвать эту вкладку необходимо выполнить следующую команду: Info>Child Windows>Report Window>Show.

Для визуализации сетей можно использовать не только Классификации/Partitions и Векторы/Vectors, но и другие инструменты такие, как Компоненты, Ядра/K-cores, Кластеры. Эти инструменты позволяют определить наиболее плотные участки сети. О них мы поговорим подробнее.



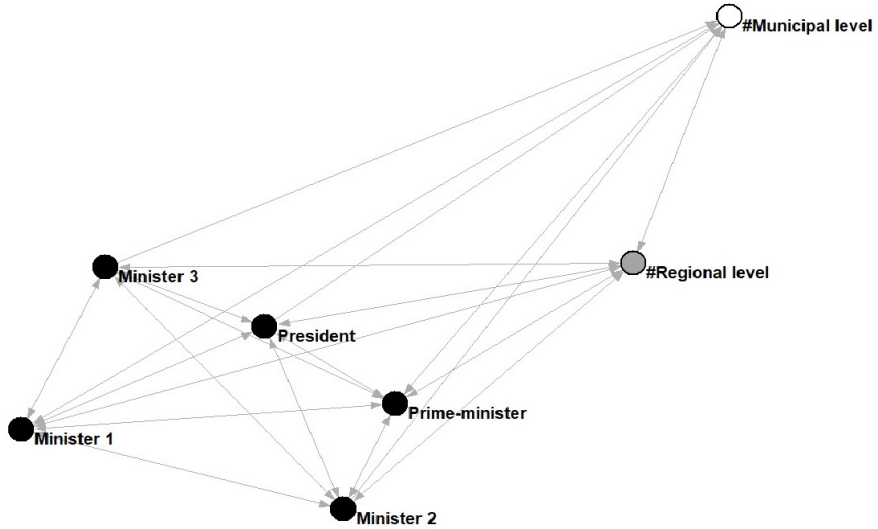


*Рис. 16. Отношения между акторами внутри федерального уровня власти с другими уровнями власти*

**Определение плотных участков сети при помощи Pajek**

Плотные участки сети связаны с тем, что акторы, образующие сети, в своих подгруппах могут быть связаны не только взаимодействием, но и солидарностью, групповыми нормами, общими интересами. Более тесные контакты внутри подгрупп сети связаны с явлением гомофелии или ассортивности, когда подобные акторы чаще контактируют с себе подобными.

Прежде чем перейти к объяснению визуализации плотных участков сети, определим простые числовые характеристики сетей. Начнем с плотности сети. Плотность сети представляет собой отношение числа всех существующих взаимосвязей в сети ко всем возможным. Полная сеть – это такая сеть, в которой все вершины взаимосвязаны. Ее плотность равна единицы. Она берется за эталон. Все остальные сети сопоставляются с этой эталонной сетью и их плотность колеблется в пределах 0,05–0,6. Множественные линии между вершинами служат индикатором более плотных участков сети. С увеличением размера сети увеличивается ее плотность.

Следующей базовой числовой характеристикой является степень вершины, которая определяется количеством линий, связанных с вершиной. Это более полезная характеристика сети, так как не зависит от ее размера. Необходимо отметить, что вершины с большей степенью

находятся в более плотных участках сети. Более высокая степень вершин ведет к увеличению плотности сети, так как такие вершины содержат больше связей. Еще одним базовым показателем является средняя степень. Он используется для измерения структурной связанности сети. Подобно степени вершины данный показатель не зависит от размера сетей и может быть использован для сравнения сетей.

Для расчета плотности сети и средней степени вершин используем команду [Main] Network>Info>General. Если нам интересна только плотность и средняя степень, то вводим 0. Pajek показывает плотность с петлями и без. Нас интересует показатель плотности сети без петель.

Для того, чтобы определить среднюю степень сети для ненаправленной сети преобразуем HappyLand.net в HappyLand_symmertized.net, используя команду Network>Create New Network>Transform>Arc->Edges>All. В диалоговом окне нам предлагают указать одну из 6 опций, выберем «1». Сохраняем новую сеть HappyLand_symmertized.net при помощи команды File>Network>Save. Далее мы можем рассчитать степень группы, используя команду: Network>Create Partition>Degree>submenu: Input, Output, All. Поскольку сеть ненаправленная, мы можем выбрать любой из предложенных вариантов (Input, Output, All). Для того, что посмотреть степень группы, используем команду Partition>Info. Номер кластера отражает степень группы, но мы не можем рассчитать среднюю степень группы, поэтому считаем среднюю степень сети. Считаем среднюю степень для ненаправленной сети Attiro_symmertized.net, используя команду Network>Info>General. Во вспомогательной вкладке получаем плотность сети 0,65, средняя степень сети 9,07. Это означает, что плотность сети довольно высока, и в среднем представители власти довольно часто контактируют друг с другом.

Плотные участки сети можно идентифицировать с использованием ряда команд Pajek и получить не только их числовые характеристики, но и визуализировать. Первой характеристикой плотных участков сети служат компоненты. Они бывают слабые, т.е. представляют собой максимально слабо связанную подсеть, и сильные, которые представляют собой максимально сильно связанную подсеть. Для характеристики слабых и сильных компонентов сети используется понятие тропы (полутропы). Тропой называется такая ситуация, когда в сети мы можем двигаться только по направлениям дуг, не повторяя вершины. Такая сеть является сильно связанной. Если же мы пренебрегаем направлением дуг и движемся по сети в любом направлении, но не повторяя вершины, то такая сеть называется слабо связанной. Для нахождения сильных



компонентов сети используем направленную сеть Attiro.net. Команда Network>Create Partition>Component>Strong позволяет выявить одну сильную компоненту сети, которую также можно представить визуально при совместном использовании сети и классификации. Поскольку количество слабых компонентов в направленной сети равно общему количеству компонентов в ненаправленной сети, используем сеть HappyLand_symmertized.net и команду: Network>Create Partition>Component>Week. В ненаправленной сети количество слабых компонент равно единице.

Характеристика Ядра/K-cores также указывает на плотные участки сети, где K-cores – это кластеры, а K – это минимальная степень каждой вершины внутри ядра. Так 3-cores содержит все вершины, которые связаны со степенью 3 и более с другими вершинами. Таким образом, K-core – это максимальная подсеть, в которой каждая вершина имеет, по крайне мере, степень K в подсети. Данные ядра «гнездятся друг в друге». Так вершина в 3-Cores является частью 2-Cores. Однако не все члены 2-Cores принадлежат 3-Cores. Таким образом, удаление из сети ядер с более низкой степенью позволяет идентифицировать более плотные участки компоненты сети.

Еще одна характеристика плотности сети – это клика. Клика представляет собой набор вершин, в котором каждая вершина напрямую связана со всеми другими вершина, т.е. подсеть максимальной плотности состоит из трех вершин и более. Клики могут перехлёстываться, так как одна и та же вершина может принадлежать разным кликам. Минимальная клика представляет собой триаду, т.е. максимально плотный участок сети, состоящий из трех вершин. Для определения всех полных триад внутри сети, используем фрагмент сети triad_undir.net в первом окне и ненаправленную сеть HappyLand_symmertized.net. Далее найдем все полные триады в сети, используя команду Networks>Fragment (First in Second). В разделе Иерархия/Hierarchy появится новый документ. Корень иерархии – это узел, который связывает все группы. Он равен количеству вершин во всех группах: 126 групп умножить 3 вершины равно 378. Чтобы посмотреть на набор триад заходим в иконку View/Edit Hierarchy и кликаем на плюс слева около корня. После этого раскрывается набор из «126» триад, в каждой из которых можно посмотреть, какие вершины входят, кликнув на левую кнопку мыши.

### Расчеты центральностей вершин и централизации сети

Акторы в сети могут располагаться или ближе к центру, или ближе к периферии. Pajek позволяет произвести расчеты показателей центральности акторов и централизации сети. Показатели центральности

используются для определения позиций индивидуальных вершин. Показатель централизации характеризует всю сеть. Высокоцентрализованная сеть имеет четкую границу между центром и периферией. В этой сети распространение информации идет быстро, но центр незаменим для передачи информации.

Для измерения центральности используются следующие показатели: степень вершины, близость к центру, центральность по посредничеству и центральность по собственному вектору. Степень центральности вершины – это ее степень, т.е. количество линий, которые она содержит. Для анализа показателей центральности используем наш файл HappyLand.paj.

Степень вершин в Pajek вычисляется с использованием команды: Network>Create Partition>Degree>All. Степень вершин можно посмотреть при помощи команды Partition>Info. Номер кластера -–это степень вершины или количество линий, связанных с данной вершиной. В Pajek группирование (partition) – это классификации, приписывающие вершины кластерам. Однако для вычисления центральностей вершин и централизации сети используют векторы. Для того, чтобы рассчитать степень централизации сети, необходимо использовать команду Network>Create Vector>Centrality>Degree. В нашей сети степень централизации составляет 0,32. Этот показатель имеет смысл, только при сравнении с другими показателями.

Следующий показатель – это близость вершины к центру, который определяется как количество остальных вершин, деленное на сумму всех расстояний между этой вершиной и другими вершинами. Централизация близости – это вариация близости вершин к центру, деленное на максимальную вариацию близости к центру, возможную для сети данного размера. Расчет показателей близости вершин к центру и централизации близости осуществляется посредством команды: Network>Create Vector>Centrality>Closeness>All. Значение показателей близости вершин к центру можно посмотреть при помощи команды: View/Edit.

Центральность по посредничеству показывает, насколько часто актор является посредником между любыми другими двумя участниками, находясь на кратчайшем пути между ними. Централизация по посредничеству представляет собой отношение вариации центральности по посредничеству к максимальной центральности по посредничеству, которая возможна в сети данного размера. Для расчета этих показателей воспользуемся командой: Network>Create Vector>Centrality>Betweenness. Показатели централизации отражаются во вспомогательной вкладке Pajek (Report screen).



Показатель по собственному вектору демонстрирует зависимость между положением актора (например, его приближенности к центру) и центральности связанных с ним других акторов. Наибольший показатель центральности по собственному вектору будет у того актора, у которого много связей, и который связан с другим актором, у которого также много связей. Для расчета центральности по собственному вектору и показателя централизации сети по собственному вектору используем команду Netwrok>Create Vector>Centrality>Hubs-Authorities.

Еще одной характеристикой положения актора в сети является расстояние, которое определяется как количество шагов или посредников для кого-то, чтобы достичь индивида в сети. Чем меньше расстояние между индивидами, тем легче обмен информацией. Geodesic – это наиболее короткое расстояние между вершинами. В нашем примере наибольшей степенью обладает губернатор (Governor). Для расчета расстояния между данным актором и другими вершинами используем команду: Network>Create Partition>k-neighbors. Данная команда создает группы классов, указывающие на расстояние между актором Governor и другими вершинами.

Кратчайшее расстояние между двумя вершинами можно найти, используя команду: Netwrok>Create New Network>Subnetwork with Path>All Shortest Path between Two Vertices. Найдем кратчайшее расстояние между двумя крайними вершинами «President» и «Committee head 1». На вопрос «Forget values of lines?» отвечаем утвердительно, так как мы не хотим взвешивать линии в соответствии с их значениями. Кратчайшее расстояние между этими вершинами равно двум.

Таким образом, программное обеспечение Pajek дает нам богатые возможности по построению и визуализации сетей, являясь важным инструментом при проведение сетевого анализа не только в социологии, но и в политической науке. Его несомненным преимуществом является также возможность математической оценки положения акторов и определения их ролей в сети. Методология сетевого анализа является несомненно уникальным подходом к исследованию социальных и политических феноменов, и ее недостаток в отношении объяснительного потенциала будет со временем преодолён.

**Блок самопроверки**

Построить сеть дружбы из 10 акторов с использованием команды Kamada-Kawai, посчитать центральности вершин и дать им интерпретацию.

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ


1. Ермолаев О.Ю. Математическая статистика для психологов: учебник. М.: Московский психолого-социальный институт: Флинта, 2003. – 335 с.

2. Babbie E. The Basics of Social Research. Wadsworth: Cengage Learning, 2008. – 552 p.

3. Квале С. Исследовательское интервью. М.: Смысл, 2003. – 301 с.

4. Козлов М.В. Планирование экологических исследований: теория и практические рекомендации. М.: КМК, 2014. – 169 с.

5. Сергеева И.И., Чекулина Т.А., Тимофеева С.А. Статистика / И.И. Сергеева, Т.А. Чекулина, С.А. Тимофеева. М.: Инфра-М, 2016. – 227 с.

6. Ноэль Э. Массовые опросы: введение в методику демоскопии. М.: Прогресс, 1978. – 382 с.

7. Кун Т. Структура научных революций. М.: Издательство АСТ, 2003. – 605 с.

8. Масалков И.К., Семина М.В. Стратегия кейс стади: методология исследования и преподавания. М.: Академический Проект; Альма Матер, 2011. – 443 с.

9. Сартори Дж. Искажение концептов в сравнительной политологии // Полис. Политические исследования. – 2003. – № 3. https://doi.org/10.17976/jpps/2003.03.07

10. Goktug M., Ivanova N. Methods Taught in Public Policy Programs: are Quantitative Methods still Prelevant? // Journal of Public Affairs Education. Vol. 16, no 2 (Spring 2010). – P. 255–277.

11. Dunn W. Public policy analysis. 3d ed. Pearson: Prentice Hall, 2004. – 510 p.

12. Брайман А., Белл Э. Методы социальных исследований. Группы, организации, бизнес / Пер. с англ. Харьков: Изд-во Гуманитарный центр, 2012. – 774 с.

13. Поппер К. Логика научного исследования. М.: Республика, 2005. – 447 с.

14. Добреньков В.И., Кравченко А.И. Методы социологического исследования: учебник. М.: Инфра-М, 2016. – 767 с.

15. Козлов М.В. (2014). Планирование экологических исследований: теория и практические рекомендации. М.: КМК. – 169 с.

16. Бусыгина Н.П. Методология качественных исследований в психологии: учеб. пособие. М.: Инфра-М, 2014. – 304 с.

17. Бусыгина Н.П. Качественные и количественные методы исследования в психологии: учебник для бакалавриата и магистратуры. М.: Издательство Юрайт, 2017. – 423 с.





18. Наследов А.Д. Математические методы психологических исследований: анализ и интерпретация данных: учеб. пособие / А.Д. Наследов. СПб: Речь, 2004. – 384 с.

19. R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Электронный ресурс]. – URL: http://www.R-project.org/ (дата обращения 04.11.2020)

20. RStudio (2018). RStudio: Integrated development environment for R (Version 1.1.453). Boston, MA. [Электронный ресурс]. – URL: http://www.rstudio.org/ (дата обращения 14.11.2020)

21. Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014. – 572 с.

22. Hand D.J. et al. A Handbook of Small Data Sets, N.Y: Chapman and Hall, 1994. – 392 p.

23. Лонг Дж., Титор П. R. Книга рецептов: Проверенные рецепты для статистики, анализа и визуализации. М.: ДМК, 2020. – 503 с.

24. Низаметдинов Ш.У., Румянцев В.П. Анализ данных: учеб. Пособие. М.: НИЯУ МИФИ, 2012. – 288 с.

25. Faraway J. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. London: CRC Press, 2016. – 399 p.

26. Сморгунов Л.В., Шерстобитов А.С. Политические сети: Теория и методы анализа: Учебник для студентов вузов / Л. В. Сморгунов, А. С. Шерстобитов. М.: Издательство «Аспект Пресс», 2014. – 320 с.

27. Мирошниченко И.В. Сетевой подход в политических исследованиях: содержание и направления развития // Человек. Сообщество. Управление. – 2013. – № 3.

28. Rhodes R. Policy Network Analysis. In M. Moran, M. Rein and R. E. Goodin (Eds.) The Oxford Handbook of Public Policy. Oxford: Oxford University Press), 2006. – P. 423–45.

29. Rhodes R. Understanding Governance: Ten Years On // Organization Studies, 2007. Vol. 28. no 08. – P.1–22.

30. De Nooy W., Mrvar A., Batagelj V. Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software. Third Edition. Cambridge: Cambridge University Press, 2018. – 455 p.

31. Ядов В.А. Стратегия социологического исследования: описание, объяснение, понимание социальной реальности: учеб.пособие. М.: Издательство «Омега-Л», 2009. – 567 с.

32. Knoke D., Pappi F., Broadbent J., Tsujinaka Y. Comparing Policy Networks. Labor Politics in the U.S., Germany and Japan. New York: Cambridge University Press, 1996. – 288 p.

*Учебное издание*

Игнатьева Ольга Анатольевна

КОЛИЧЕСТВЕННЫЕ МЕТОДЫ
АНАЛИЗА ПУБЛИЧНОЙ ПОЛИТИКИ

*Учебно-методическое пособие*

ЦНИТ «АСТЕРИОН»
Заказ № 226. Подписано в печать 14.12.2021. Бумага офсетная
Формат 60×84$^{1}/_{16}$. Усл. печ. *л*. 3,5. Тираж 300 экз.
Санкт-Петербург, 191015, а/я 83, тел. /факс (812) 685–73–00, 970–35–70
✉ : asterion@asterion.ru 🌐 : http://www.asterion.ru
: https://vk.com/asterion_izdatelstvo